https://doi.org/10.3011/ESARDA.IJNSNP.2021.9

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

Lee Burke, Karl Pazdernik, Daniel Fortin, Benjamin Wilson and Rustam Goychayev
Pacific Northwest National Laboratory
902 Battelle Blvd, Richland, WA 99354
E-mail: lee.burke@pnnl.gov

John Mattingly
North Carolina State University
Raleigh, NC 27695

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.

Keywords: nuclear; energy; language; classification

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 prioritize 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 [1]. 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. Extant approaches to processing such information, to the limited extent it is attempted, rely heavily on manual analysis by humans, a method constrained by time and subject-matter expertise. A new approach would help the IAEA to develop capabilities toward the detection of nuclear technology use or transfer through analysis of technical publications.

The amazing progress of state-of-the-art NLP methods has opened up new opportunities for nuclear domain

Burke, L., Pazdernik, K., Fortin, D., Wilson, B., M., Goychayev, R., & Mattingly, J. (2021). NukeLM: Pre-Trained and Fine-Tuned Language Models for the Nuclear and Energy Domains. *ESARDA Bulletin - The International Journal of Nuclear Safeguards and Non-proliferation, 63, 30-40.* https://doi.org/10.3011/ESARDA.IJNSNP.2021.9

researchers to leverage powerful language models. Models like BERT [2] have shown significant improvement in NLP benchmarking metrics, such as the General Language Understanding Evaluation (GLUE) benchmark [3]. 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 [4], 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 [5]—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 NukeLM, 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 un-labeled data using a self-supervised task, then fine-tuned on a supervised task in the same domain.

Word embeddings (e.g., word2vec [6], GloVe [7], fastText [8]) 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 [9] 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 [2] learns to encode context from both left and right at once using a very large architecture of stacked Transformers [4], 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 [10] 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 [11] 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. [5] 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 [12] continues pre-training on biomedical corpora. NukeBERT [13] 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 NukeLM, 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 Centre 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 [14] obtained in November 2018, amounting to nearly two million abstracts from over 70 years of research results from DOE and its predecessor agencies. No pre-processing is performed on these abstracts; they are analyzed as they appear in the database.

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 [15], 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. These categories are all assigned to the positive class in the binary classification problem ("NFC-related", referring to the nuclear fuel cycle [NFC]), regardless of the step or steps of the Physical Model to which they correspond. 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 [16], 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. [5], 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. For the remainder of this manuscript, we use the naming convention NukeLM to define RoBERTa Large with continued pre-training on OSTI abstracts. 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. [5]). Other hyperparameters follow Gururangan et al. [5].

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. [5] by passing the final layer [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. [5].

5. Results of the Language Modelling 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. As in RoBERTa-style pre-training, one token per sample is masked randomly, without consideration of sub-word status, stop words, or other factors.

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.

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

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	Top-5 Predictions	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
RoBERTa Large	**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

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.

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. [17]

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, evidence that its large size allowed it to learn from pre-training alone some subtleties of the nuclear domain that the smaller models did not. After continued pre-training, all three models regardless of size succeed in predicting the correct answer with high confidence.

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 several times more trainable parameters.

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

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**.

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 (and, in the case of SciBERT, accuracy), 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.

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

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**.

Moreover, numerically small improvements in performance metrics belie the very large size of the datasets presented here. An analyst attempting to filter a corpus as large as OSTI into a more manageable size would be well-served to choose NukeLM over the other models discussed above; a single percentage point change could translate to thousands of relevant papers that would have been missed, or irrelevant papers requiring manual inspection. Indeed, this use-case motivates a preference for recall (the fraction of true positives predicted to be positive) over precision (the fraction of predicted positives which are truly positive), further widening NukeLM's advantage over its competitors in our quantitative assessments.

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. This metric is computed via the Kullback-Leibler divergence, a measure of dissimilarity between the true and predicted probability distributions over the output categories, averaged over the test set. Twenty repetitions with different random seeds are performed. For visual convenience, the probability density function of each of these sets of repetitions is estimated using the kernel density estimation technique, analogous to a smoothed histogram. Generally, lower log-loss indicates better predictions, and greater separation of distribution

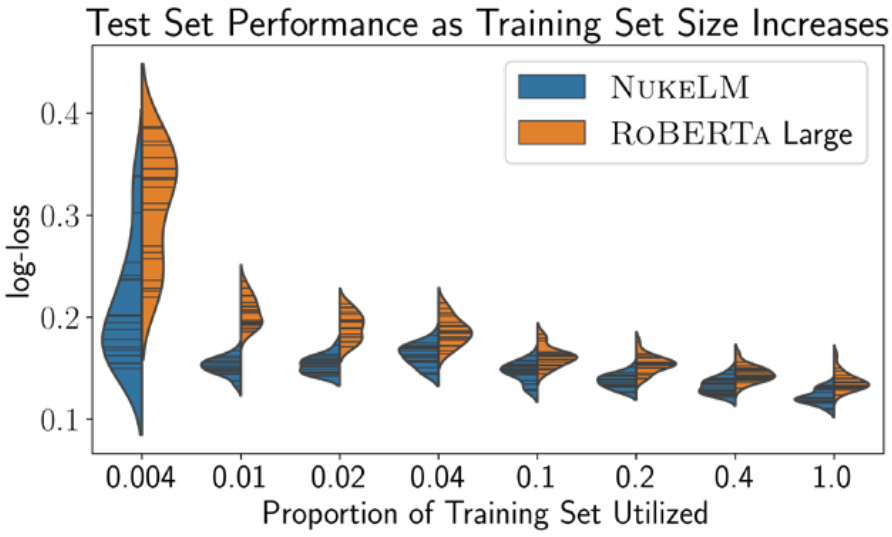

Figure 1: Binary classification performance, measured by log-loss on a hold-out test set, as the training set size is increased, for RoBERTa Large (orange) and NukeLM (blue). Hash marks are each of 20 repetitions with different random seeds, while filled areas are kernel density estimations.

density indicates more significant differences. Figure 1 summarizes the results.

While Table 4 summarizes some performance metrics of NukeLM compared to other models on the full fine-tuning data set for the binary classification task, this experiment provides insight into the potential benefit of using NukeLM on other fine-tuning tasks where the amount of labeled training data is likely to be on the order of one thousand and not one hundred thousand. The domain-adapted model achieves significantly better performance with smaller amounts of data; however, this advantage shrinks as the fine-tuning set size increases. This could be the result of continued pre-training priming the model for performance in this domain.

Therefore, the question of significance between NukeLM and other models is best understood as a function of fine-tuning training set size. Moreover, Figure 1 shows that even though the performance gain decreases with increasing amount of fine-tuning data, we still observe superior performance using NukeLM with a fine-tuning set of nearly two hundred thousand documents.

Interestingly, the disparity between models is less apparent at the lowest training set size tested (0.4% of the full corpus, or 754 documents). While NukeLM maintains its superiority, with so few examples used for fine-tuning, significant instability is observed over the repetitions. A follow-up experiment implements several strategies for stabilizing fine-tuning of large language models discussed in Zhang, et al. [18], but none have a major impact (see Appendix C for details) and are not employed further.

6.4 Qualitative Assessment

Beyond model performance on the MLM task and document classification, an important question regarding these trained language models is whether any reasonable interpretation can be made of the intermediate representations of input examples. While there is not a clear consensus on how useful these embeddings can be in providing explanations, with arguments from both sides [19, 20], 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 how prediction was improved.

As a first step toward interpreting the impact of domain-adaptive pre-training, we consider models fine-tuned on the binary classification task and visualize output embeddings from the most accurate models, RoBERTa Large both with (i.e., NukeLM) and without continued pre-training on OSTI abstracts. We use uniform manifold approximation and projection (UMAP) [21] with all default parameters to project the output corresponding to the special token [CLS] down to two dimensions, training separate UMAP projections 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 [22], 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 hierarchical density-based spatial clustering of applications with noise (HDBSCAN) [23] to cluster documents and a class-based TF-IDF (cb-TF-IDF) score for topic modeling. TF-IDF stands for term frequency and inverse document frequency, a standard method for identifying terms used unusually frequently in each document. Here, all documents within the same cluster are concatenated into a single document and then the usual TF-IDF score [24] is computed as follows:

$$\mathrm{cb-TF-IDF}_i = \frac{t_i}{w_i} \times \log \frac{m}{\sum_{j=1}^{n} t_j}$$

where t_i is the frequency of each word in class i, w_i is the 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). Recall that the clustering algorithm is applied to the embeddings after reducing their dimension to 100; visual inspection of the 2-dimensional representation may not fully reflect the shape of the BERtopic clusters. 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 samples are labeled outliers and nuclear documents are sorted into four topics. This provides evidence that continued pre-training taught the model additional knowledge of the nuclear domain, allowing it to characterize different subsets of

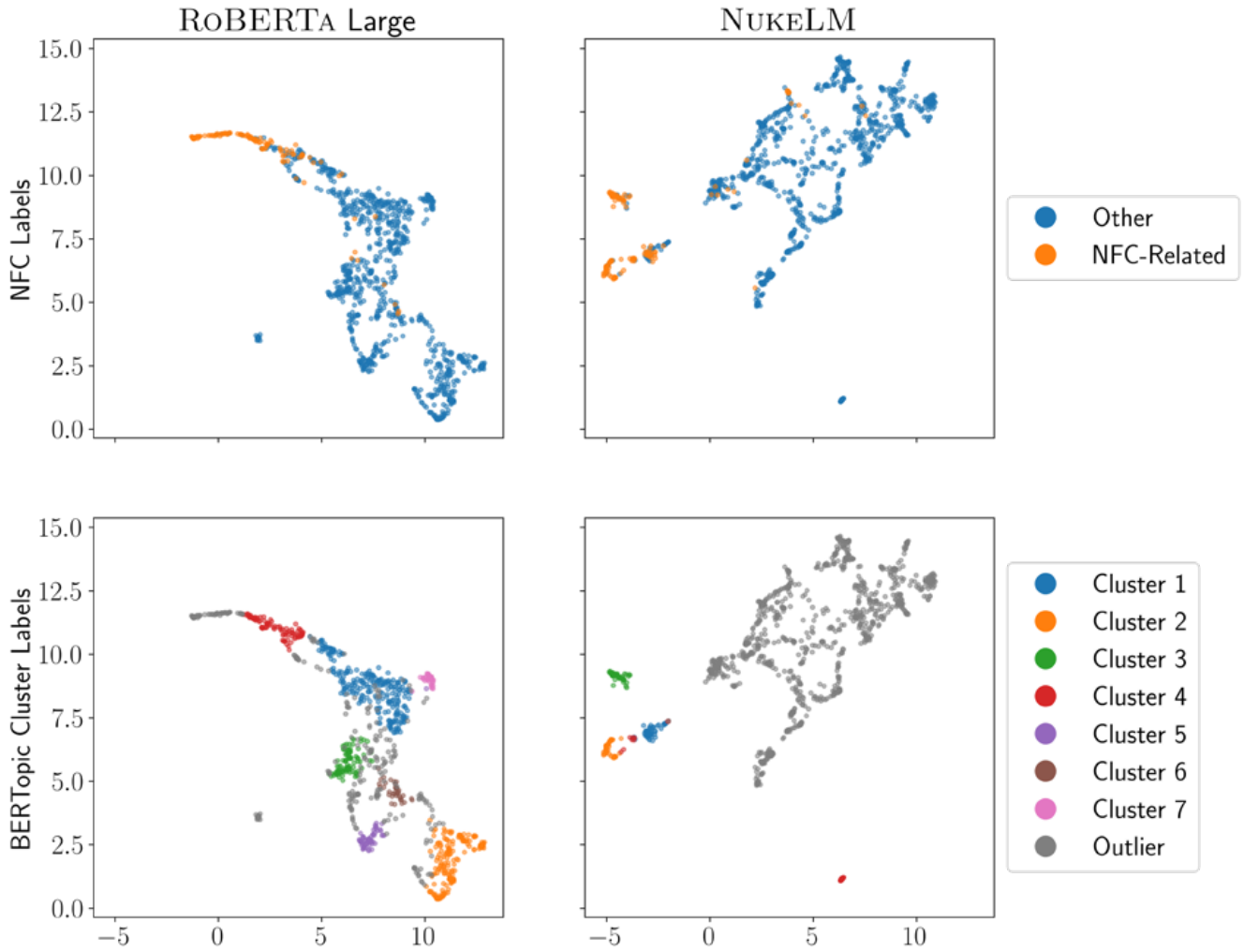

Figure 2: Visualization of UMAP-transformed output embeddings from RoBERTa Large for 1000 randomly sampled documents from the validation set after fine-tuning on the binary classification task, both without (left) and with (right) domain-adaptive pre-training on OSTI abstracts, colored by the true binary labels (top) and BERTopic clusters (bottom). Note that the cluster labels for RoBERTa Large and for NukeLM refer to different document clusters with correspondingly different topics, though they use the same colors. Each point in these plots is a low-dimensional representation of the embedding for a document's abstract.

Model	No.	Top-3 Words
RoBERTa Large	1	beam, ion, states
	2	coal, fuel, oil
	3	films, alloy, materials
	4	waste, nuclear, radiation
	5	cells, protein, cell
	6	soil, acid, conduit
	7	dust, galaxies, observations
NukeLM	1	waste, safety, SAR
	2	reactor, waste, fuel
	3	MeV, nuclei, energies
	4	Scattering, interaction, generation

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 (i.e. NukeLM) domain-adaptive pre-training on OSTI abstracts. Column two, the cluster number, corresponds with the legend in Figure 2 (bottom row).

positive examples, and recognize the irrelevance of other distinctions to the fine-tuning task.

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 down-stream 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. As an example, 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 to prioritize information or to support NLP workflows in nuclear science or nuclear nonproliferation.

The NukeLM binary classification model demonstrated superior performance for the classification task. 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.

Although the performance gains observed were minor, 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.

Additional topics for future work involve expanding the model training pipeline to include full article text and data sets other than OSTI. We will consider 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 [25].

8. Acknowledgements

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.

9. References

[1] Molas-Gallart, J; Which way to go? Defence technology and the diversity of 'dual-use' technology transfer; Research Policy; 26.3; 1997; p 367–385; DOI: 10.1016/S0048-7333(97)00023-1

[2] Devlin, J., et al.; BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding; arXiv:1810.04805; 2019

[3] Wang, A., et al.; GLUE: A Multi- Task Benchmark and Analysis Platform for Natural Language Understanding; arXiv:1804.07461; Feb. 2019

[4] Vaswani, A., et al.; Attention is All you Need; Advances in Neural Information Processing Systems; 30; 2017; p 5998–6008

[5] Gururangan, S., et al.; Don't Stop Pretraining: Adapt Language Models to Domains and Tasks; Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics; Online: Association for Computational Linguistics; 2020; p 8342–8360; DOI: 10.18653/v1/2020.acl-main.740

[6] Mikolov, T., et al.; Distributed Representations of Words and Phrases and their Compositionality; Advances in Neural Information Processing Systems; 26; Curran Associates, Inc.; 2013; p 3111–3119

[7] Pennington, J., R. Socher, and C. Manning; Glove: Global Vectors for Word Representation; Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP); Doha, Qatar: Association for Computational Linguistics; 2014; p 1532–1543; DOI:10.3115/v1/D14-1162

[8] Mikolov, T., et al.; Advances in Pre-Training Distributed Word Representations; Proceedings of the

International Conference on Language Resources and Evaluation (LREC 2018); 2018

[9] Peters, M. E., et al.; Deep contextualized word representations; arXiv:1802.05365; 2018

[10] Liu, Y., et al.; RoBERTa: A Robustly Optimized BERT Pretraining Approach; arXiv:1907.11692; 2019

[11] Beltagy, I., K. Lo, and A. Cohan.; SciBERT: A Pretrained Language Model for Scientific Text; arXiv:1903.10676; 2019

[12] Lee, J., et al.; "BioBERT: a pre-trained biomedical language representation model for biomedical text mining". en. In: Bioinformatics (Sept. 2019). Ed. by Jonathan Wren, btz682. ISSN: 1367-4803, 1460-2059. DOI: 10.1093/bioinformatics/btz682.

[13] Jain, A., N. M. Meenachi, and B. Venkatraman; "NukeBERT: A Pre-trained language model for Low Resource Nuclear Domain". In: arXiv:2003.13821 [cs, stat] (Aug. 2020).

[14] OSTI; The Department of Energy (DOE) Office of Scientific and Technical Information (OSTI); https://www.osti.gov; Accessed: 2018-11

[15] Liu, Z. and S. Morsy; Development of the Physical Model; 2001; URL: http://inis.iaea.org/Search/search.aspx?orig_q=RN:33045150; Accessed: 2020-12-14

[16] Wolf, T., et al.; HuggingFace's Transformers: State-of-the-art Natural Language Processing; arXiv:1910.03771; 2019

[17] Wikipedia; Pressurized heavy-water reactor; Page Version ID: 996963562; Dec. 29, 2020; URL: https://en.wikipedia.org/w/index.php?title=Pressurized_heavy-water_reactor&oldid=996963562 Accessed: 2020-12-29

[18] Zhang, T., et al.; Revisiting Few-sample BERT Fine-tuning; arXiv:2006.05987; 2021

[19] Jain, S. and B. C. Wallace; Attention is not Explanation; arXiv:1902.10186; 2019

[20] Wiegreffe, S., and Y. Pinter; Attention is not not Explanation; arXiv:1908.04626; 2019

[21] McInnes, L., et al.; UMAP: Uniform Manifold Approximation and Projection; The Journal of Open Source Software; 3.29; 2018; p 861

[22] Grootendorst, M.; BERTopic: Leveraging BERT and c-TF-IDF to create easily interpretable topics; version v0.4.2; 2020; DOI: 10.5281/zenodo.4430182

[23] Campello, R. J. G. B., D. Moulavi, and J. Sander; Density-Based Clustering Based on Hierarchical Density Estimates; Advances in Knowledge Discovery and Data Mining; Berlin, Heidelberg: Springer; 2013; p. 160– 172. DOI: 10.1007/978-3-642-37456-2_14

[24] Teller, V.; Speech and language processing: an introduction to natural language processing, computational linguistics, and speech recognition; Computational Linguistics; 26.4; 2000; p 638–641

[25] Conneau, A., et al.; Unsupervised Cross-lingual Representation Learning at Scale. arXiv:1911.02116; 2020

Appendix A: OSTI Subject Categories

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

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

Model	Learning Rate	Batch Size	Accuracy	F1-Score	Loss
RoBERTa Large	1×10^{-5}	16	**0.9545**	**0.9537**	0.1173
		64	0.9506	0.9502	**0.1081**
	$2 \times 1\ 0^{-5}$	16	0.9397	0.9409	0.1568
		64	0.9524	0.9523	0.1118
	5×10^{-5}	16	0.9206	0.9097	0.2260
		64	0.9363	0.9338	0.1699
RoBERTa Large + OSTI	1×10^{-5}	16	**0.9573**	0.9568	0.1127
		64	**0.9573**	**0.9570**	**0.0967**
	2×10^{-5}	16	0.9520	0.9516	0.1340
		64	0.9557	0.9559	0.0977
	5×10^{-5}	16	0.9328	0.9279	0.2093
		64	0.9525	0.9518	0.1108

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

Appendix 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 1×10^{-5}, 2×10^{-5}, and 5×10^{-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.

Appendix C: Stabilizing Few-Shot Fine-Tuning

A further hyperparameter tuning experiment is performed on the binary classification task using RoBERTa Large, both with and without domain-adaptive pre-training, and restricted to only 0.4% of the training set (754 documents). Following Zhang, et al. [18], we perform a grid search over the number of training epochs (3, 6, 12, and 24) and over the number of layers to reinitialize (0 through 6). The layers are chosen from the bottom of the model, nearest the final classification layer, which is always newly initialized. Twenty repetitions with different random seeds are performed. Results on the validation set are summarized in Figure 3, using the same metrics and techniques as in Figure 2.

Zhang, et al. [18], suggests that more training epochs and reinitializing several layers often stabilizes fine-tuning on very small datasets, narrowing the range of results. That does not appear to hold true here: longer training has the opposite effect, and though re-initializing three or four layers may result in a smaller range with three training epochs, the effect mostly reduces the incidence of outliers, so-called "failed runs", rather than making most runs more predictable. Therefore, we do not employ either technique in the main body of this manuscript.

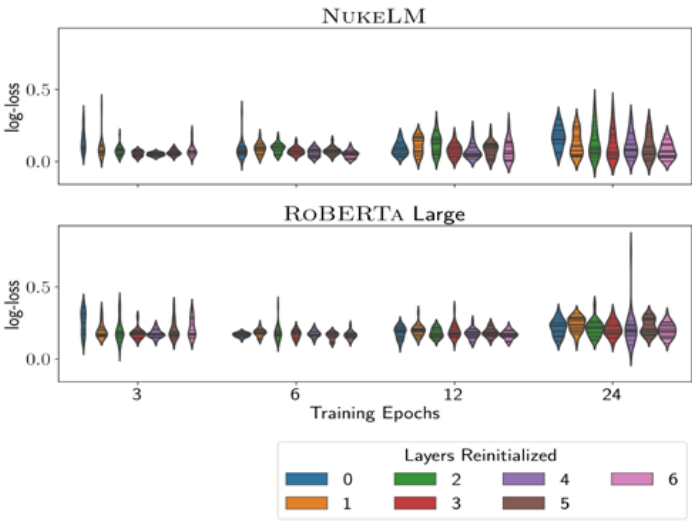

Figure 3: Results of a hyperparameter tuning experiment for number of training epochs (horizontal axis) and number of reinitialized layers (color), with log-loss shown on the vertical axis. Hash marks are each of 20 repetitions with different random seeds, while filled areas are kernel density estimations.