Name	Initial Category
Fiber Optic Cable, OFNR Riser, #Fibers = 72, Tight buffered, Non-breakout style cable, Indoor, Single	Fiber Optic Cable
Fuse, Small Production	Fuse

Figure 3 is a bar chart that represents counts of the most popular categories (categories with counts over 10). Only about thirty different categories out of the 372 are represented here. The largest category, Fiber Optic Cable, has 56 entries.

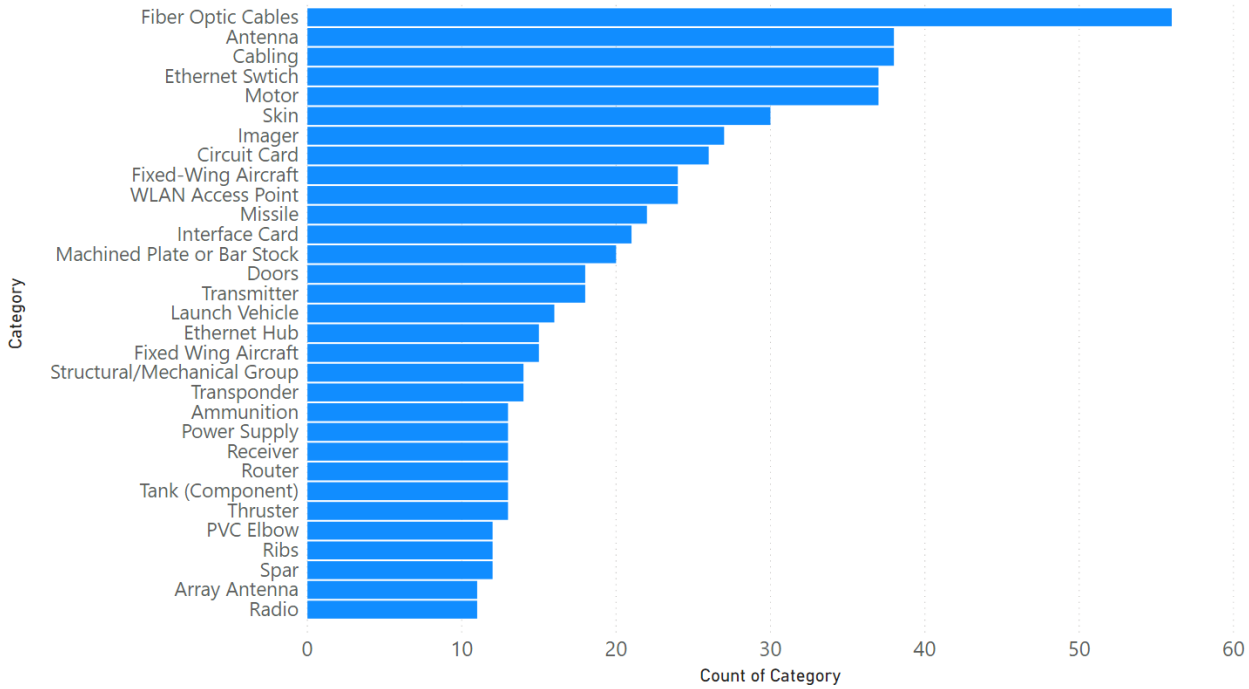

Figure 3: Name counts by Category for the dataset.

For out of 372 assigned categories, about 91% of data in the Category column (340 Categories) had less than 10 names mapped to it. About 85% of the data did not have more than 5 instances in the dataset. Over 50% of the data (201 categories) only has one datapoint associated with it. These issues ultimately led to choosing not to implement a Machine learning model, which will be detailed further in the next sections.

More detail about the distributions of counts (or Names) within the Category column is described in Figure 4.

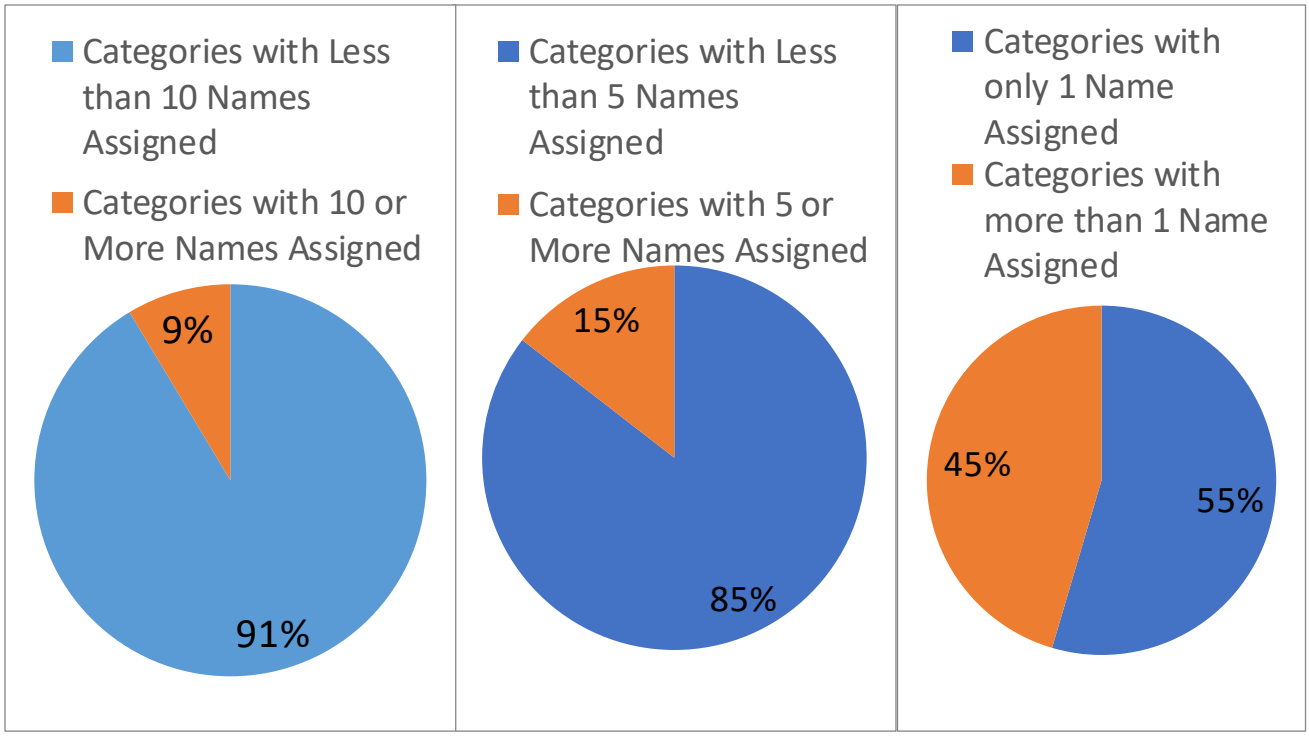

Figure 4: Charts that detail information about Name counts by Category.

Exploring Supervised Multi-Class Classification Machine Learning Models

There are plenty of considerations to make when employing a supervised multi-class classification model. The amount and variety of data supplied to the machine learning model must be robust enough to help the model "learn" to categorize text properly. The labeled data would need to be randomly split into training and test sets, which can be used to grade the accuracy of any machine learning methods tested. If the initial model does not produce satisfactory results, then the machine learning process would be adjusted.

In our example, we have about 375 possible classes (Categories) which would be assigned based on the words in the "Name" column. The machine learning model will also have to identify the label based on the text data in the Name column. It may be able to do this by individual words or groups of words within this column.

Training machine learning models have their own challenges, especially with classes that have similar terminology, as is common in the Aerospace and Defense Industry. In theory, those classes would probably be "confusable" and cause issues with training the data [18]. For example, the entries Antenna, Parabolic Reflector Antenna 25 GHz, and ARC210v Antenna Logic Unit all have the word "Antenna" in them. Would the model be smart enough to recognize that the reason why the last entry is considered an Antenna Logic Unit and not just an Antenna because it has the words "Logic Unit" in it? That seems like an obvious question, but these algorithms can sometimes have strange results.

As stated in the previous section, a majority of classes only have one datapoint assigned in the original dataset. This can cause issues in developing an ML model. The most problematic issue is that we don't have enough data to train and test the model for these Categories.

There are ways to get around datasets where certain classes/labels are over or under-represented in a dataset, which is also called "imbalanced". In this use case, a lot of data is severely under-represented, so there are several options. Of course, more data could always be collected, but this takes effort. Another option could be using a technique to generate fake training data [19] However, the most popular method, called SMOTE, cannot easily generate fake text data due to complications of how text is represented as vectors. However, text could also be randomly generated through "augmentation", which is randomly adding noise and slightly changing already existing entries [20]. Generating fake data may not be particularly helpful for future data collection. It would be impossible to anticipate naming conventions and text structure in future data. Therefore, training a model on mostly fake data is unlikely helpful.

For all of the reasons stated above, it was decided that a supervised classification ML model would not be the best path forward. Thus, other options were explored.

Part of Speech Tagging

The next natural language processing method explored was Part of Speech (POS) tagging. Part of speech tagging is when each word from a sentence gets assigned a part of speech found in a sentence: nouns, verbs, adjectives, etc. Of course, different meanings of the words can have different parts of speech. For example, the word "bug" could be a noun or a verb, so the computer program (the tagger) needs to know the difference. This process gets its "intelligence" from a text corpus - a collection of documents where words were already assigned parts of speech tags. Text corpora are either specific or interdisciplinary collections of text documents (such as novels, news articles, etc.). For further examples of corpora, see section 1.6 here [21].

For this dataset, it would be helpful to identify which words are nouns and which words are adjectives (so either component types or subtypes, respectively). Part of speech tagging was done using Python code with the nltk module [21]. The specific tagger used was the conll2000, which uses tagged sentences from Wall Street Journal articles [22]. Although this is not an ideal corpus for aerospace and defense, it was initially chosen because of its chunking capability (see section below). Plus, there appear to be no tagged corpora in the nltk.org list exclusively meant for engineering disciplines.

Below is a table which contains the normalized Names and the corresponding part of speech tags taken from [23]:

Name	Part of Speech Tags
Accelerometer	[('accelerometer', 'Noun')]
Aileron	[('aileron', 'Noun')]
Wgf Aileron	[('wgf', 'Noun'), ('aileron', 'Noun')]
9mm, Full Metal Jacket, Round Nose, Ammunition	[('9mm', 'Cardinal Digit'), ('full', 'Adjective'), ('metal', 'Noun'), ('jacket', 'Noun'), ('ammunition', 'Noun')]
40, S&w, Full Metal Jacket, Ammunition	[('40', 'Cardinal Digit'), ('s', 'Adjective'), ('w', 'Noun'), ('full', 'Adjective'), ('metal', 'Noun'), ('jacket', 'Noun'), ('ammunition', 'Noun')]
Amplifier	[('amplifier', 'Noun')]
Antenna	[('antenna', 'Noun')]
Arc210v Antenna Logic Unit	[('arc-210', 'Adjective'), ('v', 'Noun'), ('antenna', 'Verb, 3rd Person Singular'), ('logic', 'Adjective'), ('unit', 'Noun')]
Parabolic Reflector Antenna 25 GHz	[('parabolic', 'Adjective'), ('reflector', 'Noun'), ('antenna', 'Verb, Past Tense'), ('25', 'Cardinal Digit'), ('ghz', 'Noun')]
Cf Spine Skin	[('cf', 'Noun'), ('spine', 'Noun'), ('skin', 'Noun')]
Radome Assembly	[('radome', 'Noun'), ('assembly', 'Noun')]
Battery Assembly	[('battery', 'Noun'), ('assembly', 'Noun')]
Thermal Battery	[('thermal', 'Adjective'), ('battery', 'Noun')]
Electronics, Attitude Control	[('electronics', 'Plural Noun'), ('attitude', 'Verb, Sing. Present'), ('control', 'Noun')]
Fuse, Small Production	[('fuse', 'Adverb'), ('small', 'Adjective'), ('production', 'Noun')]

An initial run of POS tagging was fairly unsuccessful. The word "metal" in both ammunition Names is marked as a noun, but it would be viewed as more of a descriptive term than component type. In "Fuse, Small Production", the word "Fuse" was tagged as an adverb! These mistakes probably happened because Parts of Speech Taggers are meant to tag full sentences, not fragments of sentences.

The initial failures brought up an interesting point: do there need to be text corpora focused solely on documents in the cost estimation/aerospace and defense fields? While there is no concrete evidence that aerospace and defense terminology is unique, Crossley et. al. did a

study that stated not only does different subsets of STEM science have a unique vocabulary (such as science versus engineering), but subsets of science (such as mechanical or industrial engineering) should also be treated differently as well [24]. Although this research did not consider aerospace engineering, it is reasonable to assume it is unique as well.

Proposed Solution: Lexicon

After conversation with others in the cost estimation field, it was decided that tagging single words would not be sufficient. After all, words have different meanings when together or apart. For example, the phrase "wing antenna" has different meanings than the words "wing" and "antenna" alone. Text chunking breaks phrases into smaller chunks that can be more easily understood. But this is dependent on part of speech tagging, which was shown in the previous section to be unreliable.

Therefore, a workable solution involved developing a lexicon to pull certain text information from the original data. "A lexicon, or lexical resource, is a collection of words and/or phrases along with associated information such as part of speech and sense definitions" [21]. A keyword list was developed based on general knowledge of the aerospace and defense industry. Primary Keywords were mostly based on component types. Secondary keywords were descriptive words. Both can help organize data for any type of analysis.

A couple of Python modules were written that could detect the presence of the Primary or Secondary Keyword in the Name column. An initial run was done with about 120 phrases. Note that the best practice was to convert the "Name" column from the original data and the "Term" column from our lexicon to lowercase because the Python code was case-sensitive. Below is a small subsample of those:

Term	Key Name Type
accelerometer	Primary
access point	Primary
accessory drive	Primary
accumulator	Primary
actuator	Primary
aileron	Primary
airborne	Secondary
airframe	Primary
altimeter	Primary
aluminum	Secondary
amplifier	Primary
antenna	Primary

For this method, punctuation was stripped from the "Name" column in an attempt for higher accuracy. Here are the results, including some of the data shown in previous sections:

Name	Primary Name	Secondary Name
Accelerometer	['accelerometer']	
Aileron	['aileron']	
Wgf Aileron	['aileron']	
9mm, Full Metal Jacket, Round Nose, Ammunition	['ammunition']	['metal', 'round']
40, S&W, Full Metal Jacket, Ammunition	['ammunition']	['metal']
Amplifier	['amplifier']	
Antenna	['antenna']	
ARC210v Antenna Logic Unit	['antenna', 'logic unit']	
Parabolic Reflector Antenna 25 GHz	['antenna', 'parabolic reflector antenna', 'reflector antenna']	
Cf Spine Skin	['skin']	
Radome Assembly	['radome', 'assembly']	
Battery Assembly	['battery', 'assembly']	
Thermal Battery	['battery']	['thermal']
Electronic, Attitude Control	['attitude control']	['electronic']
Fiber Optic Cable	['cable']	['fiber optic']
Fiber Optic Cable	['cable']	['fiber optic']
Fuse, Small Production	['fuse']	

Of course, this methodology has its own issues. Some of the entries seem a bit redundant, such as in the "Parabolic Reflector Antenna". It is marked as a "parabolic reflector antenna", "antenna" and a "reflector antenna". Not every single word in the "Name" column is categorized, especially proper nouns such as "ARC-210v". In addition, this methodology is also case-sensitive. As of now, the Python code cannot distinguish between plural and singular words. Note that this is a first pass, and it will be impossible to categorize every single word that might be associated with a hardware component in a large field such as aerospace and defense.

Therefore, it would still be recommended to double-check the work of the lexicon extraction method. Plural words were normalized from the beginning, so more work would have to be done if differentiation between the two was key for data analysis. This method also wouldn't catch spelling mistakes. Despite the pitfalls, a lexicon could reduce the time that it takes to initially categorize new data.

Lessons Learned

This case study presents interesting points about using natural language processing and machine learning in cost estimation. Machine learning is not all about coding and understanding algorithms. Like any mathematical model, an important part of machine learning is understanding limitations and what situations these tools should be used in. Additionally, for a machine learning classification model to be successful, classes need to be well represented. Unbalanced classes lead to poor models due to issues with algorithm training. Unfortunately, the data in the use case was severely imbalanced, so other methods in natural language processing had to be used. A lexicon approach could help reduce the time it takes to normalize data, but it is not perfect.

Future Directions

This study was oversimplified because it only looked at one text-type column of the dataset ("Name"). This was done because the other text columns did not have much information. As mentioned earlier in this research, if additional data was collected to make the dataset more balanced in terms of label representation, then a machine learning model may be a possible approach.

Not all possible avenues were explored. Due to time constraints not all possible machine learning algorithms were researched, so there still may be an existing model that could lead to a better solution. A more balanced dataset could be generated to make a machine learning method possible. Also, making a dataset with multiple text labels assigned to each datapoint (called "multi-label") would be more complicated and could be part of a future study.

Conclusion

This paper reviewed natural language processing and machine learning, along with exploring methods within the two. This use case showed why using machine learning techniques would not be successful with this specific dataset. Instead, using natural language processing could be a practical solution. More specifically the proposed method of using lexicons could be a way to pre-process data. This methodology won't catch all mistakes and issues, such as spelling mistakes. It is also case sensitive. It could, however, be an effective first step for classification when ingesting new data.

Special thanks to Adam James of Technomics, Inc. (ajames@technomics.net)

Works Cited

[1] N. J. Nilsson, Introduction To Machine Learning.

[2] W. Thompson, H. Li and A. Bolen, "Artificial intelligence, machine learning, deep learning and beyond," SAS, [Online]. Available: https://www.sas.com/en_us/insights/articles/big-data/artificial-intelligence-machine-learning-deep-learning-and-beyond.html#/.

[3] IBM Cloud Education, "Deep Learning," [Online]. Available: https://www.ibm.com/cloud/learn/deep-learning.

[4] K. Mourikas, N. Hanov, J. King and D. Nelson, "Machine Learning & Non-Parametric Methods for Cost Analysis," Boeing, [Online]. Available: https://www.iceaaonline.com/ready/wp-content/uploads/2018/07/TI04-NOT-APPROVED-FOR-PUBLIC-Machine-Learning-and-Non-Parametric-Methods-Mourikas.pdf.

[5] J. Kilgore, "Intersections of AI and Cost Estimating: Explainability," [Online]. Available: https://www.dhs.gov/sites/default/files/publications/intersections_of_ai_and_cost_estimating_explainability_kilgore.pdf.

[6] C. O'Neil, Weapons of Math Destruction.

[7] Google, [Online]. Available: https://developers.google.com/machine-learning/crash-course/framing/ml-terminology.

[8] Wikipedia, [Online]. Available: https://en.wikipedia.org/wiki/Multi-label_classification.

[9] Google, [Online]. Available: https://developers.google.com/machine-learning/crash-course/training-and-test-sets/splitting-data.

[10] Wikipedia, "Linguistics," [Online]. Available: https://en.wikipedia.org/wiki/Linguistics.

[11] By: IBM Cloud Education, "Natural Language Processing," [Online]. Available: https://www.ibm.com/cloud/learn/natural-language-processing.

[12] "DS Foundation," [Online]. Available: https://datascience.foundation/sciencewhitepaper/natural-language-processing-nlp-simplified-a-step-by-step-guide.

[13] K. Roye and C. Smart, "Beyond Regression: Applying Machine Learning to Parametrics," [Online]. Available: https://www.iceaaonline.com/ready/wp-content/uploads/2019/06/ML06-Paper-Beyond-Regression-Applying-Machine-Learning-Roye.pdf.

[14] M. Johnson and D. Shafer, "Don't Be Scared, Machine Learning is Easy!," [Online]. Available: https://www.iceaaonline.com/ready/wp-content/uploads/2019/06/ML04-Dont-Be-Scared-Machine-Learning-is-Easy-JohnsonMary.pdf.

[15] K. Mourikas, J. Lemus and E. Serrot, "Machine Learning and Natural Language Processing for Cost Analysis," [Online]. Available: https://www.iceaaonline.com/ready/wp-content/uploads/2019/06/ML05-Machine-Learning-and-Natural-Language-Mourikas.pdf.

[16] D. J. Brown and D. Geraghty, "Machine Learning Assisted Data Extraction and Normalization," [Online]. Available: https://www.iceaaonline.com/ready/wp-content/uploads/2019/06/ML02-Machine-Learning-Assisted-Data-Extraction-BrownJonathan.pdf.

[17] V.-S. Ionescu, H. Demian and I.-G. Czibula, "Natural Language Processing and Machine Learning Methods for Software Development Effort Estimation," *Studies in Informatics and Control,* 2017.

[18] M. R. Gupta, S. Bengio and J. Weston, "Training Highly Multiclass Classifiers," *Journal of Machine Learning Resesarch,* no. 15, 2014.

[19] J. Brownlee, "8 Tactics to Combat Imbalanced Classes in Your Machine Learning Dataset," [Online]. Available: https://machinelearningmastery.com/tactics-to-combat-imbalanced-classes-in-your-machine-learning-dataset/.

[20] Towards Data Science, [Online]. Available: https://towardsdatascience.com/how-i-handled-imbalanced-text-data-ba9b757ab1d8.

[21] S. Bird, E. Klein and E. Loper, "2. Accessing Text Corpora and Lexical Resources," [Online]. Available: https://www.nltk.org/book/ch02.html.

[22] E. T. K. Sang. [Online]. Available: https://github.com/teropa/nlp/tree/master/resources/corpora/conll2000.

[23] Geeks For Geeks, [Online]. Available: https://www.geeksforgeeks.org/part-speech-tagging-stop-words-using-nltk-python/.

[24] S. A. Crossley, D. R. Russell, K. Kyle and U. Romer, "Applying Natural Language Processing Tools to a Student Academic Writing Corpus: How Large are Disciplinary Differences Across Science and Engineering Fields," [Online]. Available: https://core.ac.uk/download/pdf/141671635.pdf.

[25] Y. Kothiya, "How I handled imbalanced text data," Towards Data Science, [Online]. Available: https://towardsdatascience.com/how-i-handled-imbalanced-text-data-ba9b757ab1d8.