Investigating Non Linear Relationships In Global Terrorism

By Shalin Singh

December 7 2021

1 Abstract

With terrorism on the rise, and causing mental and physical damage to the general population, it's important to understand why and how these tragic events have occurred. With the amount of data we have floating around today, we can use the attack data to understand why/when and how terrorist attacks happen. We can also use Data Mining techniques to classify what terrorist organization is responsible for the attacks. With the research that has been done, we've examined how some attacks lead to suicide, classify organizations that are responsible for the attacks, model the language data to highlight specific keywords, and use graph based language data to create a terrorist language data attack search engine. With these findings, we can narrow down features/conditions that lead to suicide. We can classify modern day terrorist attacks to see which organizations are responsible for these attacks. Lastly we could use new language data to see if the outcomes are the same as previous attacks.

2 Why Terrorism?

When looking at a large file of terrorist data, there are a variety problems we want to solve. The first problem we want to tackle is prioritizing terrorist data and cluster them into groups. These clusters could be used to see which organizations/attacks are in a high severity and need more analysis. We'll also want to see if we can classify what terrorist groups are responsible for a certain attack. Lastly, we can use the given language data to make more accurate classifiers and model relationships between certain words/subjects in terrorist data. These problems are of interest because with new data, we can see if the attack is a true threat based on the severity clustering. We can also see what terrorist group is responsible for the attack before the authorities find the responsible organization. With the provided language data, we can create a knowledge graph based search engine for finding patterns in terrorist attacks. Previous research around this data-set has focused on classification/clustering the terrorist data. Novelties explored in this research is around the capabilities with the language data. With the provided language data, knowledge graphs and entity recognition have been implemented to extract specific people involved in the attacks and search techniques to highlight patterns with terrorist attack language data. The objective is extract a higher level meaning using natural language processing algorithms and verfiy previous classification techniques done by other researchers. What will be unique in our research is the focus on Natural Language Processing of the attack summary notes. The summary notes column will be used to create language based classifiers that can detect similar patterns in attacks that haven't been documented.

3 Background Research

The topic of global terrorism for data mining has been examined by other institutions. The first article by Bart Schuurman at the institution of security and global affairs, investigated the constitutions of terrorism. This paper focused on classifiers that can determine whether an attack was actions of terrorism or not. This is different from the classifications that are examined in this paper. Ben Schuumran focused on the issues that come with classifying attacks as terrorism wrongfully. This problem is interesting but couldn't be implemented in this paper because the data was all considered "terrorism". The next article "Building the Global Terrorism" by researchers at the "University of Maryland", documented the process of how the global terrorism was built. This paper talks about the methods, evaluation of PGIS data, Comparison across other databases and future projects. Some of the methods examined were data collections, evaluation and statistics. For the data collection, the institute developed a web based interface for adding attacks that have happened recently. After the web interface was developed, data entry for the attacks allowed the database to grow. The next point of focus was examining the PGIS data. The PGIS data is a collection of open source terrorist data collected by multiple organizations. This has pros and cons. The pros of the PGIS is the open access to all terrorist data for exploration and validation. The cons are possible inaccuracies during the data collection. Since PGIS is open source, attacks that are not "terrorism" could be recorded. These are the main issues "University of Maryland" focused on for data integrity and verified data collection done by PGIS. The last article "Open Access Article Quantitative Analysis of Global Terrorist Attacks Based on the Global Terrorism Database" by the "University of China Mining and Technology" focuses on measuring the degree of terrorist attacks. This research focused on using property value and "Harmed/Killed" stats to measure the impact of the terrorist attack on the specified location. This research article used classification algorithms like K Nearest Neighbors/ K means and Logistic regression to measure what factors effect terrorist attacks. This article was focused on analyzing the attributes of the data set and found correlations between columns that effect the outcome of terrorist attacks. This article had the biggest influence on some of the algorithms implemented in this paper. In the next section, we'll examine how the data was preprocessed and manipulated.

4 Examining the Data

The data used in this research was the GTD subset from Kaggle.com. This data set contained all the terrorist attacks that have happened since 1970 to 2017. The data set had important columns such as location, terrorist organization (that committed the attack), date of occurrence, property value damage, amount of people harmed and killed and lastly, the summary notes of the attack. When downloading the data set, a conversion from utf-8 to Latin-1 was needed for pandas to recognize the data frame. We selected certain columns like, date, harmed, killed, location and summary for the main data frame. We dropped columns that weren't needed such as data source/label encoded values. Functions for viewing snapshots of the data were also prepared. The snapshots just displayed the first 10 columns of data to get a high level understanding before looking at thousands of columns. Overall, the main attributes we're looking at are, location, harmed/killed, property value damage and terrorist attack summary notes. In the next section, we'll examine the data processing for the algorithms implemented.

5 Algorithms

Based off the initial data processing, classification/clustering and natural language processing were the best fit for the chosen data set. The first algorithm implemented was K means for severity clustering. The objective was to create severity clusters based off the attributes "Harmed/Killed" and "Property Value Damage". The clusters are supposed to provide a birds eye view of terrorist attacks ranging from severity levels of "High", "Medium" and "Low". Our visual goal is to have dense clusters with three data points that represent these categories. The next algorithm used was "K Nearest Neighbors". This purpose of this algorithm is to classify what terrorist organization is responsible for an attack based off of the attributes "Harmed", "Killed", "Location" and "Group Name". Using these attributes, the model will pick distances associated with the class name "Group name" to identify a terrorist organization. Along with KNN and Clustering, Decision tree's were used to provide a flow chart view of how terrorist attacks can lead to suicide. The predicted classes for this model were "0" and "1". "1" represents cases where the responsible terrorist organization committed suicide during the attack and "0" represents cases that did lead to a suicide. The features/columns used were "Attack Type", "Weapon type", "Property Value Damage". The objective was to see how the tree generated rules that lead to suicide of a terrorist organization during an attack. Along with the use of these classification algorithms, Natural Language processing algorithms such as named entity recognition and knowledge graphs were used to work with attack summary notes section of the data set. Named entity recognition will be used to extract import entities such as people, locations, races and other English language parts of speech. The use of Knowledge graphs is applied for creating "Search Engine" like capabilities for terrorist attack language data. The graph acts as a collection of nodes, the nodes represent words and the edges are the action words associated with it. In the next section, we'll examine how the data was prepared for algorithms.

6 Data Analysis

The settings/environment of this research can be replicated on most computers with at least 4GB of ram. The specific specs was Ubuntu 20.04 8GB ram on Jupiter Notebooks running multi notebook kernels. The data set takes 10.65 seconds to load on the initial data frame. We'll go over the data processing for each algorithm implemented. First was K means clustering for severity clustering. We wanted to create severity clusters based off the attributes "Harmed/Killed" and "Property Value Damage". On the initial data processing, we gathered the Harmed/Killed and Property damage on a yearly basis. After plotting the initial data points, I collected the Harmed/Killed and property value on a monthly basis to add more dense clusters on the graph for K Means Clustering, lastly we label encoded the data in case text values occurred in the "Property value" column. Overall, this was the data processing involved for K means clustering. The next implemented algorithm was "K Nearest Neighbors". The first thing that was done was renamed the columns to camel case to make it easier to identify specific columns. Then we filtered for our specific columns which were "Harmed", "Killed" and "Terrorist Group" to build the model. Then I used "StandardScaler" (Another form of label encoding) from sci-kit learn. I used the Scalar to encode all the columns into each column. Then we created a loop of 31 iterations and fed in K values on an incremental basis. Then I graphed the max key value each time the classifer was called. The data was prepared in this format to create a elbow like curve when view-

ing KNN. The next algorithm was a Decision Tree to view which attacks lead to suicide. The end of the tree is supposed to lead to 2 options. "1" Is the class that represents Suicide and "0" is the class that represents "Not suicide". The first thing I did was grab the columns that I wanted such as "Harmed/Killed", "Weapon Type" and "Attack Type". Just like the other algorithms, we label encoded the columns, then used the Decision Tree classifier from sci-kit learn to plot the results. Lastly, for the Natural Language processing algorithms, we created functions to aggregate all the summaries into large paragraphs. The libraries used were Spacy and NLTK for the natural language algorithms. The algorithms I used were named entity recognition and a knowledge graph. Both these algorithms used the sentence aggregation function to prepare the language data I needed. In the next section we'll examine the results from the data prepared for the algorithms.

7 Results

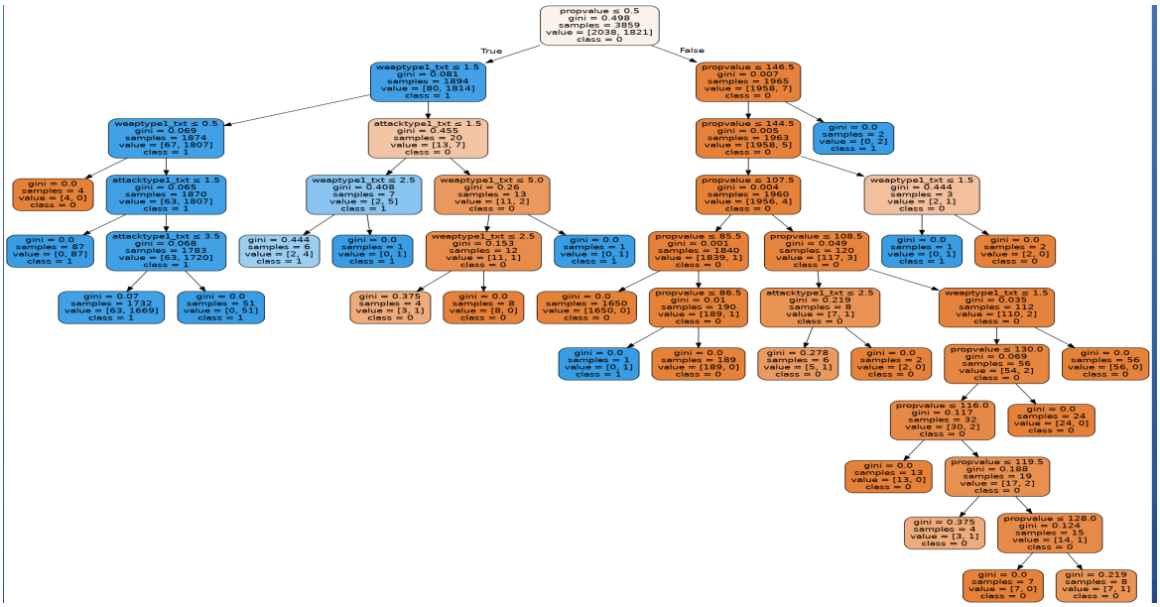

Figure 1: Decision Tree

We'll first look at the results of the decision tree. Trees tend to over fit and learn the data "to well" due to the logic it generates. In the figure above, the blue nodes represent attacks where the terrorist organization committed suicide during or after the attack. The orange nodes are attacks that didn't lead to suicide. When looking at the data without a fitted classifier, it's common to assume that attacks lead to suicide if there are bombings/explosions involved. What this tree shows is that property value damage and weapon type also have an effect on the outcome of suicide. Based off the label encoded values for the weapon type, we can see that incendiary devices are key factor in attacks leading to suicide. We can also see that Weapon Type and Property value are the first two decision nodes when evaluating the outcome. The overall tree shows us that the suicides were effected by weapon type more than the property value damage. The property damage rules isolated attacks that didn't lead to suicide. We can see this due to the tree having a visible separation between blue and orange nodes. Overall, this is the least reliable model due to over fitting and presenting oddly high accuracy from the sci-kit implementation.

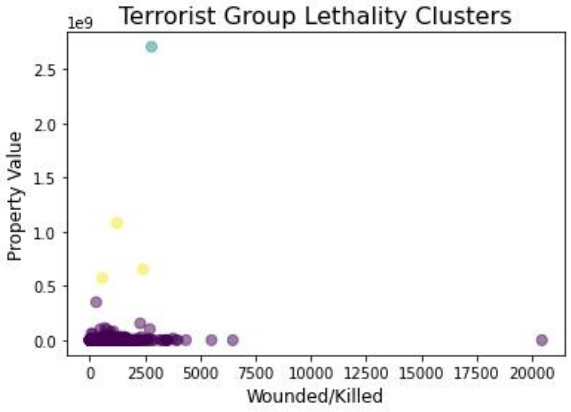

Figure 2: Severity Clusters

Now we'll examine the results of K means clustering. In the figure above, we can see that the purple data points represent the low severity clusters, the yellow points represent medium severity and blue represents high severity. The graph shows that property value damage is a highly relevant factor for clustering data. This graph is using X/Y values based on a monthly basis to create denser clusters on the plot. Overall, we can see that there is a way to prioritize which attacks are the most severe. This can be used as a way to examine highly lethal terrorist organizations.

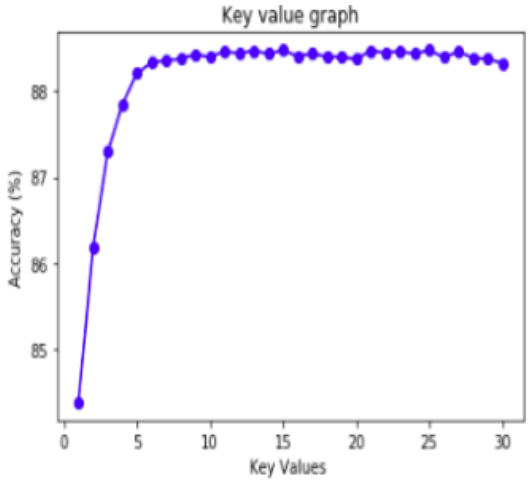

Figure 3: KNN

Now we'll look at the results of the KNN (K Nearest Neighbor) classifier. In the previous section, it was mentioned that the classifier function was called on iterations of 1 to 31. Each point on the graph represents a max key value derived from the results. When the KNN classifier is called, it takes in a key value that represents the amount of neighbors specified for the model. Since this is running multiple times, each iteration value represents the neighbor value that was specified. We can see that the accuracy increases as the neighbor keys approach closer to the 31. We can continue to optimize the results if we have a larger data set and create denser data points. Overall, this algorithm was effective in isolating and classifying terrorist organizations responsible for the attack.

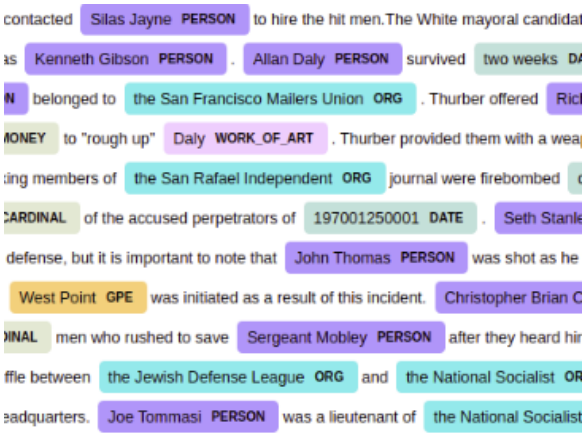

Figure 4: Named Entity Recognition

Now we will look the results of named entity recognition using the "spacy" natural language processing library. Entity recognition allows to extract interesting information such as people, organizations, racial attributes and other parts of speech. We can see in the figure above that blue represents an organization or place where the attack happened. Purple represents a specific person involved in the attack. The data set provides us with only locations for target types. The "person" entity recognition allows us extract more details about the target like the name of the person, racial attributes etc. With the entities provided, we could narrow down specific people and investigate political

motivation or why the attack happened. Overall, this algorithm gives us a better summary of why or how the attack may have happened.

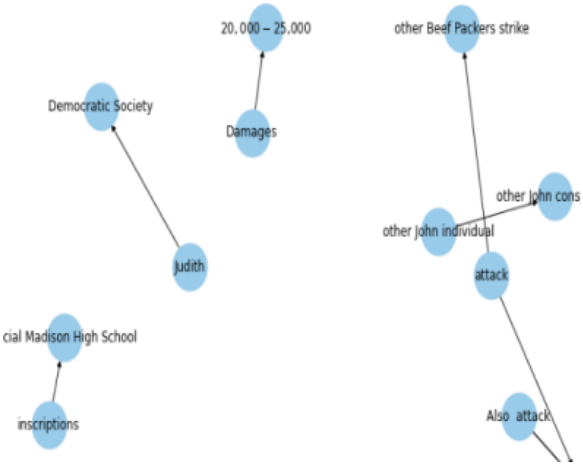

Figure 5: Knowledge Graph

Now we will take a look at the results of the knowledge graph that was implemented. As mentioned before, the knowledge graph is supposed to act as a birds eye view of how the language data is related to other words. We can see in the figure above, the blue nodes represents the words we're creating relationships for. The black line is the edge that contains the "Action" word related to each other. The figure shows us how cardinal values like property value and dates are related to subjects like people, places or things. My idea for this algorithm was model the infrastructure of a search engine. Knowledge graphs are known for being used in famous search engines such as Google, Duck Duck Go etc. With this graph, a terrorist language search engine could be used to stop active threats using language data from social media. In conclusion, we can see that the knowledge graph is a great data structure/model for seeing how the language data is related in parts of speech and how certain actions are related to nouns in the terrorist attack summaries.

8 Conclusion

Overall, crime data sets are a great way to practice data mining techniques on. With the evolution of classification algorithms, individuals attempting to use data mining on crime data sets, we'll be able to notice patterns law enforcement has not. Access to more crime data can create an open source investigation population and allow society to notice patterns in active crimes happening daily. Another important novelty, is the use of natural language processing for text mining social media and previous terrorist attack summaries for prediction or anomaly detection. This was a great project to work on and I'm glad I got the experience to have a deeper understanding of classification and natural language processing. The cons of this project is the lack of a consistent data source. This type of work is most interesting when we have continuous source of data to test how accurate the algorithms are. If you enjoyed the paper, you can follow my work at https://github.com/shaysingh818 to see other papers/projects I'm working on. In the future, I plan on implementing reinforcement learning algorithms on custom built environments.