SMU Data Science Review

Volume 5 | Number 1 Article 1

2021

Automated Analysis of RFPs using Natural Language Processing (NLP) for the Technology Domain

Sterling Beason
Southern Methodist University, sbeason@smu.edu

William Hinton
Southern Methodist University, whinton@smu.edu

Yousri A. Salamah
Southern Methodist University, ysalame98@yahoo.com

Jordan Salsman
Southern Methodist University, jsalsman@smu.edu

Automated Analysis of RFPs using Natural Language Processing (NLP) for the Technology Domain

Alex Salamah, Sterling Beason, William Hinton, Jordan Salman,
Masters of Science in Data Science, Southern Methodist University,
Dallas, TX 75275 USA
ysalamah@smu.edu , sbeason@smu.edu ,
whinton@mail.smu.edu, jsalsman@smu.edu
Acknowledgements: Dr. Kim Kukurbah Kkukurbah@gmail.com,
Amir Drusbosky Amir.Drusbosky@gmail.com for advisory and review

Abstract. Much progress has been made in text analysis, specifically within the statistical domain of Term Frequency (TF) and Inverse Document Frequency (IDF). However, there is much room for improvement especially within the area of discovering Emerging Trends. Emerging Trend Detection Systems (ETDS) depend on ingesting a collection of textual data and TF/IDF to identify new or up-trending topics within the Corpus. However, the tremendous rate of change and the amount of digital information presents a challenge that makes it almost impossible for a human expert to spot emerging trends without relying on an automated ETD system. Since the U.S. Government (USG), one of the largest purchasers of products and services, is using the Request for Proposals (RFP) to award contracts for various Information Technology and other services, this project will use Natural Language Processing (NLP) to mine the wealth of textual data that is embedded within the RFPs to develop an ETDS to identify emerging technology trends. The preliminary results show good promise that NLP may well be leveraged to mine text data and create an accurate ETDS.

1 Introduction

Albert Einstein said, “The measure of intelligence is the ability to change.” However, nowadays change is coming from all directions! Laws, regulations, technology, and social norms are all sources of change. So, detecting and adapting to change is exceedingly difficult. Organizations, especially large ones, have the most difficulties adapting to change. For example, Blockbuster Video once considered a giant in the home movie and video game rental business, had to file for bankruptcy in 2010. Blockbuster was arguably one among the most iconic brands within the family entertainment domain. In 2004, the company employed 84,000 employees and owned about 9,000 stores around the world (Goh, 2018). Blockbuster’s lack of insight and vision to detect the shift toward online digital streaming model caused its demise (Goh, 2018). On the other hand, Netflix a company that offered to sell itself to Blockbuster for only $50 million now has a 167 million subscribers and recorded $20 Billion in revenue. (Netflix Revenue and Usage Statistics, 2020).

Change is hard and takes a tremendous number of resources to implement. Consequently, the more lead time a company has in detecting change, the better their

chances are to implement adaptation. An Emerging Trend is an incongruent term. This is because a trend, to be precise, is a change that occurred in the past, and it happened over time. A trend describes history. In the example about Netflix above, the growing number of subscribers over the years is an example of a trend (Trends vs. Emerging Issues: What Is the Difference? 2016). Humans innately tend to use the past to try to extrapolate into the future, and thus people often think of trends as if they highlight the future when they are not!

Emergence or Emerging Domain, on the other hand, could be potential new technology that would revolutionize an existing domain or industry. For instance, completely autonomous vehicles might be considered an emerging domain in the vehicle manufacturing industry. It could even be a new concept or idea or social issue that, while it is currently considered eccentric, could become mainstream in the future. Whereas trends explore the past using collected data points, Emerging Domain will allow us to discover new things that may become meaningful in the future.

As a result, awareness of Emerging Domains is invaluable to companies because it will provide them with the ability to answer a series of strategic questions such as:

- Does the Emerging Domain constitute a threat to the enterprise?
- Does the corporation possess the suitable resources to adjust to the Emerging Domain if needed?
- Should the company combat or promote this Emerging Domain?
- Does the enterprise need to invest or change policies, corporate cultures, or the mix of its workforce as a result?

Nowhere is Emerging Domain more evident and rapid than within the domain of technology. There, both the volume and the velocity of change are continually accelerating to the point that no human expert can keep up.

Obtaining an RFP contract from the government could be a lengthy and rigorous process. The government proposals are allocated through different contract vehicles. For the Federal Information Technology (IT) domain, the USG relies on three types of contracts A.K.A contracts vehicle. These include the following:

- GSA Schedule: These types of contracts are managed by the General Services Administration. They are used to publish a price list for several types of products that USG customers can buy from.
- An ID/IQ vehicle is used to undertake the development of major defense systems to a large defense contractor known as “systems integrator”. Due to the elevated level of complexity of requirements a simple scheduled solicitation cannot be used. For this type of contract type, there usually is a lead company that is referred to as “prime”. The prime provide will acquire software from other technology firms, known as subs, or from their partners to integrate into the mission system that they are developing for the USG.
- A government-wide acquisition contract (GWAC) is a competition for a General Services Administration (GSA) schedule. GWAC are a form of contracts that even though they are administered by a single USG agency, they are often deployed across the entire government. The objective of these contracts is to create a short or down-selected list of provides that will meet the mandatory requirements for the product or service.

To help alleviate a number of these issues and to assist in reducing the number of systems used in the government acquisitions process, the GSA downsized and reduced the number of systems that it used for the federal acquisition and awards processes from ten online websites into one.[1] Data from the original websites, including entity registrations and historical contract data, will be migrated into the new website database. This newly consolidated website is now the official U.S. Government system for people to bid on, receive information about, and manage federal awards. The following image illustrates this migration.

Figure 1: This figure illustrates the consolidation of GSA systems into Beta.SAM.GOV

This study focuses on working with the IT vehicle dealing with a specific market for the Department of Defense (DoD) and Department of Homeland Security (DHS). This project will ingest data from the above government procurement sites, analyze ingested RFPs using NLP and look for emerging technologies that are being requested most often. Once we have this data, then this will enable a wide array of analysis. For example, using graph stores such as Neo4J or Amazon Neptune will establish lineage between technologies and IT companies that have high win rate. This lineage can then be used to perform market and competitor analysis on these companies thus enabling the formulation of strategies that are based on data that have been harvested from actual demand.

However, extracting the data out of the RFP is a complex process. This is because the RFPs are not written in a standard format, and the information contained in these documents is highly dependent on the person who is writing the RFP. Additionally,

[1] https://beta.sam.gov/ .

since RFPs are managed by different branches of the USG, this results in a lack of consistency in RFP structure. Furthermore, the USG may intentionally obfuscate specific technologies being requested in order to protect the national interest. Finally, the use of abbreviations, acronyms, and synonyms of vendors' names to describe technologies is another challenge that this project will have to overcome.

These are some of the factors that this project may have to deal with as part of the data cleansing or preparation process. Taxonomies, ontologies, and data dictionaries may need to be developed and be continually updated with new abbreviations and their definitions. Methods, such as data tagging, may need to be used so the model may properly identify technology words correctly.

2 Literature Review

The bulk of the analysis will be using Natural Language Processing (NLP) to process Request for Proposal (RFP) documents.

Meijer et al. (2014) an Automatic Taxonomy Construction from Text (ATCT) extractor that created a new taxonomy from an existing corpus. The ATCT extracted the most relevant terms for specific domains. They found that their approach worked well in more specific domains rather than ambiguous domains.

Taking the automated taxonomy creation, a step further, Hoxha, et al. (2016) introduced an unsupervised clustering technique that was able to learn from and utilize the background knowledge of a corpus' domain. This technique differed from Meijer, et al. (2014) in that there was an understanding of the domain topic of the corpus before the annotation. Hoxha et al. (2016) were successful in creating a method that found 95.75% of the concepts. The biggest gap that is present in both the automated taxonomy methods is that of ambiguous, more generalized domains. For instance, technology, the domain of interest in this research, is extremely ambiguous. For this reason, previous automated taxonomy approaches are insufficient in creating a specific and practical taxonomy.

Another approach to corpus comprehension is in keyword or key phrase extraction. Keyword extraction is not dependent on having an existing taxonomy but may be strengthened by the existence of one. Many researchers review literature to identify and extract keywords from them. This is used to gain a higher-level understanding of the corpus. Nesi et al. (2015) and Dief et al. (2017) both leveraged NLP to process substantial amounts of documents to pull out relevant keywords and phrases. Nesi et al. (2015) utilized the Hadoop data platform to build their own keyword extractor. The authors were able to process documents quicker than previous attempts. They were able to extract nine million keywords from 20,000 documents in less than 2 hours when using five nodes to cluster. The previous attempts on this data took 115 hours to extract keywords. Dief et al. (2017) had a similar approach and relied on machine learning techniques for keyword extraction. They also found that by finding similarity measures between words and then clustering on those similarities, they could speed up the extraction of the keywords. Both researchers were similarly successful in their ultimate goals of extracting key phrases from enormous amounts of documents.

Another great challenge in understanding text corpus is of data that is not clean. Documents that are not consistent throughout the corpus and may even contain missing sections. Handling this data appropriately is paramount for any research dealing with this issue. Souili, Cavallucci, & Rousselot (2015) dealt with missing, non-consistent text by filling in the blanks using an NLP process known as Natural Language Generation. They first had to do the same pre-processing techniques as others such as POS tagging, and tokenization. They created a machine learning algorithm that took the context of the sentence and predicted what was in the missing blank. Their modeling filled in the blanks for partial solutions within patents.

Alshemali and Kalita (2020) continued to emphasize the importance of clean data within modeling. They worked with Deep Neural Networks to generate language in text. Alshemali and Kalita used both clean and perturbed text within their DNN. They found that DNN can be easily compromised with unclean data. To combat this DNN's need supplemental components. One such possible component mentioned in this research is a network add-on that can decipher perturbed text. With the neural network no longer compromised they were able to recognize trends and categorize text. They were successful in showing the importance of clean text and the handling of compromised text.

Acronyms are another obstacle within NLP. Pustejovsky et al. (2001) present an innovative approach for detecting abbreviations. The solution leverages hand-built regular expressions and syntactic information to spot boundaries of noun phrases. When a short-form immediately follows a noun phrase, then the characters of the short-form are matched to the long-form. A ratio of the number of non-stop words in the long-form to the number of characters in the short form is calculated and the match is accepted only when the result is below a tolerance level of 1.5. This algorithm achieved 72% recall and 98% on "the gold standard," a small, publicly available evaluation corpus that this group created.

Building off existing taxonomies, keywords, and clean data, another researcher can achieve document classification and summarization. Hassan and Le (2020) used NLP on construction contracts to extract the essential requirements from the contracts. They used data pre-processing techniques such as stop word removal, tokenization, and parts of speech tagging to gain insight into the association of words in the contracts. Hassan and Le (2020) further used classification techniques such as Naive Bayes, Support Vector Machines, and Logistic Regression to try to identify the problem statements within the documents. They found that their SVM algorithm was successful at a rate of 95%.

Georgescu (2020) took text mining and text classification in a slightly different route while classifying the entity relationships with cybersecurity documents. Georgescu also used the same pre-processing techniques such as parts of speech tagging and stop word removal. Georgescu was able to build a machine learning model that produced computer-generated annotations that were like human-generated annotations. Georgescu was able to identify the relation type, the parent entity, and the child entity of the documents.

Additionally, Chang et al. produced an algorithm that used linear regression on a pre-selected set of features, achieving 80% precision at a recall level of 83%, and 95% precision at 75% recall on the same evaluation collection (this increases to 82% recall and 99% precision on a corrected version). Their algorithm used dynamic programming

to find potential alignments between short and long form and used the results of this to compute feature vectors for correctly identifying definitions. They then used binary logistic regression to train a classifier on 1000 candidate pairs.

Cheng and Yu (2019) and Sarhan, Spruit (2020) both leveraged transfer learning in their analysis. Transfer learning is an unsupervised type of learning that relies on past results to gain insights into new and untrained data. Sarhan, and Spruit used Open Information Extraction (OIE) to extract relationship tuples from unstructured documents. The OIE was used as the source task to transfer to other NLP tasks. They created a Recurrent Neural Network Model that used the OIE task as a sequence labeling problem. They were able to achieve better results with their labeling using the RNN model than previous research.

Kanghoong et al. (2018) considered reinforcement learning on request for proposal documents. Reinforcement learning is a part of machine learning that focuses on maximizing the cumulative award of the algorithm. They considered the optimal solution to the problem at hand and decided not to restrict cases. They chose to use an algorithm called "BOKLE", a simple algorithm that uses Kullback-Leibler divergence to constrain the set of models. They used NLP and CYK algorithms to categorize the trends and produced a viable solution in predicting the next trend for the RFP.

Kontostathis et al. (2004) focused their work on Emerging Trends Detection Systems (ETDS). They utilized a combination of semi-automatic and fully automatic ETD systems that included linguistic and statistical feature learning algorithms. Their research indicates that much progress has been made toward automating the process of detecting emerging trends but there remains room for improvements. A human with domain expertise is needed to separate emerging trends from noise in the system. As a result, projects that focus on creating effective processes to both semi and fully automate emerging trend detection can and should continue.

This research will build off the previous research referenced above. The techniques used such as machine learning, transfer learning, text annotation, and others will be utilized for the purpose of extracting technological keywords from request for proposal documents. Methods referenced above will be adapted for the use of this research.

3 Data

The data utilized in this project was sourced from RFPs that were published online within the System for Awards Management (SAM). Thanks to existing transparency and accountability regulations of the USG's acquisition process these data are made available to the public and for contractors to respond to.

This project targeted a corpus of RFP documents for the Department of Defense (DoD) and the Department of Homeland Security (DHS). These documents, on average, are between 50-75 pages of mostly natural language mixed with data that could be found in tables. There are some boilerplate pages that can be ignored like the first two pages that act as input forms, common contract acronyms descriptions, or legal disclaimers. These documents have all the information needed by a contractor to respond adequately.

Extracting the RFPs from the SAM website using an automated process proved to be challenging due to limitations imposed on the number of transactions per day as discussed in greater detail in section 4.2 below. To avoid delays in obtaining the data, a data extract that provides a list of all the RFPs posted in 2020 was obtained in the form of MS Excel.

The data extract was then filtered based on the North American Industry Classification System (NAICS) attribute. NAICS is a system that was developed for use by Federal Statistical Agencies for the collection, analysis, and publication of statistical data related to the US Economy. It is used to codify the types of industries that contribute to the US economy. The NAICS code for the Custom computing and IT services is defined to be 541 511.

Additionally, RFPs pertaining to the DOD and DHS were selected by filtering the dataset on Department/Agency attribute. Once a subset of data were defined, then solicitation numbers were randomly selected, and their corresponding RFPs were downloaded manually. Most RFPs were in PDF format, but there were other formats such as Word and Web-forms as well. Some RFP documents were scanned images of the original copies. Those were omitted as the project decided to focus on machine-generated documents to reduce the effort of reviewing the accuracy of an Optical Character Recognition (OCR) pipeline's output.

The following Figure illustrates the data flow and processing:

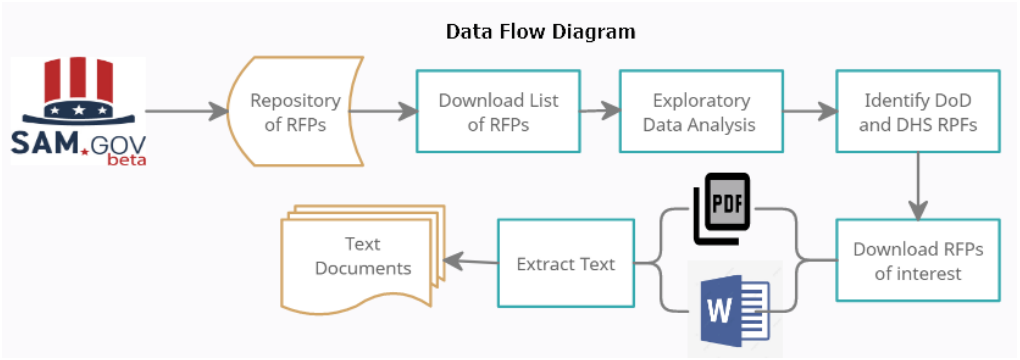


Figure 2: This figure illustrates the data flow out of GSA systems into Beta.SAM.GOV and the text extraction from the DOD and DHS RFPs

4 Methods

Using modern Natural Language Processing (NLP) techniques, the goal is to analyze a corpus of common solicitation documents, called Request for Proposals (RFPs), used by various government organizations intended for the US defense contracting industry audience. Using these techniques, it would enable us to automatically gain valuable information that led us to further relevant insights. Context vectors coupled with

identified entities allows us to distinguish ambiguous vendors by specific product lines. Extract crucial requirements to map to similar solicitations and their award adjudications. Pairing NLP and graph representations of data will allow for enhanced insights across the vast network of customers and solicitations.

Modern NLP revolves around the idea of vectorization of words, popularized by an algorithm called word2vec. The vectors together create a layer within a deep learning model called the embedding layer, which is an index the algorithm uses for each word in a sentence or sequence of word tokens. The dimensions of the vectors within the embedding layer vary, but it is common to have 300 to 600 dimensions vectors. Organizations with large resources such as Google, Facebook, and universities open-source their embedding layers and deep learning models that have been trained on sometimes billions of instances of text data sourced from places like Wikipedia or decades of newspapers. This is an expensive process and a barrier to entry for most organizations with tighter budgets. This is a big reason transfer learning is so attractive, as you do not have to spend a lot of resources to build a model from scratch. The pre-trained models are there to be leveraged for their strong feature extraction and pattern understanding that the top-most layers possess.

This research will use open-source models to leverage the context rich embedding layers and a few subsequent layers, while training the ending classification layers to this research's specific task. This method is known as transfer learning or fine-tuning. To fine-tune these models for a new task, this research requires new training data for the problem statement, which are RFPs.

To obtain the data there are a couple of approaches that could be followed:

4.1 Manual Download

This can be achieved as follows:

1. Obtain an account on the government centralized, authoritative source of federal award data website: https://beta.sam.gov/
2. The account is then used to navigate to the data bank where list of RFPs is archived and stored.
3. The list of RFPs contains a variety of records of various types. This study will focus on the solicitation and award types.
4. Using the URL found on each of the records, the RFP document could then be downloaded from the beta.sam.gov website.

4.2 Using a Webservice/API

An alternate method would be to use SAM's official Application Programming Interface (API) to obtain the RFP documents. However, non-registered entities, i.e., non-official business, are limited to only ten queries per day. This is very limiting and as a result, this method was not used as it would have been exceedingly difficult to collect enough data and the process would be excruciatingly slow. So, for this project, a manual download of the RFPs will be used.

5 Context Architecture

The following diagram represents the notional context architecture for this research.

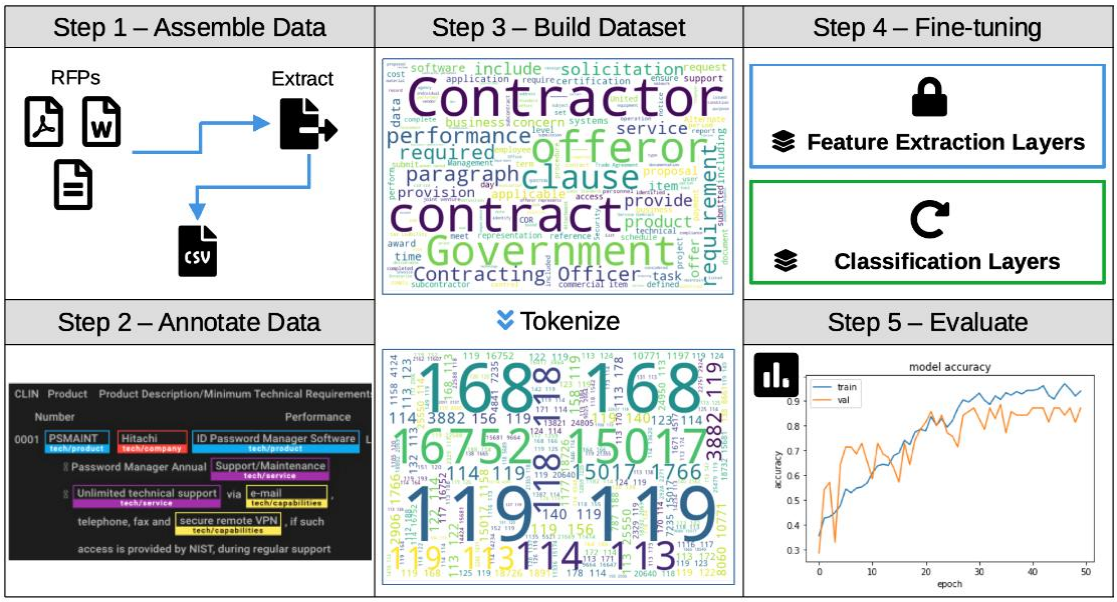
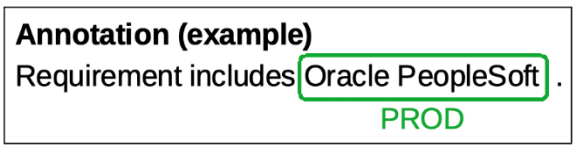


Fig. 3. Context Architecture illustrating the steps this paper will follow.

5.1 Data Annotation approach

With all the text extracted from the RFP documents, which are made up of various file formats, annotations are needed to build a dataset. Each document's contents must be addressed by a subject matter expert or an annotator with a detailed annotation guide. The guide outlines the annotation task objectives, classification options, and examples of good annotations. An open-source text annotation tool, Doccano, was used to facilitate this process. It was hosted on an Amazon Web Services (AWS) EC2 instance and managed as a container with Docker to allow input from various remote annotators. These annotations become the foundation of building a dataset that can be processed by an NLP model to fine-tune a pre-trained network's classification layers that will produce better or new classifications specifically for a task.



The result of the annotations are sequences of words for each document with a list of tags that provide the position of a tag in the sequence, its range, and its label. Further processing is required to create a final label sequence that converts the tags to an Inside-Outside-Beginning (IOB) format. This format allows the classification of multiple words as a single entity.

Fig. 4. Illustration of the Inside-Outside-Beginning (IOB) format.

In figure 3 above, the entire PROD (product) entity is “Oracle PeopleSoft”. The tag for “Oracle” is the start of the entity, so its label gets a “B-” prefix. “PeopleSoft” is not the start, so its label gets an “I-” prefix. All other words are out-of-scope or considered ‘outside’ following the IOB format. There are also other commonly used formats such as BILUO that is used by SpaCy.

5.2 Sequence Tokenization

Before the sequences of words and tags can be an input to the models, they must become numerical. These numerical representations of the words are known as tokens. Not all words will have a respective token in the model, as there are often limits on the embedding layers size. The embedding layer can be thought of as a lookup for a token’s n-dimensional vector representation that is the true input under-the-hood. For example, if there are 3000 unique words in a corpus and the embedding layer size is 1500, then the max number of unique tokens in the corpus will be 1500. Words that do not have a unique token in the embedding are replaced with a shared token value.

A token that was tagged previously in the annotation step will have a corresponding label value in the label sequence. Tokens that were not tagged will share an out-of-scope label or ‘O’ for outside, as all tokens will be classified for this task. In the training process, the NLP model will compare each classified token in the sequence with its known label. The error in this process will become a coefficient in the model's optimization of network weights.

5.3 Transfer Learning

This paper proposes using pretrained models that have been exposed to billions of natural language examples, while fine-tuning with this paper’s targeted corpus, Department of Defense RFP/RFQs. Using state-of-the-art NLP transformers that are provided by Hugging Face’s Transformers library and SpaCy, this paper will extract

the following meaningful data: named entities like technology companies, software packages, and applications. This task is known as Named Entity Recognition (NER).

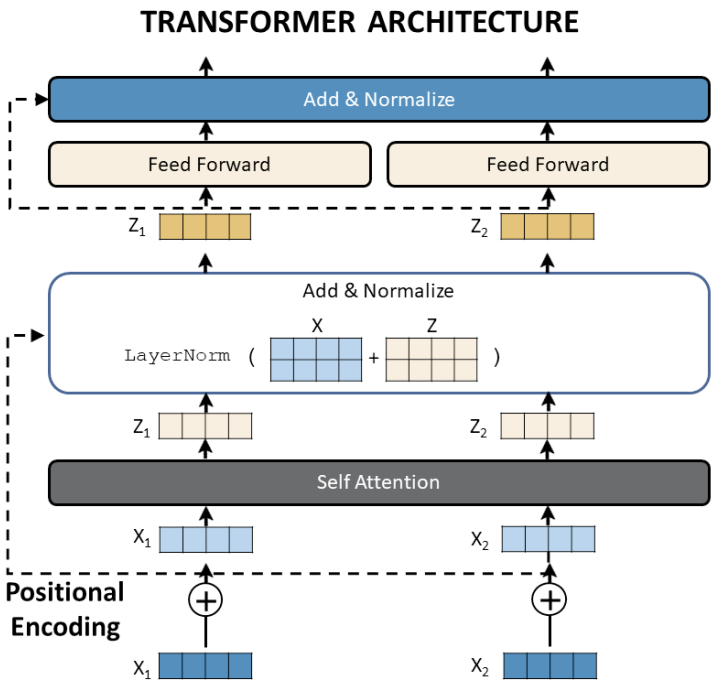


Fig. 5. Transformer architecture (https://www.thoughttrace.com/blog/why-bert/)

With the proposed transformer architecture, it goes beyond simple traditional methods like string matching and word frequencies, because it continuously keeps a contextual vector in memory for greater granularity of tokens (words) within a sequence.

As is common with fine-tuning of pre-trained models, starting layers within the model are frozen or made static, while the ending classification layers are trained with the specific corpus for a use case that targets the problem described in this paper. This way the knowledge of those billions of natural language sequences is preserved and still leveraged in the pursuit of this narrower task.

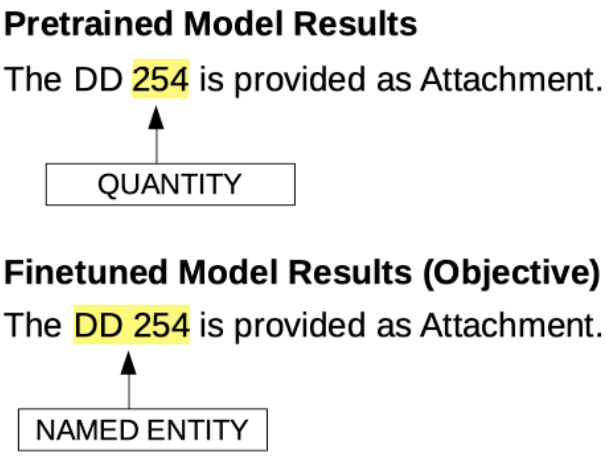


Fig 6. An Example of this Paper's objective. Fine tune pre-trained model to increase accuracy

The Department of Defense form 254 "Contract Security" (DD 254), is a common reference to a form distributed with government contracts when a requirement dictates its existence; however, a pretrained model is more generic and will not have the granularity to identify it as a complete entity or for other edge cases. This is where fine-tuning can enhance the model's ability to correctly identify and label the entities that are critical to capture for the audience of these documents. This involves custom annotations of the sequences the model will be finetuned on and the classification layers will learn to weight correctly. Correctly identifying these artifacts can enhance an organization's ability to correctly digest solicitations and focus on preparing themselves for shifting customer expectations and requirements.

6 Results

This research demonstrated the ability to extract domain-specific insight into documents that previously required extensive analysis by a human resource. While this research concluded with promising results, this paper's solution is best suited to augment an analyst's dissection of RFPs to enhance a quicker understanding of the key elements within those documents. Suffice to say, a human in-the-loop is required for any meaningful action to result like response proposals.

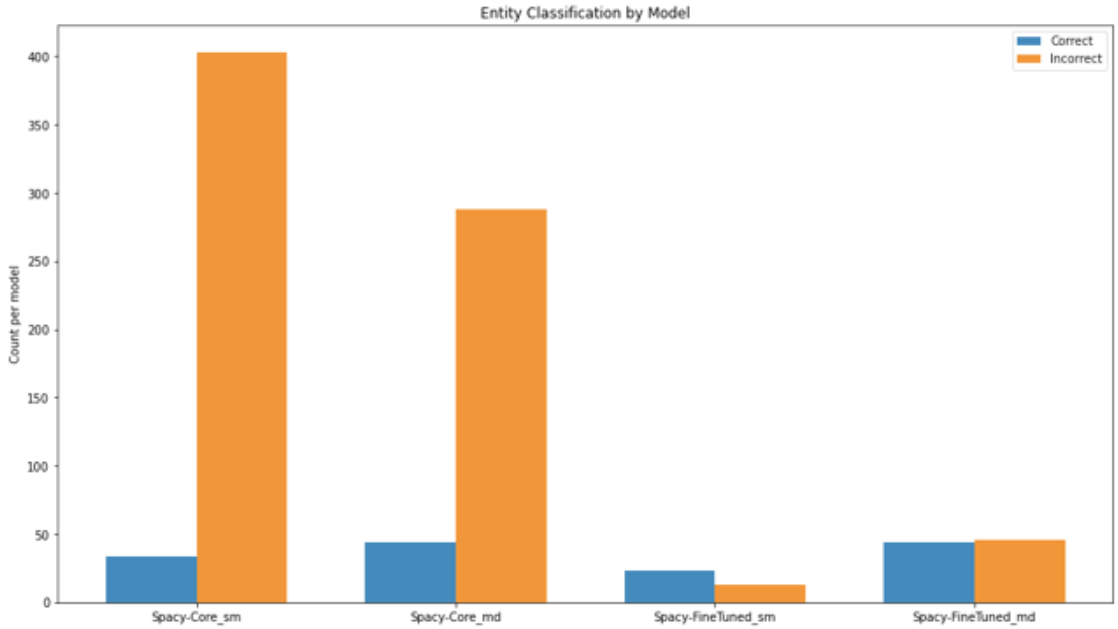


Fig. 6. SpaCy Model Results. The first two sets of bars are based on SpaCy core models and the last two sets reflect results from fine-tuned SpaCy model. Blue and Orange represents correct and incorrect classification, respectively.

The above bar chart figure shows test case classifications from the base SpaCy models on the left, and the fine-tuned SpaCy models on the right. All the fine-tuned models were trained using the custom technology taxonomy formulated for this research. The results clearly show an increase in precision with the trained models.

The small SpaCy fine-tuned model was able to identify and label more correct technological terms than incorrect ones. While the medium SpaCy model was able to achieve almost 50% correct identification and labeling. The advantage of the medium over the small model is that the medium was able to identify more technology terms overall that went undetected in the small model, in essence, the medium model was more thorough and less precise.

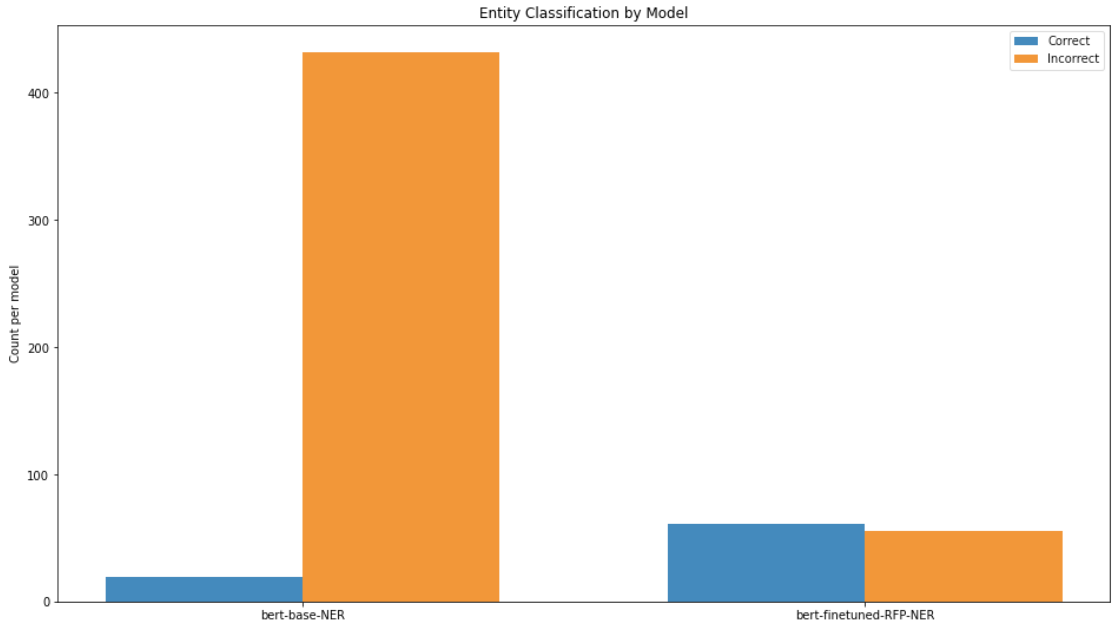


Fig. 7. Hugging Face Model Results. The first two bars are based on Bert pre-trained NER model while the last two reflect results from fine-tuned Bert model. Blue and Orange represents correct and incorrect classification, respectively.

Like the SpaCy results, a commonly used pre-trained Bert NER model provided a limited number of correct classifications and a greater number of incorrect classifications. The precision on the test cases were 0.042.

Starting from the pre-trained base Bert model and fine-tuning on this project's NER task, the incorrect classifications decreased. The correct classifications more than doubled to bring the fine-tuned model's precision to 0.526.

Table 1. Target metric, precision, by model

Model	Short Description	Precision
SpaCy-core-sm	SpaCy Core Small (generic)	0.076
SpaCy-core-md	SpaCy Core Medium (generic)	0.140
SpaCy-finetuned-sm	SpaCy Fine-tuned Small **(RFP)**	**0.675**
SpaCy-finetuned-md	SpaCy Fine-tuned Medium **(RFP)**	**0.495**
Bert-Base-NER	Bert Base NER (generic)	0.042
Bert-finetuned-RFP-NER	Bert Base Fine-tuned NER **(RFP)**	**0.526**

7 Discussion

This project used NLP to analyze Request for Proposals (RFPs) to identify Emerging Trends in the technology domain. A more mature version of this project could prove to be an unbelievably valuable tool for companies that are especially in the fast-paced technology domain. Companies that use this type of technology, may gain strategic advantages over the competitors by improving their ability to answer a series of strategic questions such as:

- Is there a shift in the marketplace driven by an emerging domain that may constitute a threat to the enterprise?
- Do we have the need resources, technological and human, to compete or beat our competitors?
- Where should we direct our monetary investment to maximize the largest ROI?

As reflected in the results section, this project illustrated the use of NLP as a potentially effective approach to identify new technology being requested as requirements by the USG. Using the fine-tuned models to extract, collect technology identities, and given enough historical RFP data; this will help identify emerging technologies.

Identifying and retrieving USG RFPs proved to be a challenge. This is due to several factors:

- Difficulties obtaining access to the RFP database
- Multiple repositories for various USG branches
- Inconsistent format of the RFP documents.

These factors along with others necessitated that data for this project be collected and processed manually. This is a laborious and tedious process that limited the data sample which potentially affected the results.

Finally, the paper highlights potential ethical questions that are discussed in the Ethics section below.

7.1 Ethics

One of the greatest ethical dilemmas in this research and in future research is in the usage of data. This research used publicly available data provided by the United States' government. All parties involved in the RFP process were made aware that this process was public and consented to such. Future research that stems from this would also need to ensure that all guidelines and regulations are being followed and the data itself is appropriate for use.

Regarding specific usage of this research, this research is meant to gain a better insight into the RFP marketplace and facilitate the RFP process. However, it is not to be manipulated into larger corporations forcing smaller corporations out by bidding on proposals they do not plan to utilize merely to cripple competition. It is also not meant for purposefully bogging down the RFP process and tying up proposals.

7.2 Future

Future application of NLP to discover patterns and extrapolate prediction from unstructured data is abound. Particularly one could envision some of the following applications:

- Identify Potential or Emerging competitors by Searching GSA compliant vendors' websites IT or Forum to see what they are commenting on. Then use this list on upcoming GSA projects to predict who might be the competitor in the future.
- Utilize user reviews data from vendor RFP and predict who might be the top vendors who will win GSA government projects over the course of a span of time. This might open business or partnership opportunities with the companies that are predicted to win these projects.
- Investigate potential emerging technology and see what projects are using these technologies today to predict what region of USA mostly uses this technology and types of government agencies.
- Technologies identified within RFPs could be cross referenced to companies' Human Resource skills to analyze the companies' ability to deliver based on the requirements of the USG. Also, companies could use this data to grow the skills of their workforce and augment it with the appropriate resources.

8 Conclusion

This project utilized NLP and transfer learning to identify technologies and technology products embedded in USG purchasing requirements. Data were downloaded and harvested manually due to the limiting factors listed above. Regular expression was used to extract the sections containing technology requirements. As illustrated in the results section, the fine-tuned models were able to identify technologies with greater accuracy and precision than the more generic pre-trained NER models; however, the extracted text had to be manually tagged and labeled to reach these results.

Additionally, based on the results, NLP has shown to be effective at extracting technology entities from complex documents such as USG RFPs. This is significant as it opens a path to a more streamlined and automated analysis of large RFP corpuses to assist in identifying emerging technologies within requirements. This predictive insight could be the difference between being a leader or falling behind and risk being crushed by the competitors.

Similar to this project, NLP could be further applied to other areas of text analytics. For example, NLP could be used to analyze how well do proposals meet the requirements defined in the RFPs, and then provide a prediction on possible win. Another possibility is to apply NLP summarization coupled with other relevant information extraction. This enables businesses to quickly determine if they are a good fit without expending their resources.

This technology might result in bigger companies that have better and more mature data scientist teams to have unfair advantages over their smaller competitors resulting in ethical considerations that need to be evaluated.

References

1. Alshemali, Basemah; Kalita, Jugal Elsevier B.V: Improving the Reliability of Deep Network in NLP: A Review. *Knowledge-Based Systems*, *191(2020)*, 105210–. https://doi.org/10.1016/j.knosys.2019.105210
2. Anderson, D. (2013). A holistic compliance model for capture teams: A grounded theory approach. ProQuest Dissertations Publishing.
3. Calahorra-Jimenez, M., Molenaar, K., Torres-Machi, C., Chamorro, A., & Alarcón, L. (2020). Structured Approach for Best-Value Evaluation Criteria: US Design–Build Highway Procurement. Journal of Management in Engineering, 36(6), 4020086–. https://doi.org/10.1061/(ASCE)ME.1943-5479.0000857
4. Cheng, Lefeng: Yu, Tao: A new generation of AI: A review and perspective on machine learning technologies to smart energy and electric power systems, Intenational journal of energy research 2019-05, Vol.42(6), p.1928-1973
5. Cropper, Andrew; Muggleton, Stephen H New York: Springer US: Learning efficient logic programs: Machine Learning, 2019-07-15, Vol. 108(7), p.1063--1083
6. Dief, N., Al-Desouky, A., Eldin, A., & El-Said, A. (2017). An Adaptive Semantic Descriptive Model for Multi-Document Representation to Enhance Generic Summarization. International Journal of Software Engineering and Knowledge Engineering, 27(1), 23–48. https://doi.org/10.1142/S0218194017500024
7. Georgescu, T. (2020). Natural Language Processing Model for Automatic Analysis of Cybersecurity-Related Documents. Symmetry (Basel), 12(3), 354–. https://doi.org/10.3390/sym12030354
8. Goh, 10 Companies That Failed to Innovate, Resulting in Business Failure. https://www.collectivecampus.io/blog/10-companies-that-were-too-slow-to-respond-to-change
9. Hassan, F., & Le, T. (2020). Automated Requirements Identification from Construction Contract Documents Using Natural Language Processing. Journal of Legal Affairs and Dispute Resolution in Engineering and Construction, 12(2), 4520009–. https://doi.org/10.1061/(ASCE)LA.1943-4170.0000379
10. J. Pustejovsky et al. "Automation Extraction of Acronym-Meaning Pairs from Medline Databases" Medinfo 2001;10 (Pt 1): 371-375
11. J.T. Chang, S. (2002). Creain H Schütze, and R.B. Altman, "Creating an Online Dictionary of Abbreviations from MEDLINE" JAMIA, 9(6), 612-620. https://doi.org/10.1197/jamia.M1139
12. Kontostathis, A., Galitsky, L., Pottenger, W., Roy, S., Phelps, D. (2004). A Survey of Emerging Trend Detection in Textual Data Mining http://dimacs.rutgers.edu/~billp/pubs/ETDArticle.pdf
13. Lee, Kanghoon; Kim, Geon-Hyeong; Ortega, Pedro; Lee, Daniel: New York: Springer Science and Business Media LLC, Machine Learning, 2018-12-19, Vol.108(5), p.765-783
14. Nesi, P., Pantaleo, G., & Sanesi, G. (2015). A Hadoop based platform for natural language processing of web pages and documents. Journal of Visual Languages and Computing, 31, 130–138. https://doi.org/10.1016/j.jvlc.2015.10.017
15. Netflix Revenue and Usage Statistics (2020). (2018, November 2). Business of Apps. https://www.businessofapps.com/data/netflix-statistics
16. Sarhan, Injy: Spruit, Marco: Can We Survive without Labelled Data in NLP, Transfer Learning for Open Information Extraction, Applied sciences, 2020-08-20, Vol.10 (17), p.5758
17. Soltys, M. (2014). Fair Ranking in Competitive Bidding Procurement: A Case Analysis. Procedia Computer Science, 35, 1138–1144. https://doi.org/10.1016/j.procs.2014.08.207

18. Souili, A., Cavallucci, D., & Rousselot, F. (2015). Natural Language Processing (NLP) – A Solution for Knowledge Extraction from Patent Unstructured Data. Procedia Engineering, 131, 635–643. https://doi.org/10.1016/j.proeng.2015.12.457
19. Xia, B., Chan, A., Zuo, J., & Molenaar, K. (2013). Analysis of Selection Criteria for Design-Builders through the Analysis of Requests for Proposal. Journal of Management in Engineering, 29(1), 19–24. https://doi.org/10.1061/(ASCE)ME.1943-5479.0000119