STUDENT SUMMER INTERNSHIP TECHNICAL REPORT

Using Natural Language Processing for Semantic Search in the Nuclear Domain

DOE-FIU SCIENCE & TECHNOLOGY
WORKFORCE DEVELOPMENT PROGRAM

Date submitted:

December 15, 2023

Principal Investigators:

Alejandro De La Noval (DOE Fellow Student)
Florida International University

Thomas Danielson (Mentor)
Savanah River National Lab

Ravi Gudavalli Ph.D. (Program Manager)
Florida International University

Leonel Lagos Ph.D., PMP® (Program Director)
Florida International University

Submitted to:

U.S. Department of Energy
Office of Environmental Management
Under Cooperative Agreement # DE-EM0005213

Applied Research Center
FLORIDA INTERNATIONAL UNIVERSITY

DISCLAIMER

This report was prepared as an account of work sponsored by an agency of the United States government. Neither the United States government nor any agency thereof, nor any of their employees, nor any of its contractors, subcontractors, nor their employees makes any warranty, express or implied, or assumes any legal liability or responsibility for the accuracy, completeness, or usefulness of any information, apparatus, product, or process disclosed, or represents that its use would not infringe upon privately owned rights. Reference herein to any specific commercial product, process, or service by trade name, trademark, manufacturer, or otherwise does not necessarily constitute or imply its endorsement, recommendation, or favoring by the United States government or any other agency thereof. The views and opinions of authors expressed herein do not necessarily state or reflect those of the United States government or any agency thereof.

EXECUTIVE SUMMARY

This research work has been supported by the DOE-FIU Science & Technology Workforce Development Initiative, an innovative program developed by the U.S. Department of Energy's Office of Environmental Management (DOE-EM) and Florida International University's Applied Research Center (FIU-ARC). During the summer of 2023, a DOE Fellow intern, Alejandro De La Noval, spent 10 weeks doing a summer internship at Savannah River National Laboratory under the supervision and guidance of Senior Scientist Thomas Danielson (Ph.D.). The intern's project was initiated on June 5, 2023, and continued through August 10, 2023 with the objective of identifying events of interest relating to worldwide state-sponsored civil nuclear power using natural language processing.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

1. INTRODUCTION

For the sake of worldwide nuclear safety, there is need for awareness about important events pertaining to the nuclear domain. Indicators for these events are open source in the form of social media posts and news articles alongside their official public documentation. Thorough extraction of all possible events indicated by a wide corpus of text data from either web posts or news articles alone is a tedious and time-consuming task. The large amount of text presents an opportunity to apply natural language processing (NLP) techniques to automate event extraction from the nuclear domain. One such technique involves tracking the contextual meaning of key terms across time, by which events can be highlighted by the semantic shifts in those contexts. In other words, the discovery of inflection points in a key term's semantic meaning indicates an event at a given timeframe. Another technique involves the comparison of word embeddings generated from respective datasets to those generated from a pool of formulated guiding sentences. Together with the addition of a model tasked with entity tagging, the foundation for a semantic search engine starts to form. With this NLP platform developed with focus on modularity and user choice, resistance to the additional incorporation algorithms and workflows should be minimal as research continues.

2. RESEARCH DESCRIPTION

The research pursued during DOE Fellow Alejandro De La Noval's internship can be split into three tasks: inflection analysis automation, semantic vector comparison, and entity tagging of text in the datasets used. Before describing each task, note that there were two separate datasets the workflows were run on. There was a Twitter dataset consisting of tweets from 2014 to 2018, and a news dataset from newsapi.ai consisting of articles with their headlines from 2014 to 2022. Most of the development was done using the Twitter data but developed workflows also had support for the news data.

Beginning with the inflection analysis automation, the pipeline outlined in [1] records semantic meaning for a list of key terms across windows of time, with each window having a separate embedding model built with the data belonging to it. The text embeddings are used to build a graph of n terms closest in (vector) similarity to the key terms provided and n terms closest to those similar terms depending on rank, where n is the number of neighbors a user wants to explore, and rank is the depth of the graph. This graph is used to compute weighted embedding vectors as per [1], after which similarity metrics are calculated for each graph and used to compare results between windows via derivation. The derivation of the metrics highlights inflection in scoring trends which in turn indicate a window of interest. Figure 1, taken from [1], shows the inflection points visualized according to their date and key term. The inflection pipeline also produces various figures to aid in analyzing similarities and co-occurrences of key terms with a corpus. Abstracting the pipeline into an executable application such that a layman user can easily run it with their particular search terms, ranking and nearest neighbor settings, and specified outputs is the goal of the automation task. The application would also be amendable enough such that addition of functionality is easily done, turning the application into more of a platform for semantic search as modules accommodate more generalized tasks and datasets in the nuclear domain.

Tackling the task from another angle, a pool of curated "target" sentences would be used to guide the search for important events. These sentences are representative of what a subject matter expert would be looking for when extracting nuclear events. By taking the embedding of these sentences as a whole (as opposed to individual word embeddings used in inflection pipeline) and comparing them to the representative sentence embedding vectors of our dataset, a user can extract all nuclear events that mirror the semantics of the guiding sentences. For example, a sentence like "[Entity X] and [Entity Y] have agreed to a deal to construct four nuclear reactors in [Z location]", would ideally bring up hits against our data pertaining to nuclear construction dealings or any dealings in nuclear between actors. That is, the search would be based on the context of the sentences rather than strict comparison of their vector representations. Thus, it is important that the vectors being compared sufficiently capture semantic context. To achieve this, a pre-trained model fine-tuned on the sentence similarity task (via a contrastive learning objective) was used to generate the embeddings. The model is based on a 6 layer version of the MiniLM-L2-H384 distilled transformer model referenced in [2]. The tokenizer and model can be loaded from the Hugging Face API under "'sentence-transformers/all-MiniLM-L6-v2" [5]. Once the embeddings are generated, two methods of extraction were developed. One simply involved calculating the cosine similarity of every sentence-piece embedding in the data against every sentence-piece embedding in the pool of guiding sentences, then filtering the data based on a threshold value that determines a minimum required similarity score needed for extraction. The other method involved using a Faiss index.

Faiss uses its own methods for calculating similarity as shown in [3]. This method allowed us to query each sentence in the pool for its n nearest neighbors, and was more efficient since the indexing is built into the dataset manager. This could therefore be the better method of performing a guiding sentence pool-based semantic search at the dataset scale.

Research related to the comparison of semantic vector brought about the idea of tagging all the data (including the guiding sentences) with their entity type, so that different words that belong to the same entity type (and thus the same context) do not confuse the model during prediction. This would lead to more hits at higher similarity scores when searching. To do this, another model was developed for the purpose of entity tagging. This model needed to be fined-tuned to our dataset, so DistilBERT [4] was chosen as it is a low resource, comparatively raw pre-trained model meant to be trained for downstream tasks. This model was trained to transform an input sentence into its tagged versions according to the IOB (inside, outside, beginning) format. This allowed us to capture entities that span across multiple words or tokens as it was specified whether a token marked the beginning, middle, or end of an entity. After prediction, tokens then needed to be merged out of the IOB format into their tags so that the processed data could be used in a more efficient semantic search as mentioned before. This involved a degree of manual annotation of some of our data, as this workflow was supervised and thus required labeled data.

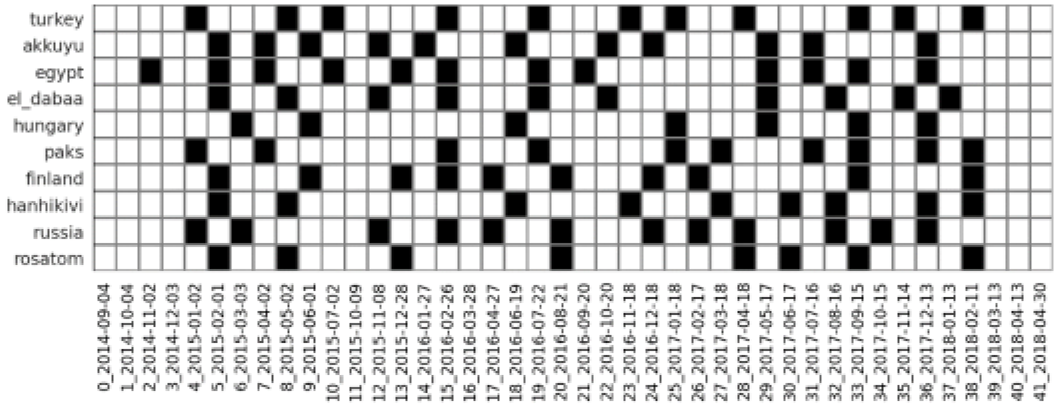


Figure 1. Contextual inflection points computed using the graph similarity metrics for the key terms of interest on the Twitter dataset [1].

3. RESULTS AND ANALYSIS

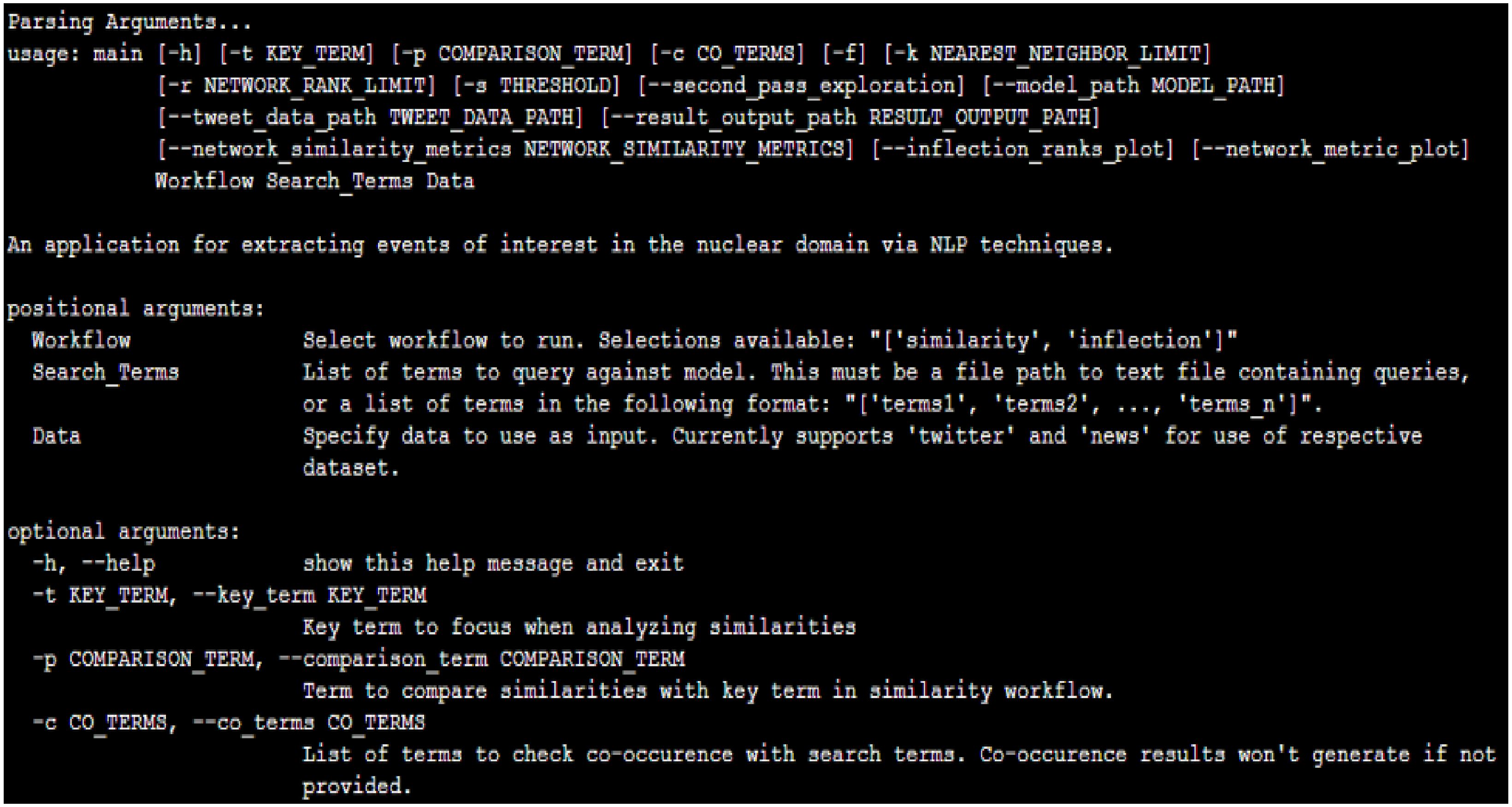
The final executable for the inflection analysis has options selectable via console arguments that determine which workflows and outputs a user would want for a particular dataset. By supplying the –h or –help flag, a user can pull up documentation on all the functionality of the application as shown in Figure 2. Through the use of arguments, a user can specify the ranking of the network graphs for querying and the threshold for similarity on the extraction. Note, the threshold only affects similarity comparison towards words outside of the vocabulary of the word embedding models. Based on defaults provided in a configuration file, or input provided by the user, output is saved at a specified directory for user inspection. The inflection workflow ultimately produces a .csv file of extracted events from the given dataset as exemplified in **Error! Reference source not found.**, while the similarity workflow produces a .csv file that shows the most similar terms to a key term in each of the window intervals from which the word embedding models were built. This is exemplified in **Error! Reference source not found.**. Both workflows have accompanying visualizations to aid in analysis. Caching implementation into the inflection pipeline has also been developed to speed up subsequent runs of the application, as computation could take a long time depending on the selected options. The caching optimization takes select parts of the code from ~800ms down to ~100ms per iteration, which adds up to a substantial time savings when providing multiple key terms to query.

Table 1. Selected Examples of Results for Key Term "Akkuyu" from Event Extraction

First_Date	*Similarity_Score*	*Capture_Via*	*Original_Tweet*	*Count*
2014-08-05T20:21:30.000Z	[0.649482250213623]	base_embedding	'Earthquake risk is hidden in Akkuyu' Earthquake expert Övgün Ahmet Ercan, for the nuclear power plant planned to be built in Mersin.. http://t.co/bzUy3Kpl0F	1.0
2014-08-23T15:51:15.000Z	[0.7018592953681946]	base_embedding	EIA Process of Akkuyu Nuclear Power Plant Project Continues Without Notifying the Public http://t.co/dZsPAnb0nQ	2.0
2015-04-18T06:20:15.000Z	[0.6212456822395325]	base_embedding	People of Mersin, do not sleep, Akkuyu thermal power plant will destroy your entire agricultural investment, may God help you. Or rather, all of us..	1.0
2016-05-17T10:15:22.000Z	[0.6192637085914612]	base_embedding	Do you expect us to believe that the nuclear power plant you will build in Akkuyu on the seaside, at 12m below sea level, will be safe?	2.0
2017-11-15T20:55:41.000Z	[0.6078699231147766]	base_embedding	Although he attributes this to the risk of a possible new embargo increase from the USA, the main problem is Akkuyu nuclei… https://t.co/DRDH373RS0	1.0

Table 2. Selected Example of Top Similar Words to Key Word "Russia"

Window_Start_Date	*Term_In_Window*	*Similarity*
9/4/2014	moscow	0.645961999893188
9/4/2014	nuclear_cooperation	0.440974622964859
9/4/2014	turkey	0.171396791934967
4/2/2015	turkey	0.414617717266082
10/9/2015	nuclear_cooperation	0.0758074671030044
1/27/2016	moscow	0.495317310094833



```
Parsing Arguments...
usage: main [-h] [-t KEY_TERM] [-p COMPARISON_TERM] [-c CO_TERMS] [-f] [-k NEAREST_NEIGHBOR_LIMIT]
            [-r NETWORK_RANK_LIMIT] [-s THRESHOLD] [--second_pass_exploration] [--model_path MODEL_PATH]
            [--tweet_data_path TWEET_DATA_PATH] [--result_output_path RESULT_OUTPUT_PATH]
            [--network_similarity_metrics NETWORK_SIMILARITY_METRICS] [--inflection_ranks_plot] [--network_metric_plot]
            Workflow Search_Terms Data

An application for extracting events of interest in the nuclear domain via NLP techniques.

positional arguments:
  Workflow              Select workflow to run. Selections available: "['similarity', 'inflection']"
  Search_Terms          List of terms to query against model. This must be a file path to text file containing queries,
                        or a list of terms in the following format: "['terms1', 'terms2', ..., 'terms_n']".
  Data                  Specify data to use as input. Currently supports 'twitter' and 'news' for use of respective
                        dataset.

optional arguments:
  -h, --help            show this help message and exit
  -t KEY_TERM, --key_term KEY_TERM
                        Key term to focus when analyzing similarities
  -p COMPARISON_TERM, --comparison_term COMPARISON_TERM
                        Term to compare similarities with key term in similarity workflow.
  -c CO_TERMS, --co_terms CO_TERMS
                        List of terms to check co-occurence with search terms. Co-occurence results won't generate if not
                        provided.
```

Figure 2. Snapshot of Inflector executable.

The semantic sentence-based search approach to extracting nuclear events from the Twitter dataset using curated guiding sentences yielded 5 hits when the similarity threshold (different from the threshold in infection pipeline) is set to 0.7, 2,332 hits when it's set to 0.6, and 73,909 hits when set to 0.5. These hits are out of the 3,390,048 total amount of tweets in the data. The strictness of the search is changed by the threshold provided. **Error! Reference source not found.** shows the extracted tweets according to a corresponding label that represents which sentence in pool of guiding sentences the tweet is most similar too. The Faiss index method of search relies on the number of closest neighbors to be extracted for each guiding sentences rather than a threshold. Faiss results are shown in **Error! Reference source not found.Error! Reference source not found.**.

The model that would be in charge of transforming sentences into their tagged counter-parts was only trained on the pool of guiding sentences mentioned earlier with their manually tagged versions according to IOB format. The pool consists only of 10 sentences. The results of entity transformation on the pool are shown in **Error! Reference source not found.**. The model predicts the tags in their IOB format and an algorithm combines respective tags to that entity spans are captured under one tag. The tag rulings used are as follows: ORG are organizations and companies, CNT are countries, LOC are locations, FLOC are foreign locations, NCNT are nuclear capable countries, PER are persons, and NUC encompasses nuclear miscellaneous terms. The performance of the model itself, however, is far from desirable. This is to be expected given the amount of data it was trained on. The tail-end of the internship was spent researching methods of automatic (and unsupervised) labeling (tagging) of data such that there is less work to do when manually labeling the data, or there is no need to even use the entity model since automatic labeling would have good enough accuracy. A promising method was found that used a spaCy English model in combination with gazetteers for context lookup and heuristics for rules or patterns in some entities (like grammar that traditionally makes up a person's name). This combination would be used to augment the data for training of the main model. Further work into this area is left as future work.

Table 3. Most Similar Sentences to Target in Pool of Guiding Sentences (or Corresponding Label)

Tweet_Index	*Translated_Tweet*	*Similarity_Score*	*Corresponding_Label*
197123	A nuclear agreement would boost foreign investment, which creates a solid base for improving diplomatic relations over years. @IMaSpiv	0.7080909609794617	[Country X nuclear capable country] and [country Y country with no little/nuclear capability] have been expanding their diplomatic relationship over the past year
1533404	even in China nuclear construction faces delays. Seems to be a key feature of nuclear: always delayed, always more…	0.7140467166900635	Head of [ABC Nuclear Power Co.] says that there have been significant delays in the construction of [nuclear reactor I] at [location Z] as a result of ongoing events in [location DEF]
1874326	Advanced nuclear announcements just keep on coming. Latest - X-Energy moves to conceptual design phase	0.721978485584259	[Entity X] has newly expressed interest in expanding nuclear power for electricity generation
2008308	Global Nexus Initiative calls for new approach to enable #nuclear power expansion	0.7061082720756531	[Entity X] has newly expressed interest in expanding nuclear power for electricity generation

2394345	“Internationally, relationships between the nuclear-weapon states have deteriorated, in particular between the US...	0.7104021310806274	[Country X nuclear capable country] and [country Y country with no little/nuclear capability] have been expanding their diplomatic relationship over the past year

Table 4. Tagging Results from Entity Model

Original_Sentence	*Token_Labels*	*Entity_Model_Output*
Vladimir Putin and Recep Tayyip Erdoğan have agreed to a deal to construct four nuclear reactors in Akkuyu, Mersin.	['[B-PER]', '[I-PER]', 'and', '[B-PER]', '[I-PER]', '[I-PER]', '[I-PER]', '[I-PER]', '[I-PER]', '[I-PER]', '[I-PER]', 'have', 'agreed', 'to', 'a', 'deal', 'to', 'construct', 'four', '[B-NUC]', '[I-NUC]', 'in', '[B-LOC]', '[I-LOC]', '[I-LOC]', ',', '[I-LOC]', '[I-LOC]', '.']	[PER] [PER] [PER] [PER] construct four nuclear reactors [NUC] [PER] [FLOC] [PER] [NUC] [NUC]
Russia and Turkey held diplomatic meetings in Akkuyu Tuesday to discuss Turkey's critical infrastructure.	['[B-CNT]', 'and', '[B-NCNT]', 'held', 'diplomatic', 'meetings', 'in', '[B-FLOC]', '[I-FLOC]', '[I-FLOC]', 'tuesday', 'to', 'discuss', '[B-NCNT]', '[I-NCNT]', '[I-NCNT]', 'critical', '[B-NUC]', '.']	[CNT] [CNT] turkey [NCNT] diplomatic meetings in [FLOC] [FLOC] turkey [NCNT] s [NCNT]
Egypt will fund the construction of four nuclear reactors in El Dabaa.	['[B-CNT]', 'will', 'fund', 'the', 'construction', 'of', 'four', '[B-NUC]', '[I-NUC]', 'in', '[B-LOC]', '[I-LOC]', '[I-LOC]', '[I-LOC]', '.']	egypt [CNT] fund the construction of four nuclear [NUC] [NUC] [LOC]
Rosatom takes 50% stake in nuclear reactors in Rostov.	['[B-ORG]', '[I-ORG]', 'takes', '50', '%', 'stake', 'in', '[B-NUC]', '[I-NUC]', 'in', '[B-LOC]', '[I-LOC]', '.']	[ORG] [ORG] [ORG] % stake in nuclear reactors [NUC] rostov .
Construction permit has been granted for Hungary to build reactors in Paks.	['construction', 'permit', 'has', 'been', 'granted', 'for', '[B-CNT]', 'to', 'build', '[B-NUC]', 'in', '[B-LOC]', '[I-LOC]', '.']	construction permit has been granted for hungary [CNT] build reactors [NUC] [LOC] [LOC]
Finland has newly expressed interest in expanding nuclear power for electricity generation	['[B-CNT]', 'has', 'newly', 'expressed', 'interest', 'in', 'expanding', '[B-LOC]', '[I-LOC]', 'for', '[B-LOC]', '[I-LOC]']	[CNT] [CNT] newly [NCNT] interest in expanding nuclear [NUC] electricity generation
President of the United States paid a visit to the site of Dungeness in England	['[B-PER]', 'of', 'the', '[B-CNT]', '[I-CNT]', 'paid', 'a', 'visit', 'to', 'the', '[B-NUC]', 'of', '[B-FLOC]', '[I-FLOC]', '[I-FLOC]', 'in', '[B-CNT]']	[CNT] [PER] the united [CNT] a visit to the site [NUC] [FLOC] [FLOC]
Russia and Turkey have been expanding their diplomatic relationship over the past year	['[B-CNT]', 'and', '[B-NCNT]', 'have', 'been', 'expanding', 'their', '[B-NUC]', '[I-NUC]', 'over', 'the', 'past', 'year']	[CNT] [CNT] turkey [NCNT] been expanding their diplomatic [NUC] the past year
Head of Rosatom says that there have been significant delays in the construction of nuclear reactor at Akkuyu as a result of ongoing events in Ukraine	['[B-PER]', 'of', '[B-ORG]', '[I-ORG]', 'says', 'that', 'there', 'have', 'been', 'significant', 'delays', 'in', 'the', 'construction', 'of', '[B-NUC]', '[I-NUC]', 'at', '[B-FLOC]', '[I-FLOC]', '[I-FLOC]', 'as', 'a', 'result', 'of', 'ongoing', 'events', 'in', '[B-NCNT]']	[ORG] [PER] [ORG] [ORG] there have been significant delays in the construction of nuclear reactor [NUC] as [FLOC] ongoing events in ukraine

Table 5. Nearest Neighbor to Each of the Guiding Sentences in Pool

Full_Date	*Guiding_Sentence*	*Tweet*	*Score*
2014-11-11T18:27:46.000Z	[Entity X] and [Entity Y] have agreed to a deal to construct four nuclear reactors in [Z location].	Russia Signs Reactor Deal With Iran: Russia said it signed a deal to build two new nuclear-reactor units in Ir...	0.7903071641921997
2016-04-01T14:00:25.000Z	[Entity X] and [Entity Y] held diplomatic meetings in [Z location] Tuesday to discuss [Entity Y's] critical infrastructure.	the impetus provided by high level discussions have contributed to enhanced bilateral cooperation in related issues	0.9016052484512329
2017-10-20T18:19:45.000Z	[Entity X] will fund the construction of four nuclear reactors in [location Y].	DOE Announces New Funding Opportunity to Support Advanced Nuclear Reactor Power Plants	0.7737705111503601
2016-12-22T14:51:43.000Z	[Entity X] will build, own, and operate four nuclear reactors in [location Y].	Tokyo, London Sign Memorandum of Understanding on Reactor Construction #News #Investing	0.8069419264793396
2015-05-17T21:13:00.000Z	[Foreign investment firm X] takes 50% stake in nuclear reactors in [location Y].	Nuclear insurance pool: Foreign firms interested to pitch in, says GIC - The Economic Times	0.8240773677825928
2016-04-28T16:50:46.000Z	Construction permit has been granted for [Entity X] to build reactors in [location Y].	Permit for a new nuclear reactor in N.J. OK'd by feds	0.7733869552612305
2015-06-25T02:06:19.000Z	[Entity X] has newly expressed interest in expanding nuclear power for electricity generation	A new look for nuclear power	0.6243684887886047
2016-06-28T05:55:54.000Z	President of [X location] paid a visit to the site of nuclear reactors in [Y foreign location]	@PhyllisCopeland Omar Mateen's company provided security to nuclear reactors. We're lucky he wasn't on assignment to a nuclear reactor.	0.8029593825340271
2018-01-18T18:38:42.000Z	[Country X nuclear capable country] and [country Y country with no little/nuclear capability] have been expanding their diplomatic relationship over the past year	@apipkin11 @SpeakerRyan Your idea of diplomacy is why he is such a nuclear power now. As far as China and Russia, t…	0.6642178893089294
2016-01-28T08:20:19.000Z	Head of [ABC Nuclear Power Co.] says that there have been significant delays in the construction of [nuclear reactor I] at [location Z] as a result of ongoing events in [location DEF]	Decision on new nuclear power plant 'delayed' - BBC News	0.629618227481842

4. CONCLUSION

With the goal of event extraction from the nuclear domain, two methods of extraction were pursued. One involved using word embeddings to build network graphs of closest neighbors to given key word queries, computing vector similarities and their derivations in time to highlight inflection points in key word semantic context. This pipeline has been packaged into an executable and modular platform which automates the workflow. The other method involved using sentence embeddings to compare guiding sentences to those in the data using vector similarity methods, producing a workflow that acts much like a search engine. To improve the amount of possible search hits in our data using this method, work was done to preprocess the sentences into their tagged versions so they would be more similar when the time came to compare vectors. To accomplish this, a supervised BERT model was developed to tag sentences according to IOB format; however, the model underperforms due to a lack of labeled (tagged) data. Future work therefore is needed for research and development of automatic labeling using unsupervised (or weakly supervised) methods.

5. REFERENCES

1. Danielson, Thomas L., & Deschaine, Larry M.. Natural Language Processing for Text Based Event Extraction: Identifying Events of Interest Related to Worldwide State-Sponsored Civil Nuclear Power. United States. https://doi.org/10.2172/1962589
2. Wang, W., Wei, F., Dong, L., Bao, H., Yang, N., & Zhou, M. (2020). MiniLM: Deep Self-Attention Distillation for Task-Agnostic Compression of Pre-Trained Transformers. ArXiv, abs/2002.10957.
3. Johnson, J., Douze, M., & Jégou, H. (2017). Billion-Scale Similarity Search with GPUs. IEEE Transactions on Big Data, 7, 535-547.
4. Sanh, V., Debut, L., Chaumond, J., & Wolf, T. (2019). DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. ArXiv, abs/1910.01108.
5. Hugging Face. (n.d.). Huggingface.co. sentence-transformers/all-MiniLM-L6-v2 https://huggingface.co/sentence-transformers/all-MiniLM-L6-v2