Data Science

Dual Use Nuclear Fuel Cycle Item Detection: NLP and Data-Driven Approach

Tyler Russin-Mirabito

Natural Language Processing (NLP) and data driven methods for detecting dual use nuclear fuel cycle items is an attractive option for regulatory agencies due to their ability to forgo the current manual process of detection in exchange for automated detection pipelines. However, it seems that NLP and data-driven approaches have yet to be shown as an effective solution to the problem. This research aims to explore possible solutions in this space and identify limitations within trade transaction datasets. Methods include the construction of an NLP pipeline that interprets trade data descriptions, as well as data driven approaches that attempt to identify patterns in transactions. The approach to utilize NLP as a substitute for the manual detection process of dual use nuclear fuel cycle items showed to be an effective one.

1 INTRODUCTION

Regulatory agencies are subject to the intense process of detecting dual use nuclear fuel cycle items in trade transaction data. The tracking of these “dual use” items, such as centrifuges and elemental Strontium, helps improve the safety and security of nations with respect to potential nuclear threats. The current method of detecting these transactions is manual. The process requires transliterate analysts with domain knowledge to scan individual documents in search of suspicious activities. Given the time and complexity of this process a better solution remains an open challenge.

Government branches, like The United States Government Accountability Office (GAO), have spent years implementing systems to prevent trades of dual use items through projects like the licensing program to legally transport dual use items (*GAO*, 1983). Efforts such as these have shown some efficacy by complicating logistics, and thus creating roadblocks for illegal operations. However, according to GAO investigations, the programs in place fall short in their effort and current approach, stating how easy it is to purchase dual use items from fronted companies with "fictitious identities" (GAO, 2009). Methods such as simple heuristics have been adopted by these analysts and agencies to create a firewall to further control dual use activities through separation of peaceful nuclear activities and nuclear threats (*Dalton*, 2017). But, they are still overwhelmed by the volume of data that requires processing.

Promising avenues to improve results in this domain exist and pertain to the realms of Natural Language Processing (NLP) and data driven methods to generalize and automate the detection process. NLP has the unique ability to process large and complex amounts of textual data, drawing useful and general information from it. Data driven models are designed to learn from features within data to reveal hidden patterns that can be used to make predictions on future datasets. Given this we ask ourselves; can NLP or data driven techniques be used effectively to isolate transactions involved in the dual use nuclear fuel cycle? The following research sets out to investigate this question.

2 TRAINING DATA

Data defines direction. Initial explorations were performed on a trade dataset with a length of sixty million rows and several hundred columns. Some of the features included in the dataset were, but not limited to Harmonized System (HS) Codes, freight description, weight, cost, and quantity. The data also contained identifiers for entities involved in the transactions, namely buyer, seller, transport, as well as banks. There were no features that gave clear information on whether a transaction was flagged as being involved in the dual use nuclear fuel cycle. Additionally the size of the dataset proved problematic, influencing how exploration and testing was conducted.

The data cleaning began with exporting randomly sampled subsets from Amazon S3 Buckets into python notebooks. Google Collab along with Amazon’s Sagemaker services were utilized in analysis, and modeling. Initial analysis of the dataset revealed much of the data was in different languages, many of the features were duplicated with slightly modified names, and for most features a substantial amount of the data was missing. Analysis showed that for 70% of features anyware from 30% - 90% of the data was missing, leaving roughly 30 features to be utilized for any kind of predictive modeling.

For features where reading the data was necessary for understanding, Google Translate and Yandex Translate services were used to translate data. Translation services use NLP to translate complex sentence structures which can lead to inaccurate translation, or important context to be lost. To compensate for this, translations were done on a word by word

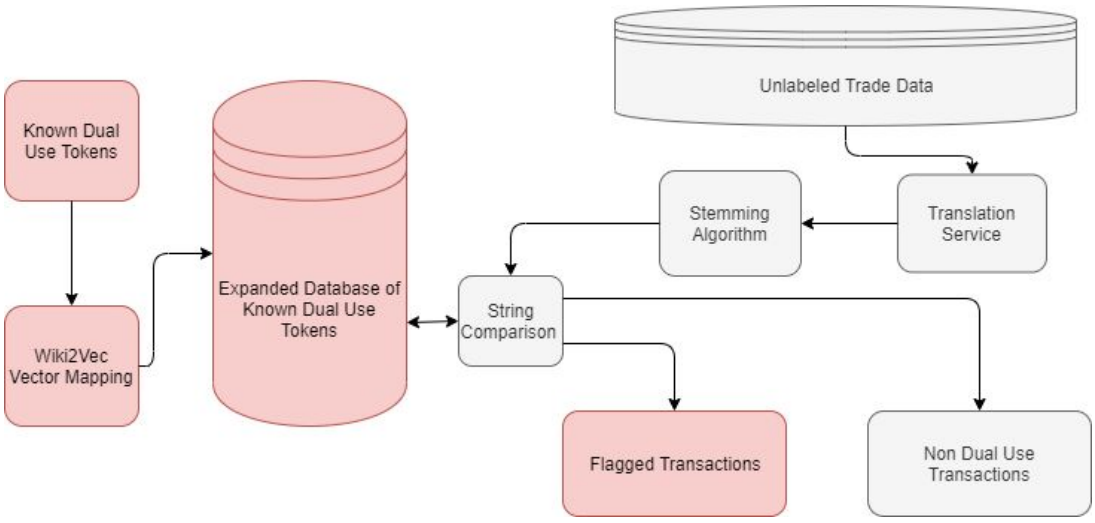

Figure 1: This diagram depicts our pipeline within the scope of processing transactions with NLP techniques. Two inputs go into the system: the known dual use tokens, and the unlabeled trade transaction data. The data then flows through the pipeline as shown in the visuzation and as described in section 3 of this paper.

basis. This proved effective given most of the sentence structure was systematic (e.g., “qty aluminum pipes: 74”) leading to a majority of useful translation.

Handling duplicate features was performed through column comparison using the Pandas library. If columns were identical one was deleted, if unique data existed in one column and null values in the other, the columns were merged. Missing data introduces bias and inaccuracies in the training process. Dealing with missing data was conducted through a process of evaluating the data feature’s potential usefulness to predictive modeling as well as setting a threshold for acceptable amounts of missing data.

The model had access to data containing HS Codes and a separate known archive of data on companies known to have been flagged as partaking in dual use nuclear fuel cycle transactions. Less than .001% of the data had nuclear HS Codes and less than 1% of the data had flagged companies. Of those flagged companies only a small fraction of their transactions included any nuclear activity. Also, given no features existed that gave clear information to whether a transaction was involved in the nuclear full cycle process, the team set out to manually label data. During this process most of the available features could not be used; things like weight and price did not provide enough information for manual labeling. We were limited to freight description and were forced to look for keywords within each description. A list of dual use item tokens was constructed and used as a reference point for comparison of freight descriptions. This method flagged 2% of the 60 million rows as dual use, around 1.3 million transactions. However, despite the success of the newly created data there existed an imbalance of more non-dual use transactions than dual use. Regardless, the model now has a base dataset that has been labeled such that dual use labeled transactions have a high chance of actually dual use association.

The initial data analysis and cleaning concluded with a standardized dataset consisting of 30 features. This work was critical to moving forward as the modeling tests would be split between several people that all need to compare results and findings.

3 NATURAL LANGUAGE PROCESSING PIPELINE

Within the dataset the freight description feature provided a rich amount of textual data for analysis and NLP. During the data exploration phase a simple string comparison method was implemented to compare tokens within the data against a list of known tokens to be associated with dual use transactions. This basic approach allowed for the successful capture of roughly 1.3 million transactions that had a high chance of dual use activity due to their association with key dual use tokens. This provided a direction to develop stronger more complex NLP pipelines to further process the data to a state where dual use transactions could be identified.

The list of dual use tokens initially contained 90 high quality values consisting of key machine parts and elements known to be used in nuclear reactors. This proved to be an effective approach. Building a larger list of relevant nuclear associated tokens became the clear priority. To do this, the wiki2vec word clustering algorithm was adopted and integrated into the heart of the pipeline.

Algorithm 1: The wiki2Vec model (**Φ**) was used. The model was trained on thousands of wikipedia articles, providing a large and complex vector mapping of words clustered based on similarly. The output token's value shown beside each token represents the token's proximity to the input token in vector space.

```
Function Vecotor_Mapping(token, threshold):
    /* Produce similar tokens */
    wiki2vec.most_common(token)

    /* Measure threshold % given token */
    For i in token_array:
        i >= threshold
    end
    Return token_array
end
Output:
    /* Clustered token, threshold % */
    [
     (<Word radiation>, 0.9999999),
     (<Word ionizing>, 0.920335),
     (<Word ionising>, 0.8600097),
      (<Word nonionizing>, 0.84354764),
     (<Word irradiation>, 0.8429501)
    ]
```

The algorithm makes use of context surrounding particular words, then uses that information to group words together that possess similar context. Experimentation with this model showed that similarity thresholds, when set in the range of 80% to 100%, provide a rich addition of dual use tokens. Tokens, such as “radiation”, when mapped by similarity with these thresholds, resulted in previously unidentified tokens, namely “radioisotope” and “irradiated”. Overall, when all the known dual use tokens were mapped, a list of roughly 360 high quality nuclear fuel cycle associated tokens was obtained.

Many different words can convey the same idea, as was shown through our word vector mapping strategy. However, as the known dual use token list expands the need to generalize words becomes increasingly apparent. Proposed is a non-traditional tactic of utilizing translation services like Google Translate and Yandex Translate to take in a variety of similar meaning words in one language and translate them to one general word in another language. This is a feasible approach given the technologies implemented by these translation services; namely, NLP, Machine Learning, and Deep Learning. Recent breakthroughs in these fields, and within companies such as Google, have resulted in this effective multilingual generalization translation tool (*Sommerlad*, 2018). In practice we find that the effectiveness of this method varies given the language of origin and the final language translation of the words. Given our dataset, this tool was vital in the success of the generalization of many different tokens of the same origin into a single token. It was found that in many cases a single dual use token in the list could exist as six varying tokens in the initial language.

Additionally, words can exist in a variety of states. A single word can exist as plural, singular, past tense, and so on; all while mostly meaning the same idea. In order to catch the majority of these tokens a method needed to be established to simplify a given token to its most base stem. This would increase the chances that a token would match a known dual use token in our list. This was done through the addition of the Snowball Stemmer Algorithm which aims to cut off ends of tokens to reveal their base state, turning tokens such as “irradiated” into “irradiate” and “enrichment” into “enrich”. This functionally worked well in capturing a large percentage of previously missed hits with only few downsides; some tokens were over-stemmed, like “hammer” to “ham”.

Overall, the use of the NLP strategies in combination with this specific trade dataset proved to be effective in identifying potential dual use transactions. The NLP aspect of the research was able to identify over 6 million transactions that had a high association to the dual use nuclear fuel cycle. Exploration into other data driven techniques was performed to further aid in the identification of these transactions.

4 DATA DRIVEN MODELS

Initially, the goal was given the dataset’s features including it’s dual use boolean feature. We set out to devise a model capable of predicting dual use transactions. However, given an unbalanced dataset (currently estimated at 98% non-dual use, 2% labeled dual use) the accuracy evaluation metric is not suitable.

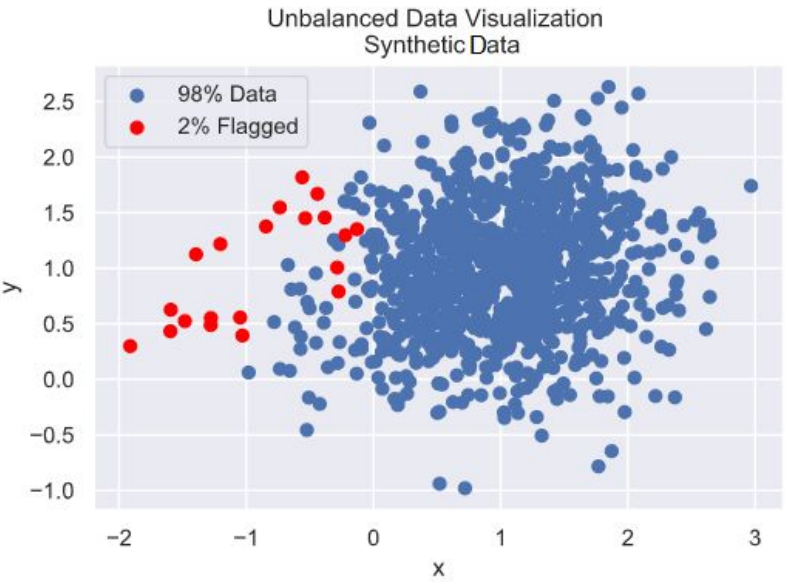

Figure 2: Represents the unbalanced state of our transaction data (Artificial subsample, true data is proprietary, x y values are arbitrary). The phenomenon in place is that given the volume of non-dual use transactions, a predictive algorithm, when evaluated by accuracy, will have a bias towards non-dual use as a prediction. This is because it will learn to be accurate 98% percent of the time if it only ever chooses non-dual use as a prediction. To the algorithm, predicting a transaction as dual is a "risk" that will result in a lower accuracy score.

Traditionally, in the given case a researcher may explore the metrics of a confusion matrix as well as ROC AUC scores. Unfortunately, given the nature of the labeling of the data, there was no certainty that a given transaction sample was labeled correctly. The method implemented for labeling only gave generalizations about the dataset as a whole and breaks down upon magnification. Regardless of these clear issues, experimentation and testing was still performed using Random Forest and XG Boost classifiers. The results were interpreted using non-traditional heuristics due to the clear flaws in the data. It was theorized that conditioning the model to aim for a high recall score and a precision score of around 50%, meaning around 8-10% of the data would be labeled as dual use. These thresholds were based on the structure of our labeling method. The results ultimately did not reach the threshold and a clustering analysis went on to show that the data consisting of weights, costs, etc. simply did not have the properties necessary to be used in a predictive algorithm.

5 CONCLUSION

NLP can effectively be used to isolate transactions involved in the dual use nuclear fuel cycle. This was observed through the successful implementations of NLP techniques; namely, string comparisons, word-vector mapping, deep learning in translation services, and stemming. The results of these processes provide a rich list of potential dual use transactions that can then be further investigated.

Given the trade dataset it was shown that data driven techniques cannot effectively be used to isolate transactions involved in the dual use nuclear fuel cycle. This is largely due to the poor quality of data within the dataset. Over half of all data were missing, and the available data provided no substance for the algorithms to make accurate predictions. In order for data driven techniques to be properly utilized in the detection of dual use nuclear items the data collection processes have to be altered.

Overall, implementing NLP methods on trade data provides marked improvement to how governments and regulatory agencies detect dual use transactions, giving them a much needed edge in the fight against nuclear threats.

HONORABLE MENTIONS

The research reflected in this paper was done by a team of four data science students at the Data Science, Lambda School in collaboration with the C4ADS organization. The scientists involved in this project by name are as follows: **Tyler Russin-Mirabito**, **Emma Rose**, **Jan Jaap de Jong**, and **Baisali Sant**. Notable contributions were also made by the project Team Lead, **Tim Dill**. Operations were done under the supervision of C4ADS's Program Director of Data & Technology, **Patrick Baine**.

REFERENCES

U.S. Government Accountability Office. (1983, October 13). Controlling Exports of Dual-Use Nuclear-Related Equipment. U.S. Government Accountability Office (U.S. GAO). https://www.gao.gov/products/NSIAD-83-28.

U.S. Government Accountability Office. (2009, June 4). Military and Dual-Use Technology: Covert Testing Shows Continuing Vulnerabilities of Domestic Sales for Illegal Export. U.S. Government Accountability Office (U.S. GAO). https://www.gao.gov/products/GAO-09-725T.

Toby Dalton, W. H. (2017, March 20). Toward a Nuclear Firewall: Bridging the NPT's Three Pillars. Carnegie Endowment for International Peace. https://carnegieendowment.org/2017/03/20/toward-nuclear-firewall-bridging-npt-s-three-pillars-pub-68300.

Sommerlad, J. (2018, June 19). The remarkable way Google Translate actually works. The Independent. https://www.independent.co.uk/life-style/gadgets-and-tech/news/google-translate-how-work-foreign-languages-interpreter-app-search-engine-a8406131.html.