ANL312

Text Mining and Applied Project

Formulation

End-of-Course Assignment

July 2020 Presentation

Prediction of Terror Attacks using Text Mining

Ho Zhong Ta Benjamin (W1711125)

22 October 2020

1 Introduction

Terrorism and its devastating impacts have been a concern for many countries, regardless of wealth and power. Terrorism, as defined by Hannah Ritchie et al (2013), refers to the use of violence on civilians, designed to instil fear and psychological repercussions, which are motivated by either politics, religion, economic crisis, or social crisis (Ritchie, Hasell, Appel, & Roser, 2013). As terrorist attacks become more global, sophisticated, and well-coordinated, there remains a need to be constantly vigilant against such prominent threats. Such attacks have significant economic consequences (Ruiz Estrada, Park, & Khan, 2018), and psychological consequences (Rubin & Wessely, 2013). The consequences might be long-lasting and persistent, dividing communities and creating social stigmas against specific groups of people (Rubin & Wessely, 2013). Thus, these attacks must be prevented prior to its occurrence. Rapid developments in artificial intelligence, data mining and machine learning allow these processes to take centre stage in the fight against terrorism. Information, in the form of historical data, chat logs and social media fuels these processes and are used to prevent, predict, and possibly identify perpetrators prior to its occurrence (Thuraisingham, 2003). Text Mining aims to discover new knowledge by automatically extracting relevant structured data from textual information (Zanasi, 2007). The structured data then will be available for further modelling, such as classification or clustering. Given the massive amount of textual information available online, the application of text mining on these data might lead us to the possible prediction of possible attacks in the future. Thus, this study proposes a predictive text mining model to predict the possibility of successful terrorist attack using historical data of terrorist attacks as provided by the Global Terrorism Database (GTD).

2 Literature Review

As the Information Age progresses, massive improvement in computing power and the reduction in the cost of data storage has caused data volumes to skyrocket. Physical data is gradually stored on computers instead, leading to an overload of data. Majority of these data are unstructured in nature, which made them costly to analyse and eventually serving little to no purpose other than providing a historical reference. Hidden in all these data were valuable information which is left untouched and undiscovered. However, things changed in 1995 when Ronen Feldman and Ido Dagan (1995) proposed a computerised method to handle all these data, which was initially termed as Knowledge Discovery in Databases (KDD) (Feldman & Dagan, 1995). In their paper, Feldman and Dagon (1995) proposed a data structure, known as the Concept Hierarchy. Each article in the unstructured data will be annotated with a concept,

thus providing the data with a structure. KDD is then used to analyse this data. This framework is then tested using the Reuters-22173 text categorisation test collection, made up of newswires from 1987. With this framework, Feldman and Dagan were able to determine the average topic distribution of different countries, which gave them a numerical overview of their text, which was able to provide them with logical summaries, such as Columbia having more emphasis on coffee than other countries in South America. With this framework, they were able to give structure to unstructured data, which allowed for proper analysis of the once useless data. In 1998, Feldman (1998) then took this framework and coined the term "Text Mining" (Feldman, et al., 1998). Since the introduction of Text Mining, there have been plenty of applications, such as on Customer Relations Management (CRM) and Human Resource Management (Gupta & Lehal, 2009).

After the September 11 attacks, Terrorism became a new focal point for most government across the globe. Recent years, the introduction of the Islamic State of Iraq and Syria (ISIS) and their gruesome attacks gave rise to even more studies on counterterrorism. In his paper, Alessandro Zanasi (2009) stated that these terror groups are going global and turning to the Internet for recruitment and coordination purposes, which gave rise to increased interests in detecting their online presence. The challenge comes with the sheer amount of information available online to be dug through to detect their presence, identify such groups and preventing future attacks. Instinctively, this inspired the application of Text Mining on Counterterrorism (Zanasi, 2009). In his paper, Zanasi proposed the use of Text Mining as a form of information extraction, with the Internet as a source of information. New information, such as data from websites, e-mails, and social media, provides ample inputs for Text Mining, which might provide insights on possible names and relationship detection, money laundering, and even social network links detection. In a similar vein, Norshuhani Zamin and Alan Oxley (2012) suggested a combination of two Text Mining methods, the Unapparent Information Revelation (UIR) which is a graphical method and the Concept Chain Graph (CCG) to discover interesting linkages between topics within unstructured texts (Zamin & Oxley, 2012). Using this method, they were able to create a model which is highly robust in discovering new information, using the 9/11 report in the study. There were several challenges faced, such as noisy texts and the size of the data which will affect the accuracy and the speed of the model. Most current studies have suggested Text Mining as a method to determine interesting relationships within terrorism datasets, which only determine a certain pattern of terrorist activities and their operations and their proposals revolve around more analysis and study of their behaviour. There is a lack of study on actual prediction of possible terrorist attacks based upon the historical data of their

previous attacks. This research gap has inspired the possibility of using a predictive model combined with Text Mining method to predict the possibility of future terrorist attacks and identify the factors behind such attacks.

3 Methodology

3.1 CRISP-DM Framework

The Cross Industry Standard Process for Data Mining (CRISP-DM) Framework will be used to discuss the methodology of the proposed model. The CRISP-DM framework provides a standardised framework to data mining projects regardless of the industries or techniques used (Wirth & Hipp, 2000). The usage of this framework encourages efficiency, by making it large data mining projects faster, more manageable, and less costly. The CRISP-DM framework consists of six phases which are reiterative, as shown in Figure 1 below.

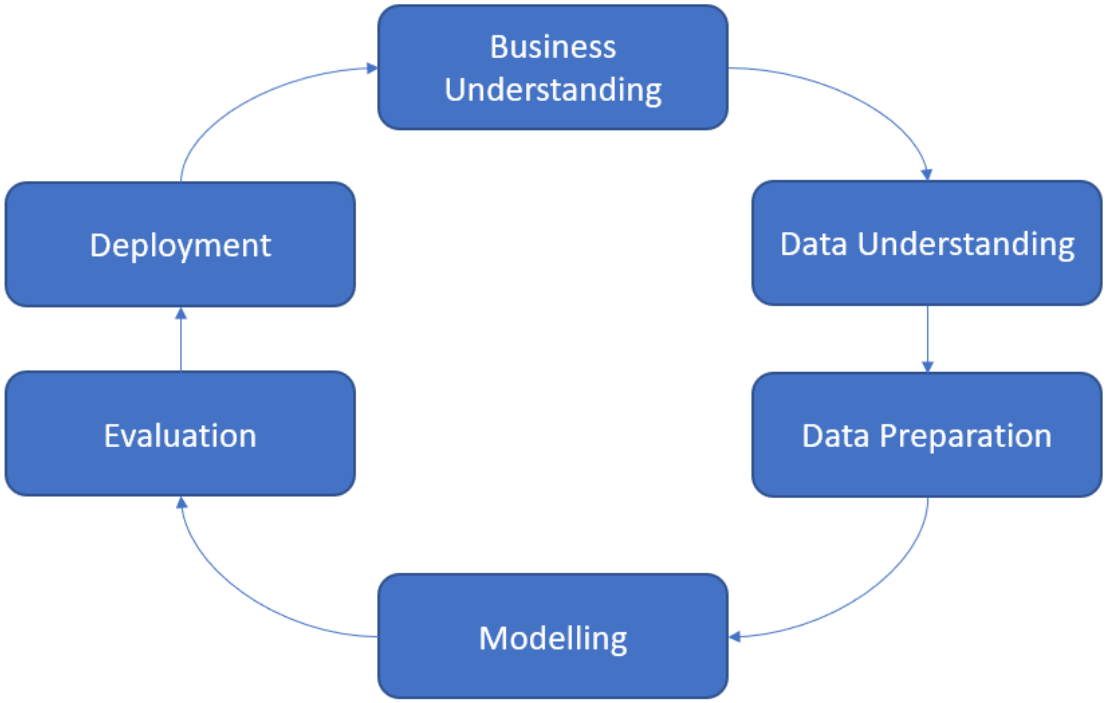


Figure 1: CRISP-DM Framework

The Business Understanding phase of the CRISP-DM seeks to determine the goals and outputs for the data mining project. The Data Understanding phase focuses on the data acquisition, data exploration and data quality verification steps to further understand the data to be used. The Data Preparation phase focuses on cleaning, integrating, preparation and

selection of data to be used for the data mining project. The Modelling phase focuses on model building, model assessment and parameter optimisation of the model. Using the results gathered from the assessment, the model will be evaluated in the Evaluation phase, which will determine the results gathered from the model, processes reviewed and the final model to be deployed will be selected. Upon selection of the final model, the Deployment phase focuses on the deployment plan and monitoring and maintenance of the deployed model (Wirth & Hipp, 2000). The model will be constantly monitored and reviewed and new insights generated from the model will might be able to be inputs in other data mining projects in the Business Understanding phase, thus the emphasis on the iterative nature of CRISP-DM.

3.2 Business Understanding

3.2.1 Objective

This objective of this study is to predict future terrorist attacks based on historical data set of terrorist attacks. The objective will be achieved by the usage of a classification model using results from text mining the summary of attacks as inputs, alongside structured data inputs provided in the GTD database.

3.2.2 Method/Dataset to be Used

The Data Understanding, Data Preparation, Modelling and Evaluation will be done using the IBM SPSS Modeller. For the text mining of the attack's summary will be done using the IBM SPSS Text Analytics function.

The dataset to be used will be from an open-source database called the GTD, which is created by the National Consortium of Study of Terrorism And Responses to Terrorism (START) (START, 2020). This dataset provides a record of information on terrorist events globally since the 1970 to the 2018.

3.2.3 Assumptions to be made

To predict "possible" terrorist attacks, there are several assumptions to be made.

The first assumption to be made is that the attacks recorded in the database are terrorist in nature. As stated by Hannah Ritchie et al (2013), it is hard to determine if an attack is considered a terror attack. For the GTD, there are three criteria to be present for an attack to be considered of a terror nature (START, 2019). Firstly, the goal of the attack must be either political, economic, religious, or social in nature. Secondly, there must be evidence of intention to persuade, intimidate or convey a message to the masses, other than the immediate victims. This means that the attack is carried out with the intention to bring a message across, not just at the specific victim. Lastly, acts of war does not count as terrorism in this database. Although these are criteria provides a boundary of what falls under terrorism, there might be separate attacks that falls through the crack and have the intention of terrorising. However, this database does provide an in-depth coverage of the nature of these terrorist attacks, which might be useful for the prediction of future attacks. Thus, this assumption of criteria behind terrorist attacks should be considered in the assessment of the model.

The second assumption is the definition of a "successful" attack. The following table shows the definition of a successful attack based on the GTD codebook (START, 2019).

Table 1

Definition of Successful Attacks

Attack Type	**Success Definition**
Assassination	Target killed
Armed Assault	Assault taken place and target hit
Bombing/Explosion	Explosive detonated
Hijacking	Hijackers assumes control of vehicle
Hostage Taking	Hostage takers successfully taken control of individuals
Infrastructure Attack	Infrastructure damaged
Unarmed Assault	Victim injured

Note. These definitions are based on the GTD codebook

These assumptions are made and accounted for as the "success" variable in the dataset. However, the successful nature of a terror attack is largely debated as it is psychological in nature and hard to quantify (Ritchie, Hasell, Appel, & Roser, 2013).

The last assumption is that only successful attacks should be considered in the creation of the predictive model. The database consists of both successful and attempted attacks. This study will focus on predicting successful attacks and the factors that determines a successful attack or an unsuccessful attack. This is due to the highly destructive nature of a successful attack as compared to an unsuccessful attack. For example, infrastructural damages in an infrastructure attack and a slain influential figure in an assassination will create more psychological trauma compared to a failed detonation or a failed assassination. Also, this assumption is made as the model seeks to determine and predict future terrorist attacks, and it is more important to understand the nature of these successful attacks and taking steps to preventing it.

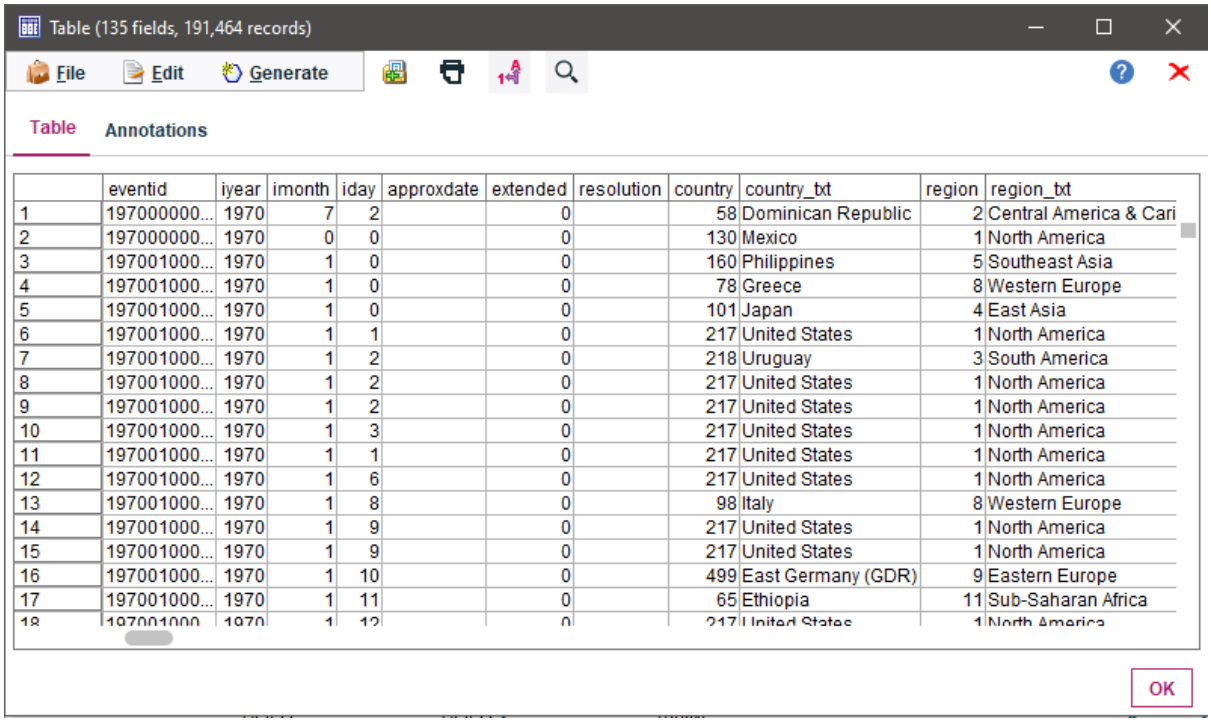
These assumptions should be considered when deploying the model to prevent any unforeseen circumstances, such as isolated incidents that does not qualify in the database's criteria.

3.3 Data Understanding

3.3.1 Overview of Data

The initial dataset downloaded from GTD is in Excel (.xlsx) format. To improve efficiency of data reading and modelling, the dataset is converted into the Tab-separated Values (.TSV) format. This decision is due to the presence of commas within entries of dataset, and the TSV format eliminate such possible read errors.

Using a Table node in IBM SPSS, Figure 2 illustrates the overview of the dataset.



	eventid	iyear	imonth	iday	approxdate	extended	resolution	country	country_txt	region	region_txt
1	197000000...	1970	7	2		0		58	Dominican Republic	2	Central America & Cari
2	197000000...	1970	0	0		0		130	Mexico	1	North America
3	197001000...	1970	1	0		0		160	Philippines	5	Southeast Asia
4	197001000...	1970	1	0		0		78	Greece	8	Western Europe
5	197001000...	1970	1	0		0		101	Japan	4	East Asia
6	197001000...	1970	1	1		0		217	United States	1	North America
7	197001000...	1970	1	2		0		218	Uruguay	3	South America
8	197001000...	1970	1	2		0		217	United States	1	North America
9	197001000...	1970	1	2		0		217	United States	1	North America
10	197001000...	1970	1	3		0		217	United States	1	North America
11	197001000...	1970	1	1		0		217	United States	1	North America
12	197001000...	1970	1	6		0		217	United States	1	North America
13	197001000...	1970	1	8		0		98	Italy	8	Western Europe
14	197001000...	1970	1	9		0		217	United States	1	North America
15	197001000...	1970	1	9		0		217	United States	1	North America
16	197001000...	1970	1	10		0		499	East Germany (GDR)	9	Eastern Europe
17	197001000...	1970	1	11		0		65	Ethiopia	11	Sub-Saharan Africa
18	197001000...	1970	1	12		0		217	United States	1	North America

Figure 2: Overview of Dataset (Table Node)

Based on Figure 2, the dataset has 135 fields and 191,464 records.

3.3.2 Data Exploration

Prior to data preparation, the dataset must be studied to better understand the context of the data within the dataset and determine fields which are significant in the study.

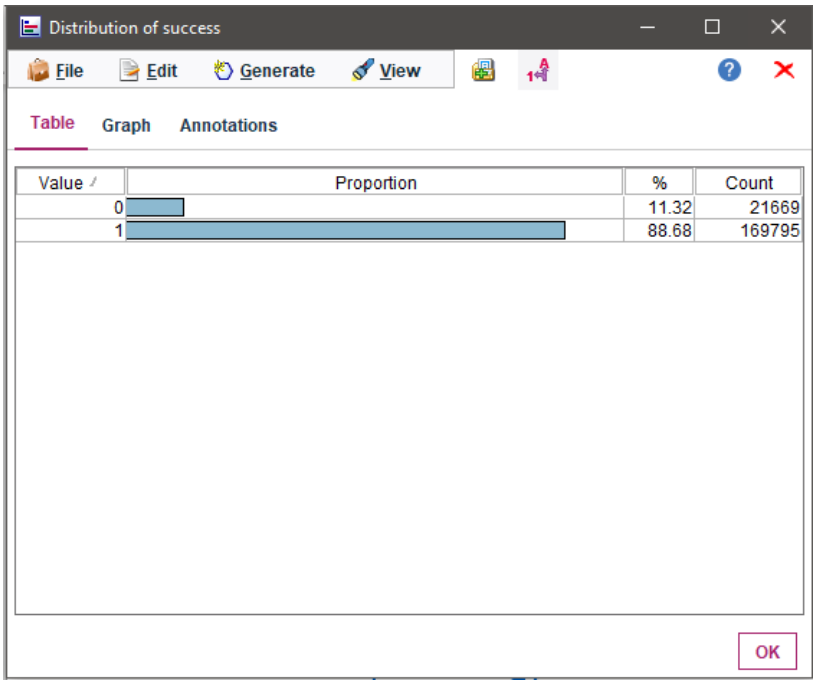


Figure 3: Distribution of success

Figure 3 shows a distribution of the "success" field created using the IBM SPSS's distribution node. This illustrates the number of successful attacks against the number of non-successful attacks, and it shows that 88.68% of attacks in the database are successful. This shows that the dataset has ample entries for successful attacks that can be used for modelling.

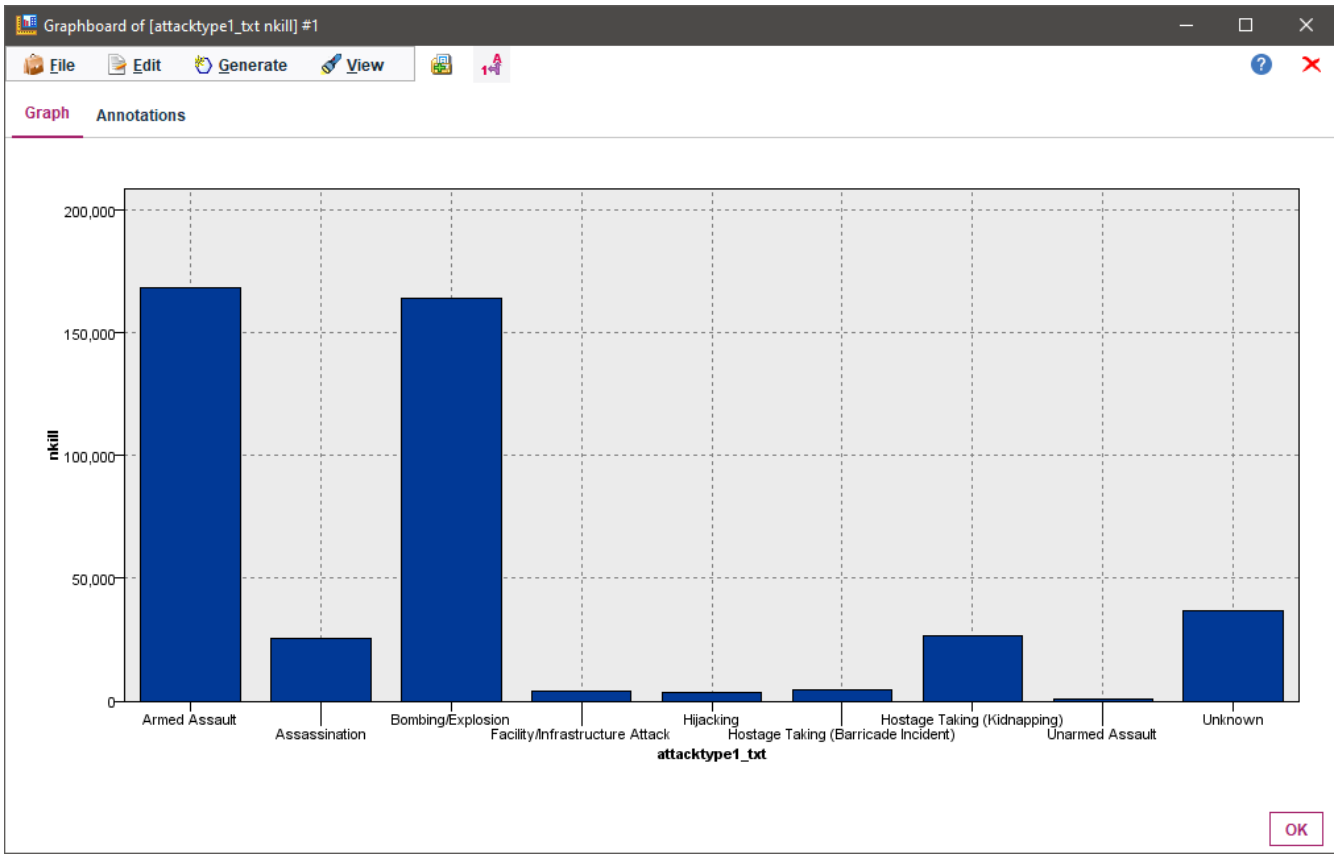


Figure 4: Bar Chart of attacktype1_txt against nkill

Figure 4 plots the "attacktype1_txt" field, which stands for the type of terrorist attack in plain text format, against "nkills", which stands for number of fatalities (START, 2019). Figure 4 shows that significant majority of fatalities are due to armed assault and bombing/explosion. Thus, this shows that significant emphasis should be done to predict occurrences of possible armed assault and bombing to prevent loss of lives.

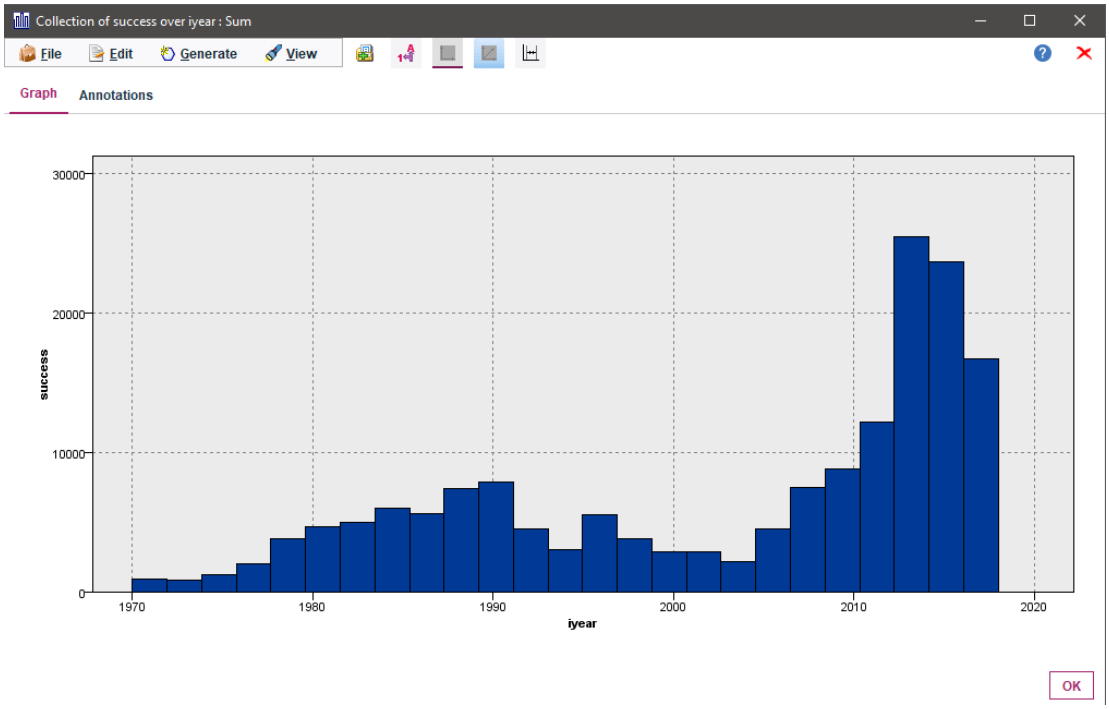


Figure 5: Histogram of success over iyear

Figure 5 illustrates a histogram of success over the years, created using the Graphboard node in SPSS. Figure 5 shows that majority of successful cases occurs after 2010, rising above the 10,000 mark counts in 2011 and remained above that until 2018. This shows that significant emphasis should be placed on attacks that happened after 2010.

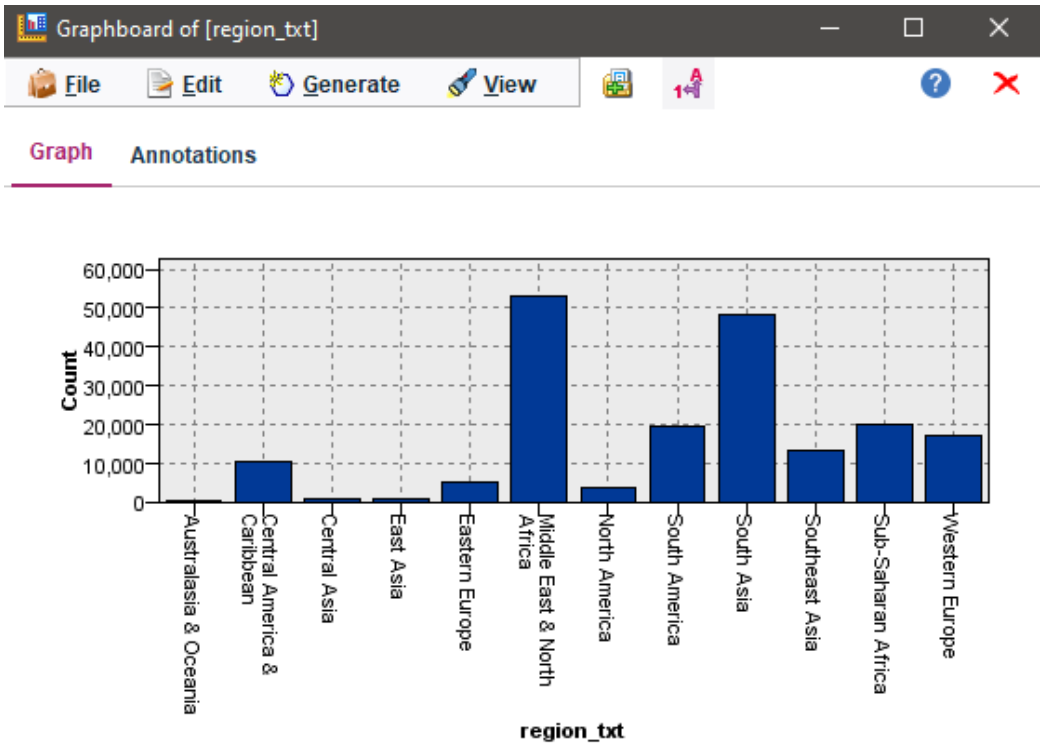
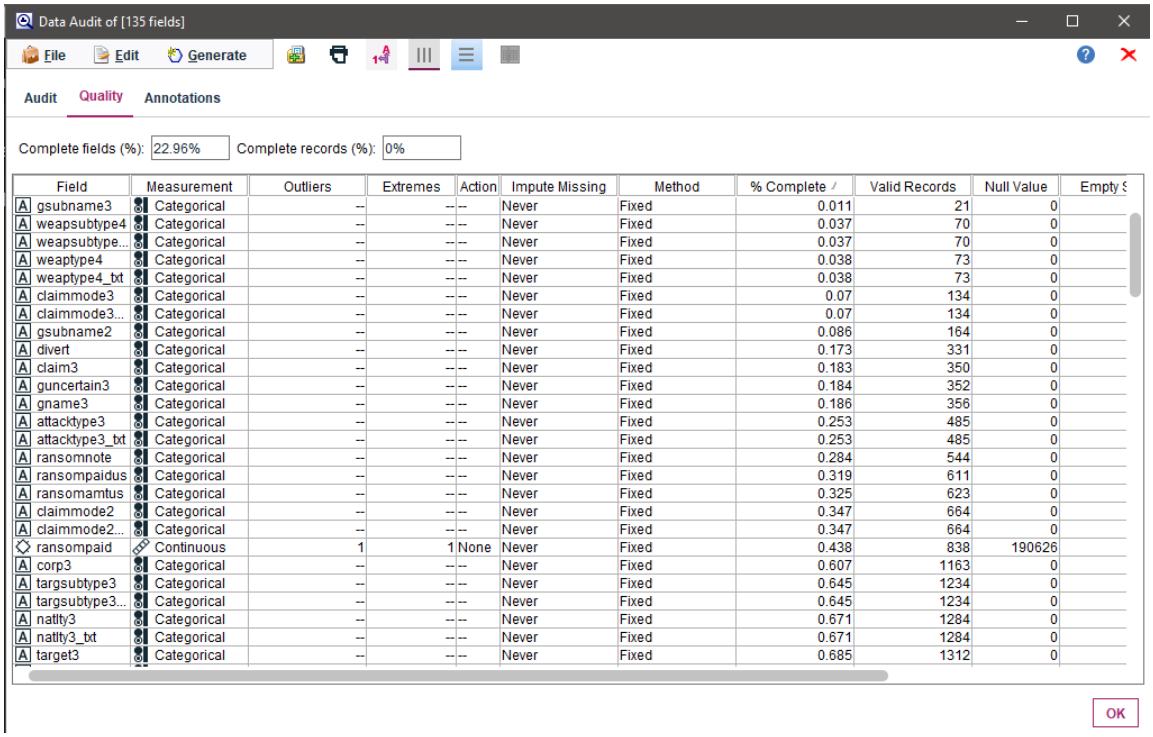


Figure 6: Count of Attacks in Different Regions

Figure 6 highlights the number of attacks in different regions throughout 1970 to 2018. From this, it can be highlighted that majority of terror attacks occurs in the Middle East, North Africa and South Asia.

3.3.3 Data Quality



Field	Measurement	Outliers	Extremes	Action	Impute Missing	Method	% Complete	Valid Records	Null Value	Empty S
gsubname3	Categorical	--	--	--	Never	Fixed	0.011	21	0	
weapsubtype4	Categorical	--	--	--	Never	Fixed	0.037	70	0	
weapsubtype...	Categorical	--	--	--	Never	Fixed	0.037	70	0	
weaptype4	Categorical	--	--	--	Never	Fixed	0.038	73	0	
weaptype4_txt	Categorical	--	--	--	Never	Fixed	0.038	73	0	
claimmode3	Categorical	--	--	--	Never	Fixed	0.07	134	0	
claimmode3...	Categorical	--	--	--	Never	Fixed	0.07	134	0	
gsubname2	Categorical	--	--	--	Never	Fixed	0.086	164	0	
divert	Categorical	--	--	--	Never	Fixed	0.173	331	0	
claim3	Categorical	--	--	--	Never	Fixed	0.183	350	0	
guncertain3	Categorical	--	--	--	Never	Fixed	0.184	352	0	
gname3	Categorical	--	--	--	Never	Fixed	0.186	356	0	
attacktype3	Categorical	--	--	--	Never	Fixed	0.253	485	0	
attacktype3_txt	Categorical	--	--	--	Never	Fixed	0.253	485	0	
ransomnote	Categorical	--	--	--	Never	Fixed	0.284	544	0	
ransompaidus	Categorical	--	--	--	Never	Fixed	0.319	611	0	
ransomamtus	Categorical	--	--	--	Never	Fixed	0.325	623	0	
claimmode2	Categorical	--	--	--	Never	Fixed	0.347	664	0	
claimmode2...	Categorical	--	--	--	Never	Fixed	0.347	664	0	
ransompaid	Continuous	1	1	None	Never	Fixed	0.438	838	190626	
corp3	Categorical	--	--	--	Never	Fixed	0.607	1163	0	
targsubtype3	Categorical	--	--	--	Never	Fixed	0.645	1234	0	
targsubtype3...	Categorical	--	--	--	Never	Fixed	0.645	1234	0	
natlty3	Categorical	--	--	--	Never	Fixed	0.671	1284	0	
natlty3_txt	Categorical	--	--	--	Never	Fixed	0.671	1284	0	
target3	Categorical	--	--	--	Never	Fixed	0.685	1312	0	

Figure 7: Data Audit of Dataset

Using a Data Audit node in SPSS, it can be shown that the dataset has only 22.96% complete fields. The dataset consists of many missing values and null values, which is large due to the absence of precise data regarding a secondary field. An example of this is the field, “weapsubtypes4” which is only 0.037% complete. This fields stands for the fourth weapon sub-type, which might not be present in every attack. Thus, these fields are left blank and attributes to the poor quality of the dataset.

3.4 Data Preparation

3.4.1 Data Reduction/Selection

The database consists of 135 fields and 191,464 records. The data needs to be reduced to a smaller dataset to create a more meaningful model. As the data spans over 48 years (1970 – 2018) and terrorism, along with political and economic struggles changes significantly throughout the years (Hannover Re, 2020), the years of data to be used should be reduced. Based on the data exploration phase, it has shown that majority of successful attacks occurs after 2010 and require more emphasis. Thus, the data to be used will be from only 2010.

Also, the dataset consists of many attacks which occurred in the Middle East, North Africa and South Asia. These regions are significantly different from Southeast Asia politically and economically. For example, majority of attacks in South Asia are due to conflicts over Pakistan and India, which has a long and complex history of unrest (Patra, 2019), which are not as relevant to Southeast Asia in general. Thus, more emphasis should be placed on the Southeast Asia Region.

The data reduction process will start with reducing the number of records based on region and year, which will be done using a Select Node as shown in Figure 8.

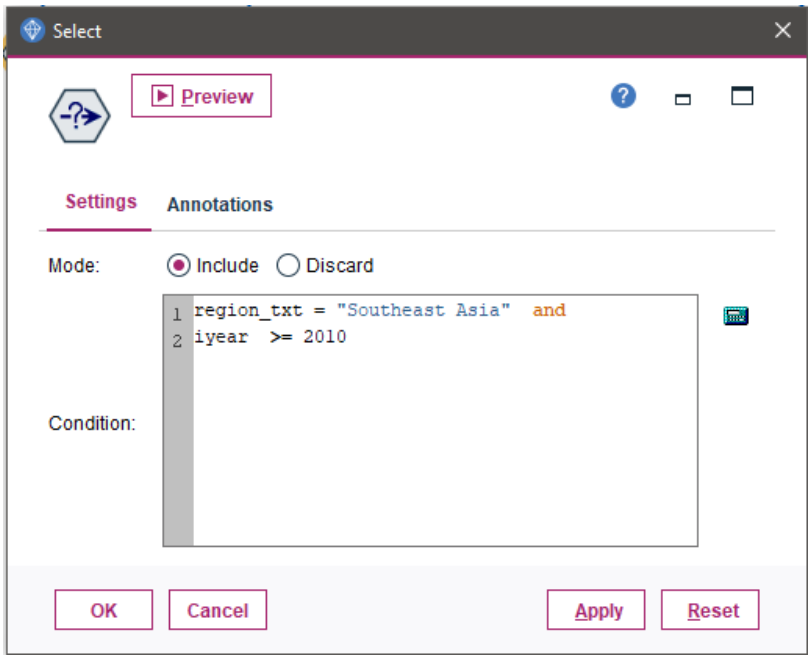


Figure 8: Filtering Based on iyear and region_txt

The large number of fields are also a concern and should be reduced to create a meaningful model. Thus, only 13 most relevant fields are chosen using the Filter Node, as shown in Figure 9.

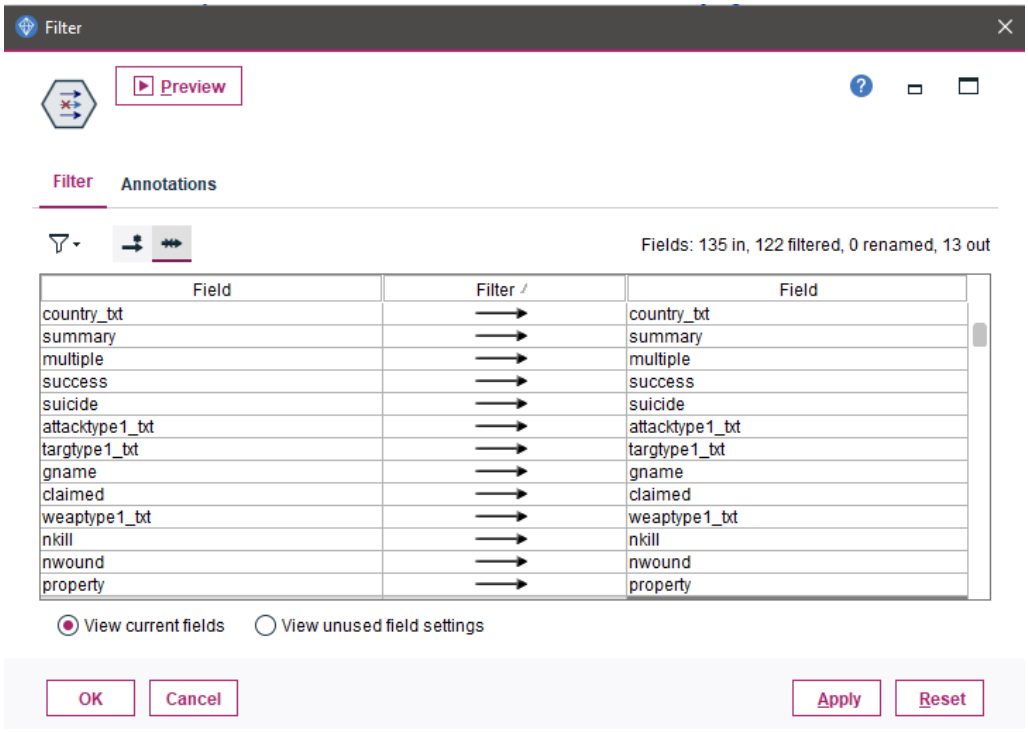


Figure 9: Filtering Relevant Fields

Table 2

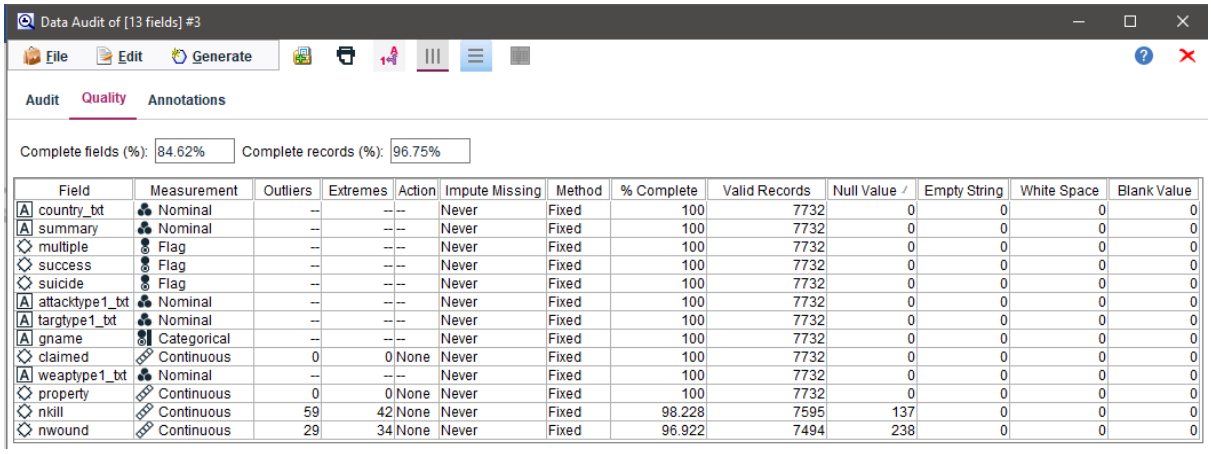
Fields Selected for Modelling

	Field Name	**Description**	**Data Type**	**Role**
1	country_txt	Name of Country	Nominal	Input
2	multiple	Multiple attacks occurred	Flag	Input
3	suicide	Suicide bombing occurred	Flag	Input
4	attacktype1_txt	Type of Attack	Nominal	Input
5	nkill	Number of Fatalities	Continuous	Input
6	nwound	Number of Wounded	Continuous	Input
7	property	Property Damage occurred	Flag	Input
8	summary	Summary of Attacks	Nominal	Input

Fields 1 to 7 are structured data which will be used in the final predictive model. Field 8 (summary) will be used for the Text Categorisation Model.

3.4.2 Data Cleaning

After data reduction, a Data Audit node is used to determine quality of the resulting dataset. The result is shown in Figure 10 below.



Field	Measurement	Outliers	Extremes	Action	Impute Missing	Method	% Complete	Valid Records	Null Value	Empty String	White Space	Blank Value
country_txt	Nominal	--	--	--	Never	Fixed	100	7732	0	0	0	0
summary	Nominal	--	--	--	Never	Fixed	100	7732	0	0	0	0
multiple	Flag	--	--	--	Never	Fixed	100	7732	0	0	0	0
success	Flag	--	--	--	Never	Fixed	100	7732	0	0	0	0
suicide	Flag	--	--	--	Never	Fixed	100	7732	0	0	0	0
attacktype1_txt	Nominal	--	--	--	Never	Fixed	100	7732	0	0	0	0
targtype1_txt	Nominal	--	--	--	Never	Fixed	100	7732	0	0	0	0
gname	Categorical	--	--	--	Never	Fixed	100	7732	0	0	0	0
claimed	Continuous	0	0	None	Never	Fixed	100	7732	0	0	0	0
weaptype1_txt	Nominal	--	--	--	Never	Fixed	100	7732	0	0	0	0
property	Continuous	0	0	None	Never	Fixed	100	7732	0	0	0	0
nkill	Continuous	59	42	None	Never	Fixed	98.228	7595	137	0	0	0
nwound	Continuous	29	34	None	Never	Fixed	96.922	7494	238	0	0	0

Figure 10: Results of Data Audit (Post-data reduction)

A Select Node is used to remove the missing and null values as shown in Figure 11.

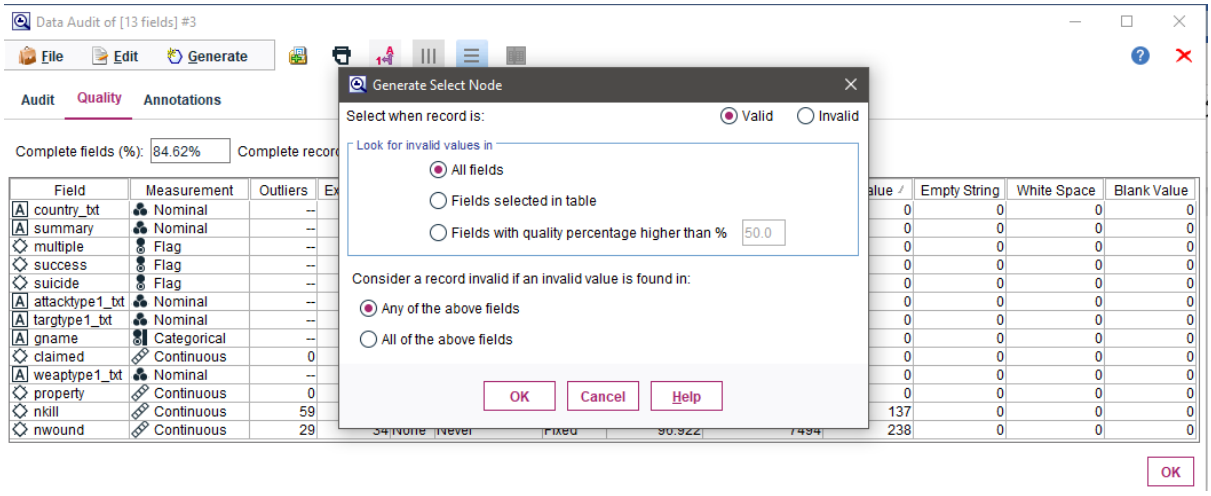
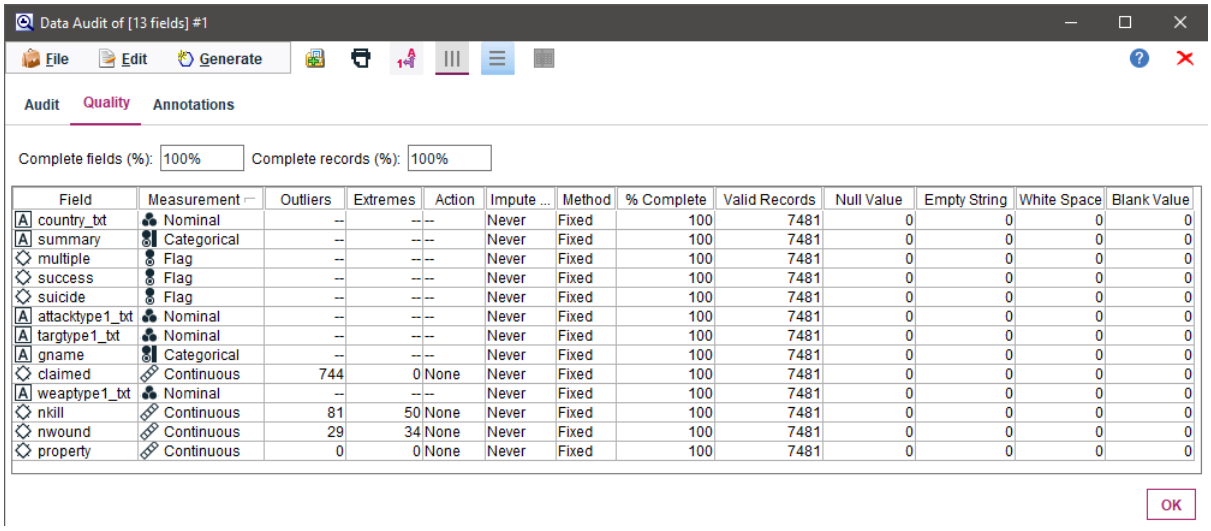


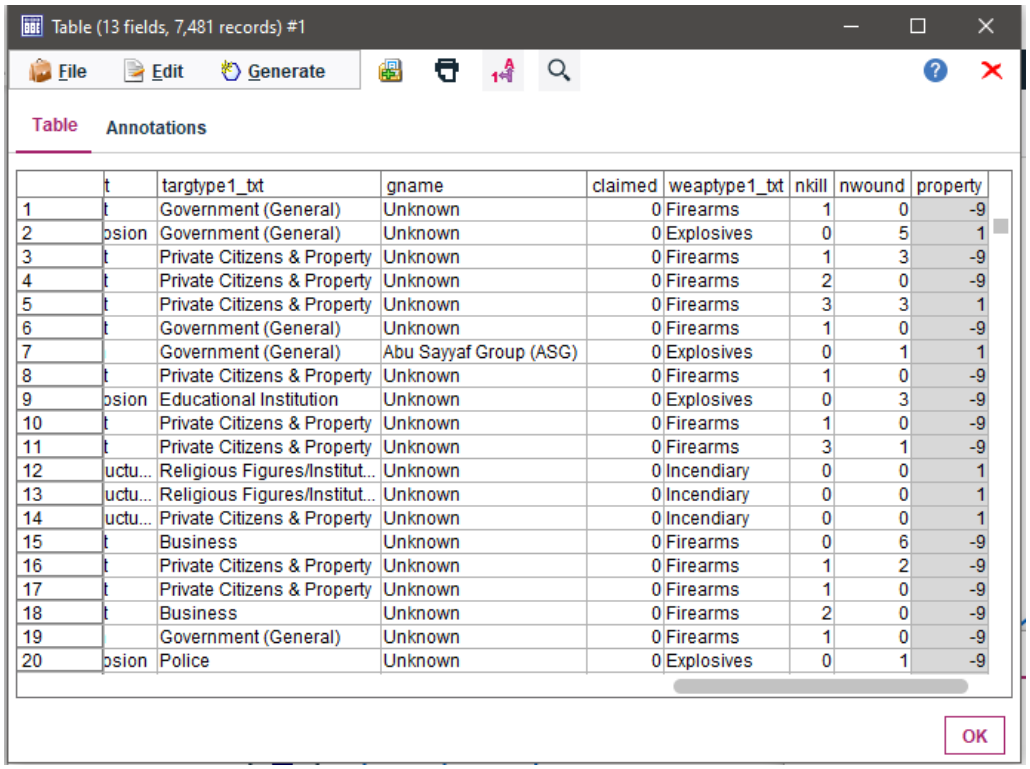
Figure 11: Selecting Records without Missing/Null Values



Field	Measurement	Outliers	Extremes	Action	Impute ...	Method	% Complete	Valid Records	Null Value	Empty String	White Space	Blank Value
country_txt	Nominal	--	--	--	Never	Fixed	100	7481	0	0	0	0
summary	Categorical	--	--	--	Never	Fixed	100	7481	0	0	0	0
multiple	Flag	--	--	--	Never	Fixed	100	7481	0	0	0	0
success	Flag	--	--	--	Never	Fixed	100	7481	0	0	0	0
suicide	Flag	--	--	--	Never	Fixed	100	7481	0	0	0	0
attacktype1_txt	Nominal	--	--	--	Never	Fixed	100	7481	0	0	0	0
targtype1_txt	Nominal	--	--	--	Never	Fixed	100	7481	0	0	0	0
gname	Categorical	--	--	--	Never	Fixed	100	7481	0	0	0	0
claimed	Continuous	744	0	None	Never	Fixed	100	7481	0	0	0	0
weaptype1_txt	Nominal	--	--	--	Never	Fixed	100	7481	0	0	0	0
nkill	Continuous	81	50	None	Never	Fixed	100	7481	0	0	0	0
nwound	Continuous	29	34	None	Never	Fixed	100	7481	0	0	0	0
property	Continuous	0	0	None	Never	Fixed	100	7481	0	0	0	0

Figure 12: Resulting Data Audit after Removal of Missing/Null Values

Figure 12 shows the final Data Audit after removal of missing and null values. The dataset has 100% complete fields.



	t	targtype1_txt	gname	claimed	weaptype1_txt	nkill	nwound	property
1	t	Government (General)	Unknown	0	Firearms	1	0	-9
2	osion	Government (General)	Unknown	0	Explosives	0	5	1
3	t	Private Citizens & Property	Unknown	0	Firearms	1	3	-9
4	t	Private Citizens & Property	Unknown	0	Firearms	2	0	-9
5	t	Private Citizens & Property	Unknown	0	Firearms	3	3	1
6	t	Government (General)	Unknown	0	Firearms	1	0	-9
7		Government (General)	Abu Sayyaf Group (ASG)	0	Explosives	0	1	1
8	t	Private Citizens & Property	Unknown	0	Firearms	1	0	-9
9	osion	Educational Institution	Unknown	0	Explosives	0	3	-9
10	t	Private Citizens & Property	Unknown	0	Firearms	1	0	-9
11	t	Private Citizens & Property	Unknown	0	Firearms	3	1	-9
12	uctu...	Religious Figures/Institut...	Unknown	0	Incendiary	0	0	1
13	uctu...	Religious Figures/Institut...	Unknown	0	Incendiary	0	0	1
14	uctu...	Private Citizens & Property	Unknown	0	Incendiary	0	0	1
15	t	Business	Unknown	0	Firearms	0	6	-9
16	t	Private Citizens & Property	Unknown	0	Firearms	1	2	-9
17	t	Private Citizens & Property	Unknown	0	Firearms	1	0	-9
18	t	Business	Unknown	0	Firearms	2	0	-9
19		Government (General)	Unknown	0	Firearms	1	0	-9
20	osion	Police	Unknown	0	Explosives	0	1	-9

Figure 13: Issue with "property" Field

Figure 13 shows a Table of the resulting dataset. One issue with the resulting dataset is the presence of "-9" in the property field. Based on the GTD codebook, it refers to unknown property damage (START, 2019). This provides no value to the field and can be considered

as a null value. Thus, the "-9" value is removed from the property field using a Select node as shown in Figure 14 below.

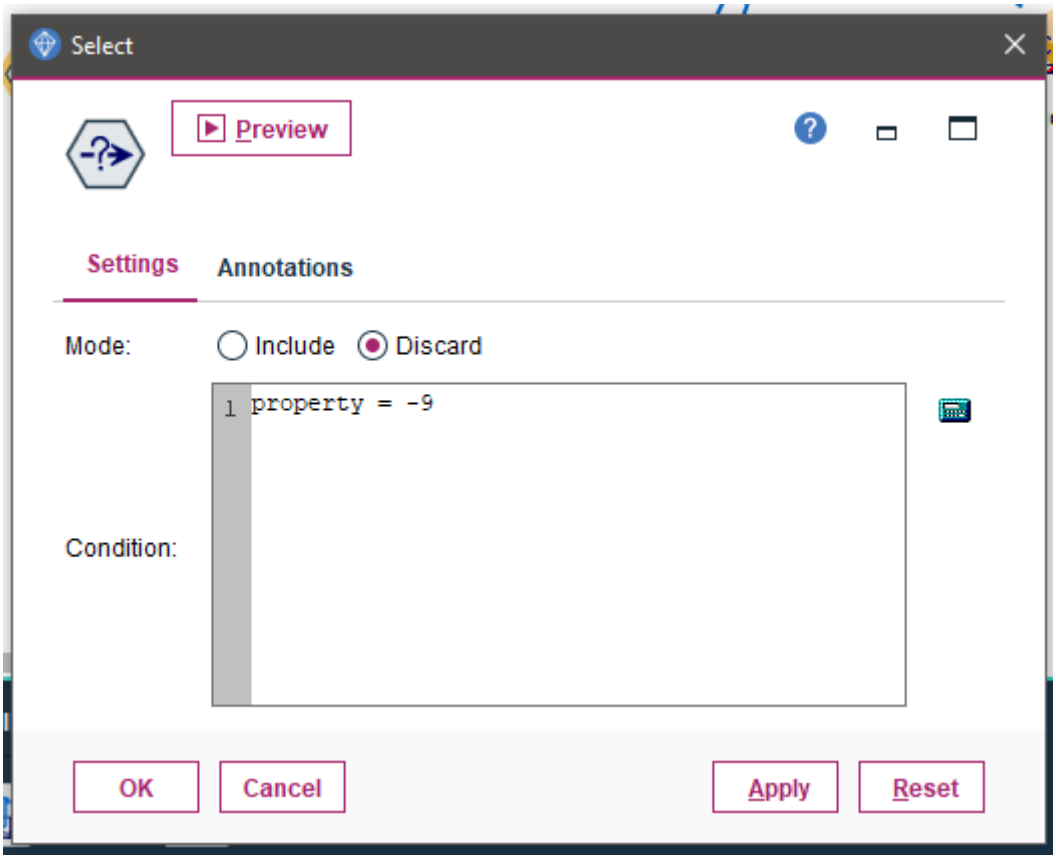


Figure 14: Removing Unknown Property Damage

3.4.3 Data partitioning

Data partitioning is required prior to the predictive model. The partition is created to split the dataset in to two different partitions, Training Set and Test Set. The training set is used to train the predictive model while the test set, which is not seen by the model, will be used to test the predictive abilities of the model. For this study, the training set will be set to 70% while the testing is set to 30%. Figure 15 below shows the setting of the Partition Node used.

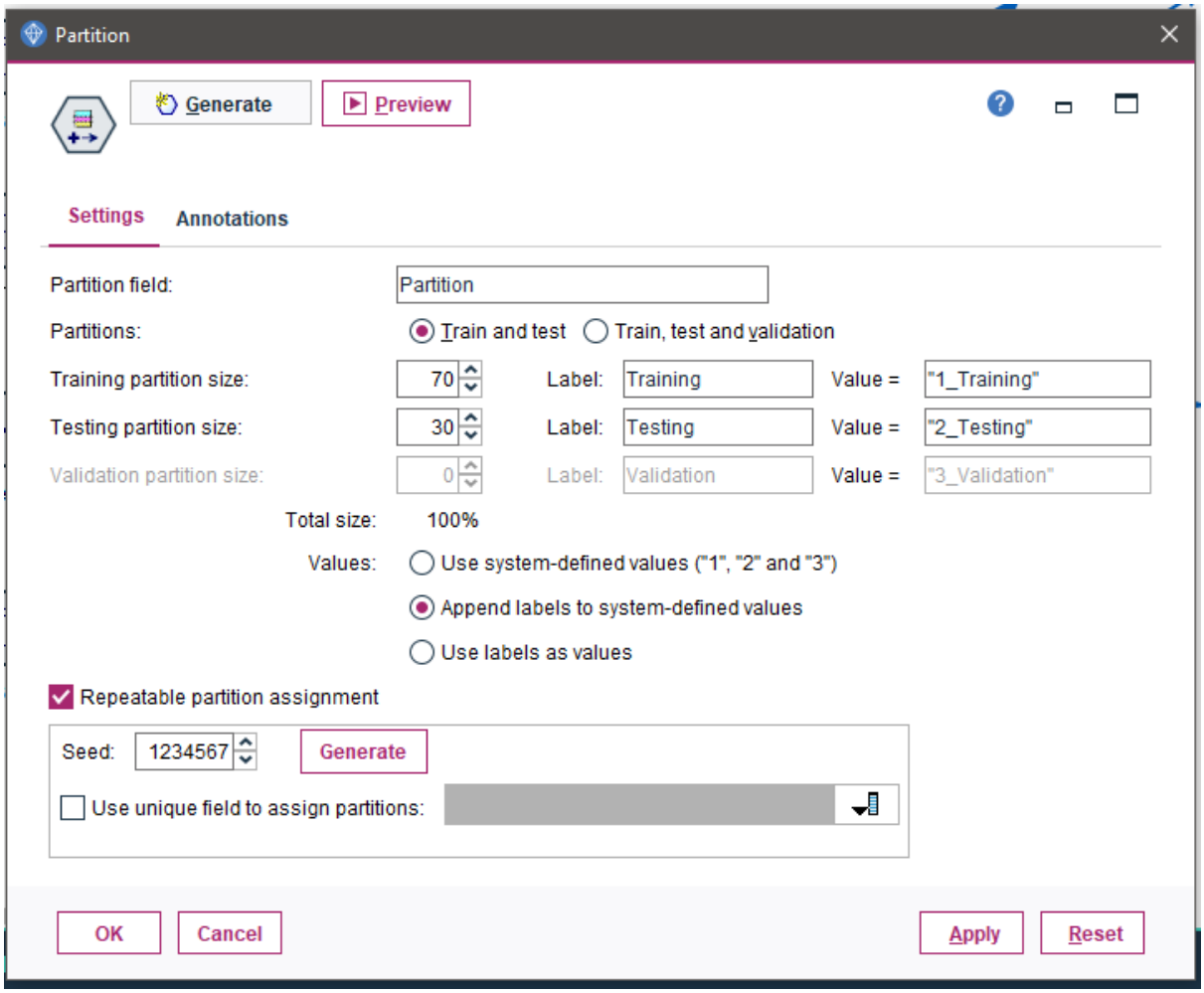


Figure 15: Settings of Partition Node

Figure 16 below shows the IBM SPSS stream that was used for the data exploration and data preparation stage

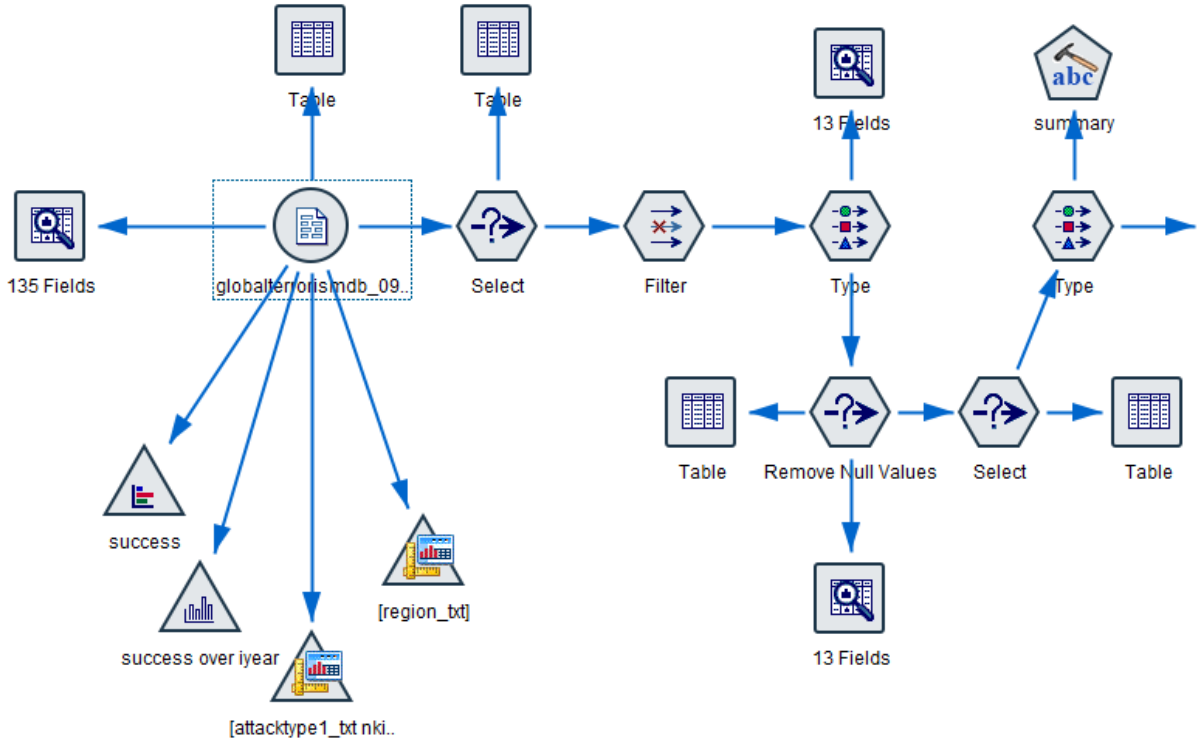


Figure 16: Stream used for Data Exploration and Data Preparation

3.5 Modelling

3.5.1 Text Mining

One data field, summary, consists of the summary of the attack in the dataset. This field exists in free text format as shown in Figure 17.

summary
01/13/2010: On Wednesday, five civilians were wounded when suspected separatist militants detonated an improvised explosive device outside a tea shop in Panarea area in Pattanis Muang district. Investigators fo...
01/18/2010: On Monday, suspected partisans using assault rifles opened fire at farmers houses in Carmen, North Cotabato, killing three villagers and wounding three others. The fatalities have been identified as H...
01/19/2010: On Tuesday evening at 2040, source details that a civilian was wounded when unidentified militants detonated an improvised explosive device outside of the residence of Lamitan City Mayor Roderick Fu...
01/22/2010: On Friday, in the Sibuhuan area of Indonesias Sumatra, in one of three related attacks, a Protestant church was burned down in a violent protest by Muslim activists. The group also burned down a secon...
01/22/2010: On Friday, in the Sibuhuan area of Indonesias Sumatra, in one of three related attacks, a Protestant church was burned down in a violent protest by Muslim activists. The group also burned down another ...
01/22/2010: On Friday, in the Sibuhuan area of Indonesias Sumatra, in one of three related attacks, the private residence of a priest was burned down in a violent protest by Muslim activists. The group also burned d...
01/29/2010: On Friday night at 0400, inn Zamboanga City, Mindanao Island, Philippines, in one of two related incidents, unknown assailants detonated an improvised explosive device (IED) on Sevilla Street. Shortly a...
01/29/2010: On Friday night at 0415, in Zamboanga City, Mindanao Island, Philippines, in one of two related incidents, unknown assailants detonated an improvised explosive device (IED) in Plaza Pershing. No casu...
01/31/2010: On Sunday morning, a six-year-old boy and his parents, aged 42 and 50, were killed as they were travelling by motorcycle in Sai Buri district of Pattani province. A bullet also grazed the couples youngest ...
02/01/2010: On Monday morning, in Sumisip, Basilan, Philippines, assailants detonated an improvised explosive device targeting an armored vehicle owned by Ben Hataman, a candidate for mayor in May elections, ...
02/01/2010: On Monday afternoon at 1630, Elmer Eusebio, a pro government militia man, was playing with his kids at his residence at Sta. Cruz village when seven rebels barged inside his home and dragged him a...
02/09/2010: On Tuesday morning at 0740, in Paluan, Mindoro Occidental, Philippines, a Globe Telecom site was attacked and set on fire by ten armed men. No group claimed responsibility, although it was widely b...
02/09/2010: On Tuesday morning, a 41-year-old Muslim senior district official was shot dead by suspected separatist militants in the Balo area of Raman district in Thailands Yala province. No group has claimed re...
02/10/2010: On Wednesday, a 45-year-old Muslim villager was killed instantly when two men on a motorcycle opened fire as he rode his motorbike home from a mosque in an unknown city in the province of Pattani, ...
02/10/2010: On Wednesday, suspected militants shot the assistant to a village leader, a 47-year-old Muslim man, as he rode home from a market in an unspecified city in Pattani province, Thailand. He later died of h...
02/11/2010: On Thursday morning, in Salvador, Lanao del Norte, Philippines, three men fired upon a municipal councilor, Subair Tawantawan, using M-16 rifles as he rode his motorcycle, killing him and damaging h...
02/13/2010: On Saturday around 0000 midnight, in Bangkok, a grenade was detonated by unidentified militants on the campus of a university near the prime minister's office. Three vehicles were damaged but no ca...
02/13/2010: On Saturday night, in Yarang, Pattani, Thailand, gunmen fired upon civilians sitting outside a residence, killing one civilian, wounding four others, and damaging the home. The dead man was identified a...
02/15/2010: On Monday evening, a 36 year-old Muslim man was shot and killed while working at rubber farm in the Rangae district of Narathiwat, Thailand. No group claimed responsibility.
02/15/2010: On Monday night, a 32-year-old militant was shot as insurgents ambushed a group of government officials and their guards on a road in Rangae, Narathiwat, Thailand. The district officials were returning...

Figure 17: Screenshot of "summary" field

Text Mining, in the form of Text categorisation, will be carried out to convert the unstructured free text into usable structured data which will be used in the predictive model. This will be done using the IBM SPSS Text Analytics tool.

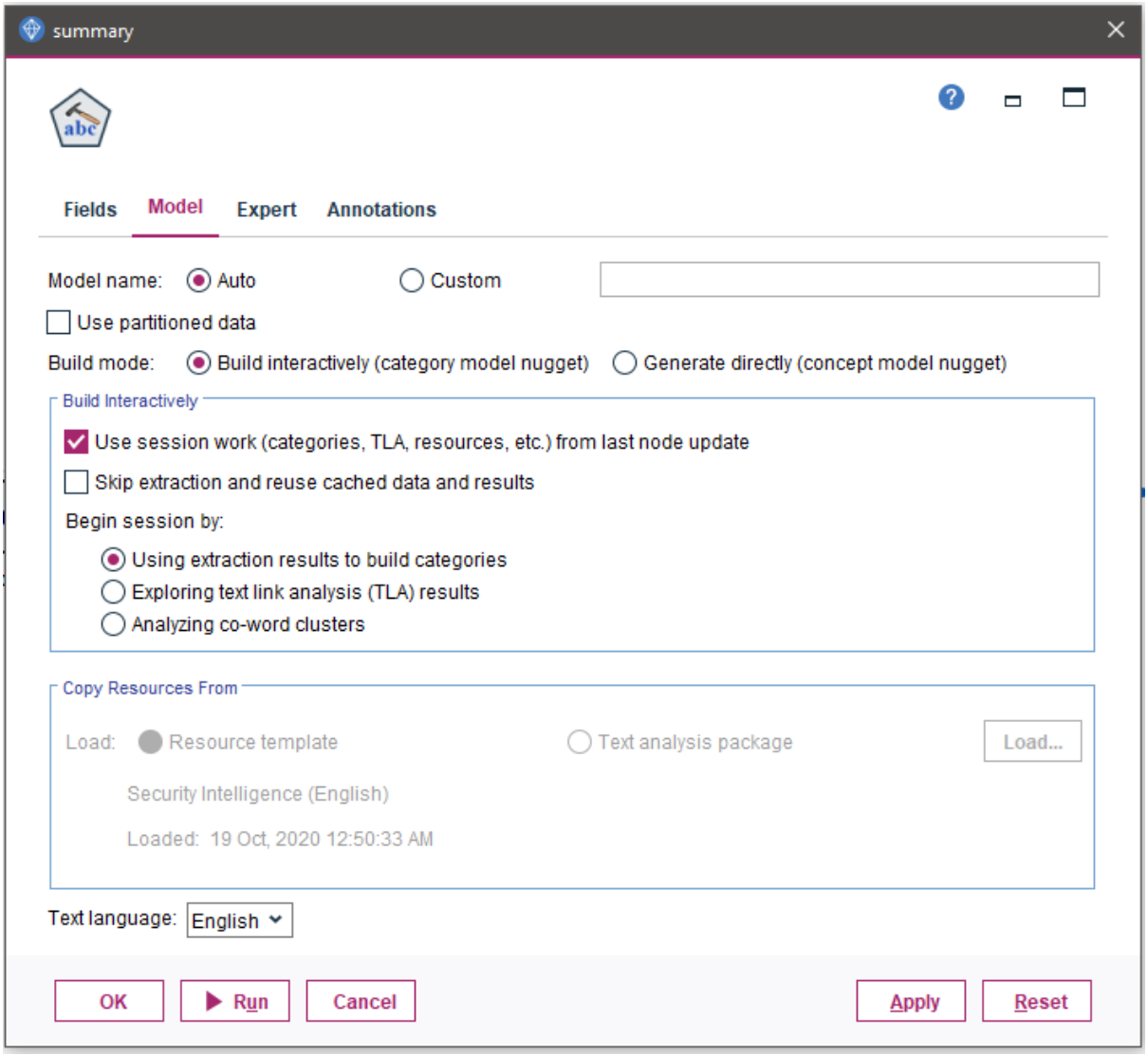


Figure 18: Settings used for Text Mining Node

Figure 18 shows the settings used for the Text Mining Node. Two resource templates were used and compared, namely the Basic Resource (English) and the Security Intelligence (English). Figure 18 shows the result of building the text mining model using the Basis Resource Template and Figure 19 shows the results of using the Security Intelligence resource template. There are only two types produced, as compared to 23 types for the latter. The Security Intelligence (English) resource template is also preferred due to the nature of the study, which revolves around security and terrorism. Thus, the Security Intelligence resource template is chosen for the model.

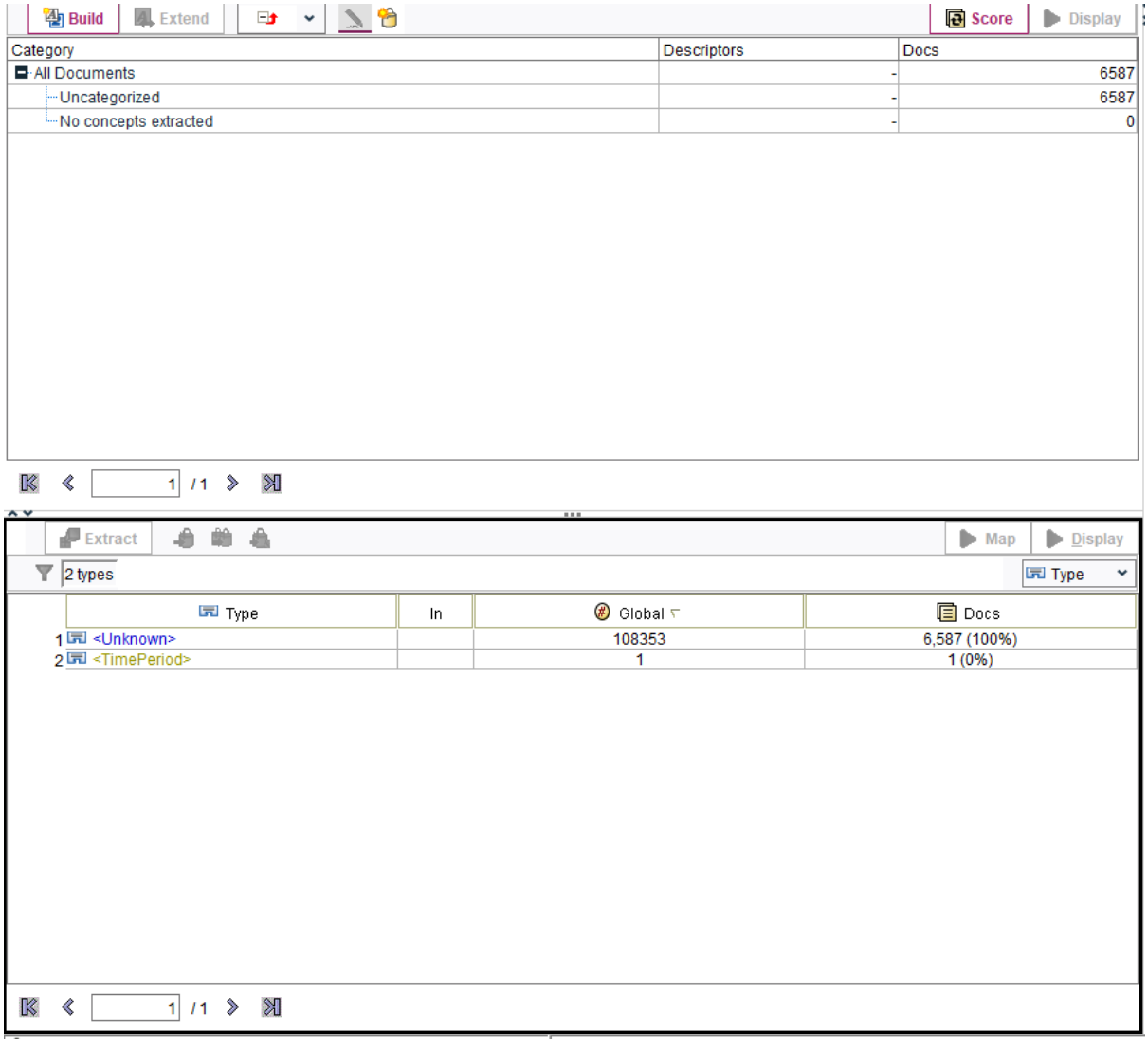
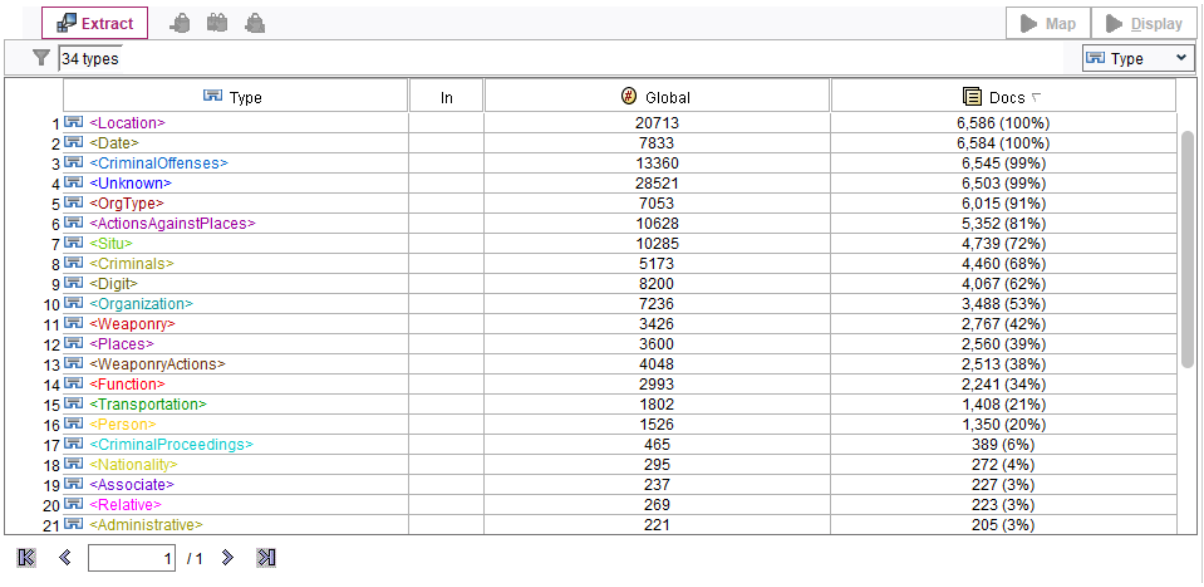
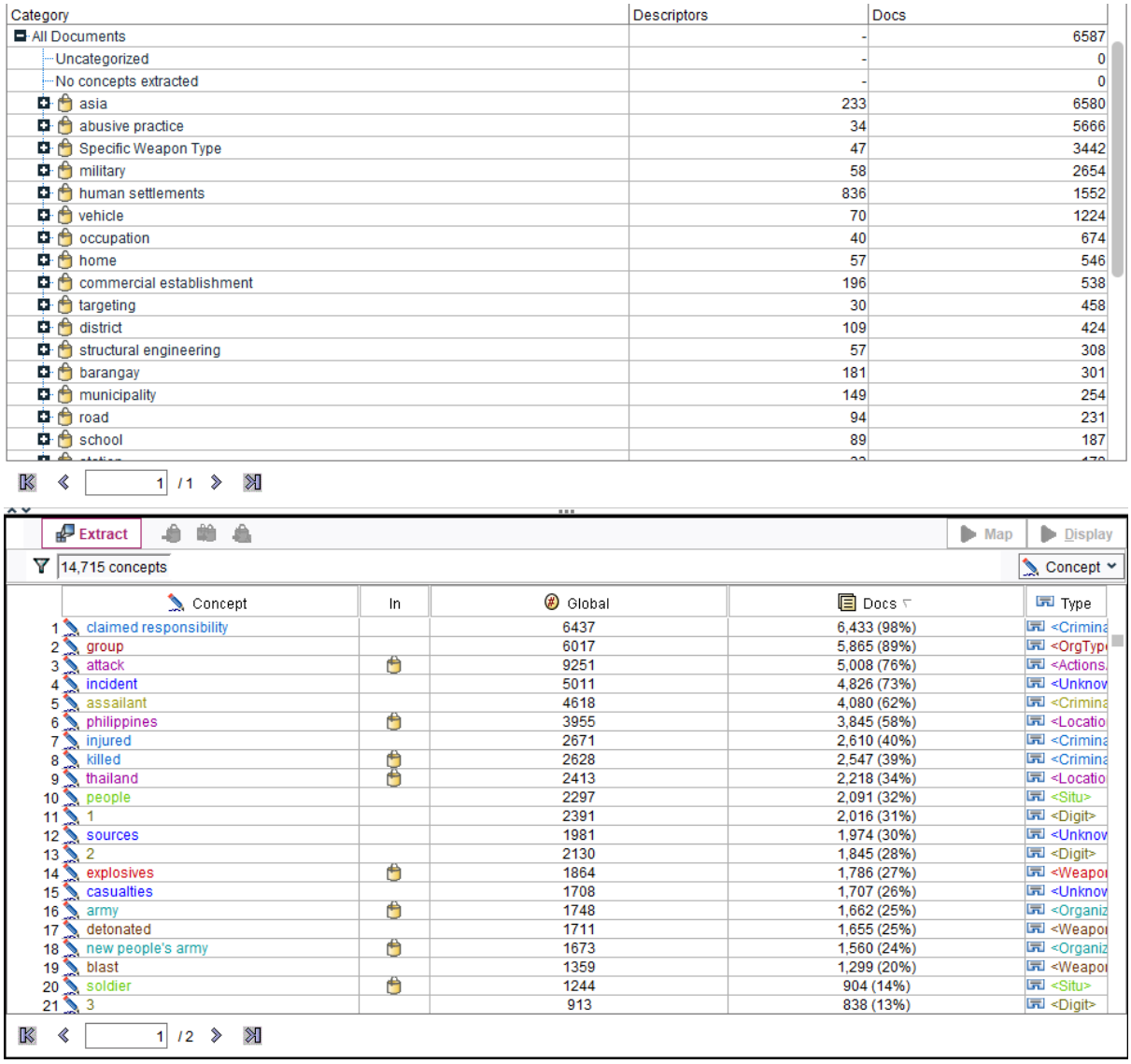


Figure 19: Result of Using Basis Resource Template



	Type	In	Global	Docs
1	<Location>		20713	6,586 (100%)
2	<Date>		7833	6,584 (100%)
3	<CriminalOffenses>		13360	6,545 (99%)
4	<Unknown>		28521	6,503 (99%)
5	<OrgType>		7053	6,015 (91%)
6	<ActionsAgainstPlaces>		10628	5,352 (81%)
7	<Situ>		10285	4,739 (72%)
8	<Criminals>		5173	4,460 (68%)
9	<Digit>		8200	4,067 (62%)
10	<Organization>		7236	3,488 (53%)
11	<Weaponry>		3426	2,767 (42%)
12	<Places>		3600	2,560 (39%)
13	<WeaponryActions>		4048	2,513 (38%)
14	<Function>		2993	2,241 (34%)
15	<Transportation>		1802	1,408 (21%)
16	<Person>		1526	1,350 (20%)
17	<CriminalProceedings>		465	389 (6%)
18	<Nationality>		295	272 (4%)
19	<Associate>		237	227 (3%)
20	<Relative>		269	223 (3%)
21	<Administrative>		221	205 (3%)

Figure 20: Result of Using Security Intelligence Resource Template



Category	Descriptors	Docs
All Documents	-	6587
Uncategorized	-	0
No concepts extracted	-	0
asia	233	6580
abusive practice	34	5666
Specific Weapon Type	47	3442
military	58	2654
human settlements	836	1552
vehicle	70	1224
occupation	40	674
home	57	546
commercial establishment	196	538
targeting	30	458
district	109	424
structural engineering	57	308
barangay	181	301
municipality	149	254
road	94	231
school	89	187

1 / 1

Extract | Map | Display

14,715 concepts | Concept

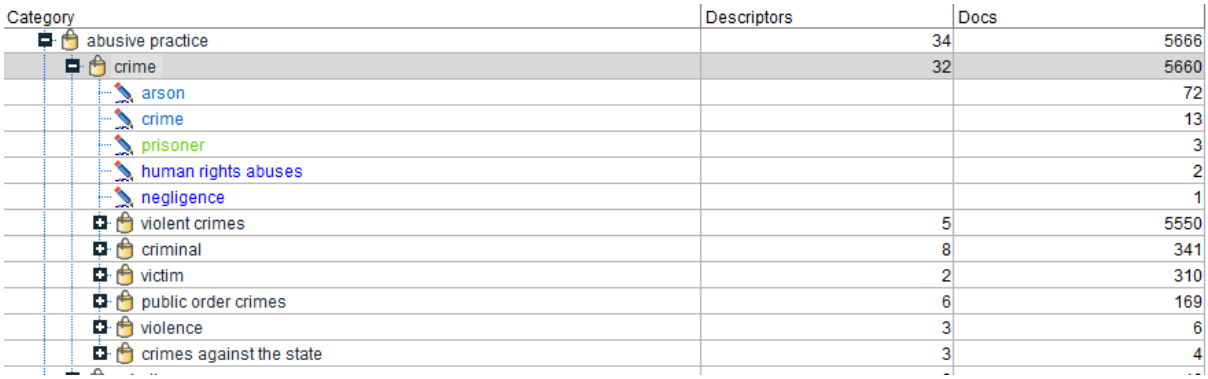
	Concept	In	Global	Docs	Type
1	claimed responsibility		6437	6,433 (98%)	<Crimina
2	group		6017	5,865 (89%)	<OrgTyp
3	attack		9251	5,008 (76%)	<Actions
4	incident		5011	4,826 (73%)	<Unknov
5	assailant		4618	4,080 (62%)	<Crimina
6	philippines		3955	3,845 (58%)	<Locatio
7	injured		2671	2,610 (40%)	<Crimina
8	killed		2628	2,547 (39%)	<Crimina
9	thailand		2413	2,218 (34%)	<Locatio
10	people		2297	2,091 (32%)	<Situ>
11	1		2391	2,016 (31%)	<Digit>
12	sources		1981	1,974 (30%)	<Unknov
13	2		2130	1,845 (28%)	<Digit>
14	explosives		1864	1,786 (27%)	<Weapo
15	casualties		1708	1,707 (26%)	<Unknov
16	army		1748	1,662 (25%)	<Organiz
17	detonated		1711	1,655 (25%)	<Weapo
18	new people's army		1673	1,560 (24%)	<Organiz
19	blast		1359	1,299 (20%)	<Weapo
20	soldier		1244	904 (14%)	<Situ>
21	3		913	838 (13%)	<Digit>

1 / 2

Figure 21: Categories Created using Concepts Extracted

After running the node, the model is built, and the model is scored. Figure 21 above shows the scored results. The third category is renamed to Specific Weapon Type to better describe the category. From the results of text categorisation model, three categories will be selected to provide a better resolution for the predictive model.

Firstly, under the “abusive practice” category lies a sub-category named “crime”, which is shown in Figure 22 below.

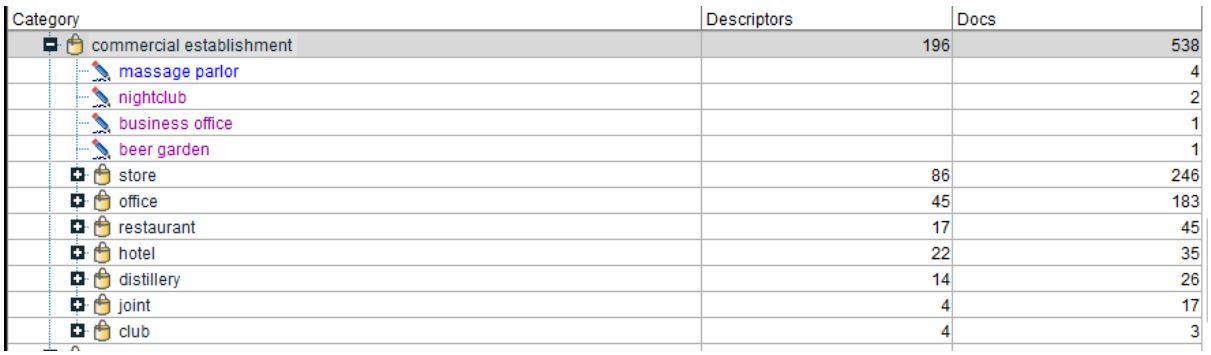


Category	Descriptors	Docs
abusive practice	34	5666
crime	32	5660
arson		72
crime		13
prisoner		3
human rights abuses		2
negligence		1
violent crimes	5	5550
criminal	8	341
victim	2	310
public order crimes	6	169
violence	3	6
crimes against the state	3	4

Figure 22: Crime Sub-category

This sub-category provides a better resolution as it provides insights on the crimes committed and crimes that influenced a certain retaliation (in the form of terrorism). This is not provided as a form of structured data and this category will be able to cover that lack of information. This category will provide more insights for the predictive model.

Secondly, the category "commercial establishment" will be used for the predictive modelling. Figure 23 shows the summary of this category.

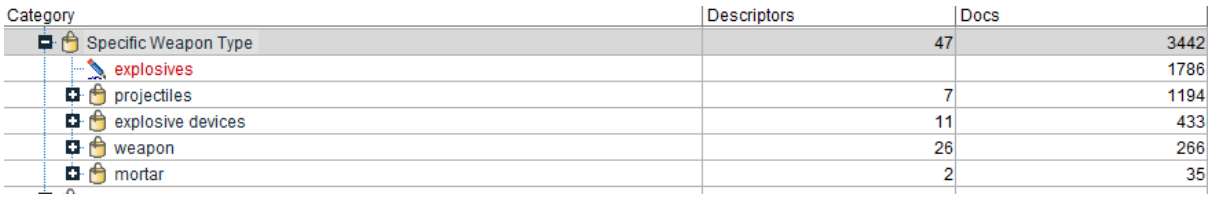


Category	Descriptors	Docs
commercial establishment	196	538
massage parlor		4
nightclub		2
business office		1
beer garden		1
store	86	246
office	45	183
restaurant	17	45
hotel	22	35
distillery	14	26
joint	4	17
club	4	3

Figure 23: Commercial Establishment Category

Amongst the structured data, there was no mention of the type of establishment which attack occurred. For example, the "massage parlour" was mentioned more than "business office" in the dataset, which might further suggest that it could be a potential hotspot for such attacks. The inclusive of this category in the predictive model might provide more insights on the type of commercial establishment that requires more protection against attacks.

Lastly, the renamed category "Specific Weapon Type" will be used in the predictive model. The summary is shown is Figure 24 below.



Category	Descriptors	Docs
Specific Weapon Type	47	3442
explosives		1786
projectiles	7	1194
explosive devices	11	433
weapon	26	266
mortar	2	35

Figure 24: Specific Weapon Type Category

Although one of the structured fields (weaptype_txt) provides a weapon type field, the field provides a general weapon type instead of a specific weapon. Knowing the specific type of weapon such as handgun or a biological weapon, might provide more insights to predicting the success rate of a terrorist attack.

3.5.2 Data Fields for Predictive model

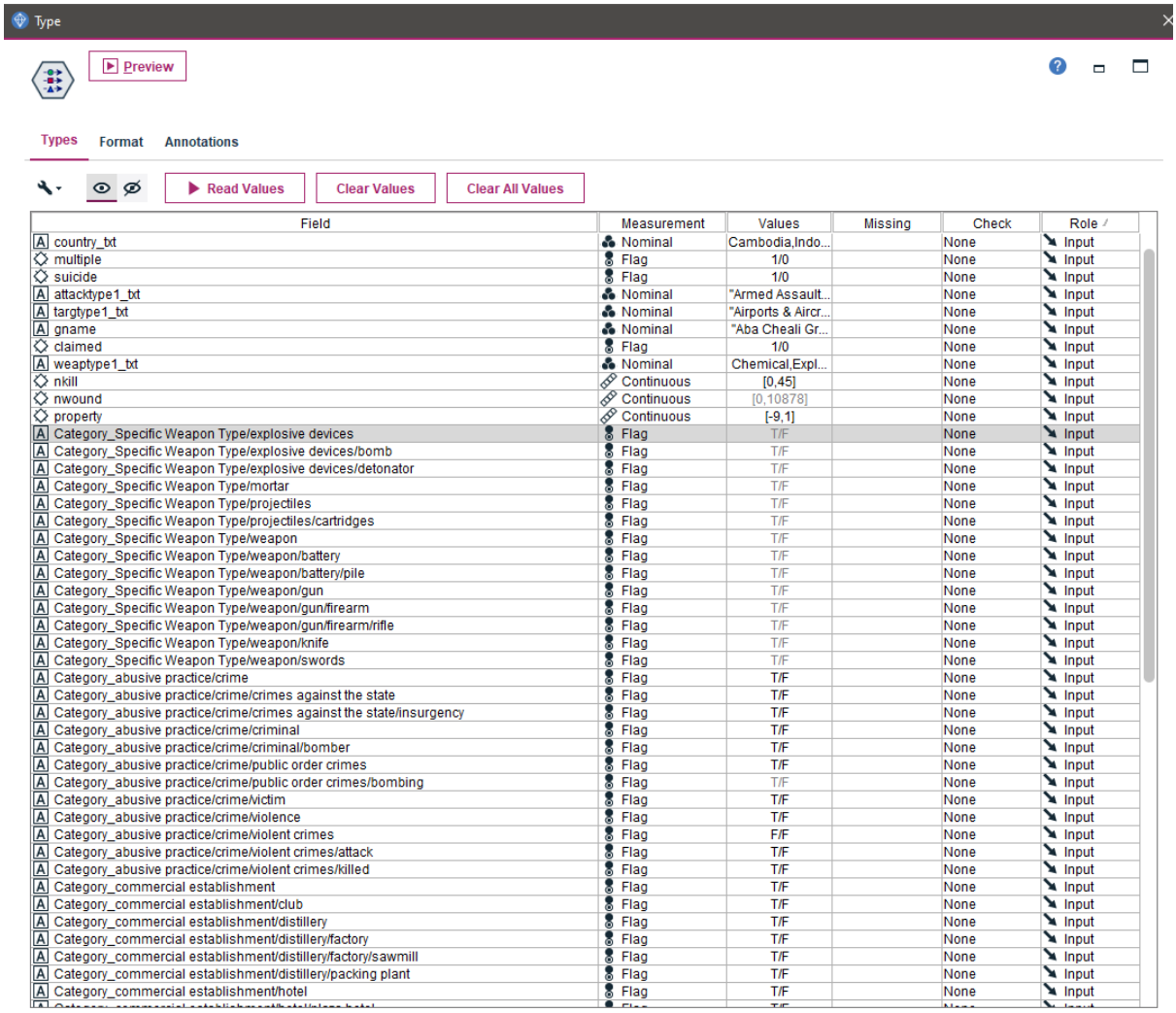
Table 3 below summarises the fields which will be used for the predictive model. Figure 25 and 26 below provides the expanded list of fields which will be used for the model.

Table 3

Fields to be Used for Predictive Model

	Field Name	**Description**	**Data Type**	**Role**
1	country_txt	Name of Country	Nominal	Input
2	multiple	Multiple attacks occurred	Flag	Input
3	suicide	Suicide bombing occurred	Flag	Input
4	attacktype1_txt	Type of Attack	Nominal	Input
5	nkill	Number of Fatalities	Continuous	Input
6	nwound	Number of Wounded	Continuous	Input
7	property	Property Damage occurred	Flag	Input
8	Text-mined category: Specific Weapon Type (13 fields)	Specific weapon types used	Flag	Input

9	Text-mined category: crime (13 fields)	Crimes that occurred	Flag	Input
10	Text-mined category: Commercial Establishment (36 fields)	Type of commercial establishment that attack occurred in	Flag	Input
11	success	If attack is considered as success	Flag	Target



Field	Measurement	Values	Missing	Check	Role
country_txt	Nominal	Cambodia,Indo...		None	Input
multiple	Flag	1/0		None	Input
suicide	Flag	1/0		None	Input
attacktype1_txt	Nominal	"Armed Assault...		None	Input
targtype1_txt	Nominal	"Airports & Aircr...		None	Input
gname	Nominal	"Aba Cheali Gr...		None	Input
claimed	Flag	1/0		None	Input
weaptype1_txt	Nominal	Chemical,Expl...		None	Input
nkill	Continuous	[0,45]		None	Input
nwound	Continuous	[0,10878]		None	Input
property	Continuous	[-9,1]		None	Input
Category_Specific Weapon Type/explosive devices	Flag	T/F		None	Input
Category_Specific Weapon Type/explosive devices/bomb	Flag	T/F		None	Input
Category_Specific Weapon Type/explosive devices/detonator	Flag	T/F		None	Input
Category_Specific Weapon Type/mortar	Flag	T/F		None	Input
Category_Specific Weapon Type/projectiles	Flag	T/F		None	Input
Category_Specific Weapon Type/projectiles/cartridges	Flag	T/F		None	Input
Category_Specific Weapon Type/weapon	Flag	T/F		None	Input
Category_Specific Weapon Type/weapon/battery	Flag	T/F		None	Input
Category_Specific Weapon Type/weapon/battery/pile	Flag	T/F		None	Input
Category_Specific Weapon Type/weapon/gun	Flag	T/F		None	Input
Category_Specific Weapon Type/weapon/gun/firearm	Flag	T/F		None	Input
Category_Specific Weapon Type/weapon/gun/firearm/rifle	Flag	T/F		None	Input
Category_Specific Weapon Type/weapon/knife	Flag	T/F		None	Input
Category_Specific Weapon Type/weapon/swords	Flag	T/F		None	Input
Category_abusive practice/crime	Flag	T/F		None	Input
Category_abusive practice/crime/crimes against the state	Flag	T/F		None	Input
Category_abusive practice/crime/crimes against the state/insurgency	Flag	T/F		None	Input
Category_abusive practice/crime/criminal	Flag	T/F		None	Input
Category_abusive practice/crime/criminal/bomber	Flag	T/F		None	Input
Category_abusive practice/crime/public order crimes	Flag	T/F		None	Input
Category_abusive practice/crime/public order crimes/bombing	Flag	T/F		None	Input
Category_abusive practice/crime/victim	Flag	T/F		None	Input
Category_abusive practice/crime/violence	Flag	T/F		None	Input
Category_abusive practice/crime/violent crimes	Flag	F/F		None	Input
Category_abusive practice/crime/violent crimes/attack	Flag	T/F		None	Input
Category_abusive practice/crime/violent crimes/killed	Flag	T/F		None	Input
Category_commercial establishment	Flag	T/F		None	Input
Category_commercial establishment/club	Flag	T/F		None	Input
Category_commercial establishment/distillery	Flag	T/F		None	Input
Category_commercial establishment/distillery/factory	Flag	T/F		None	Input
Category_commercial establishment/distillery/factory/sawmill	Flag	T/F		None	Input
Category_commercial establishment/distillery/packing plant	Flag	T/F		None	Input
Category_commercial establishment/hotel	Flag	T/F		None	Input

Figure 25: Extended Fields for Predictive Modelling (1st Half)

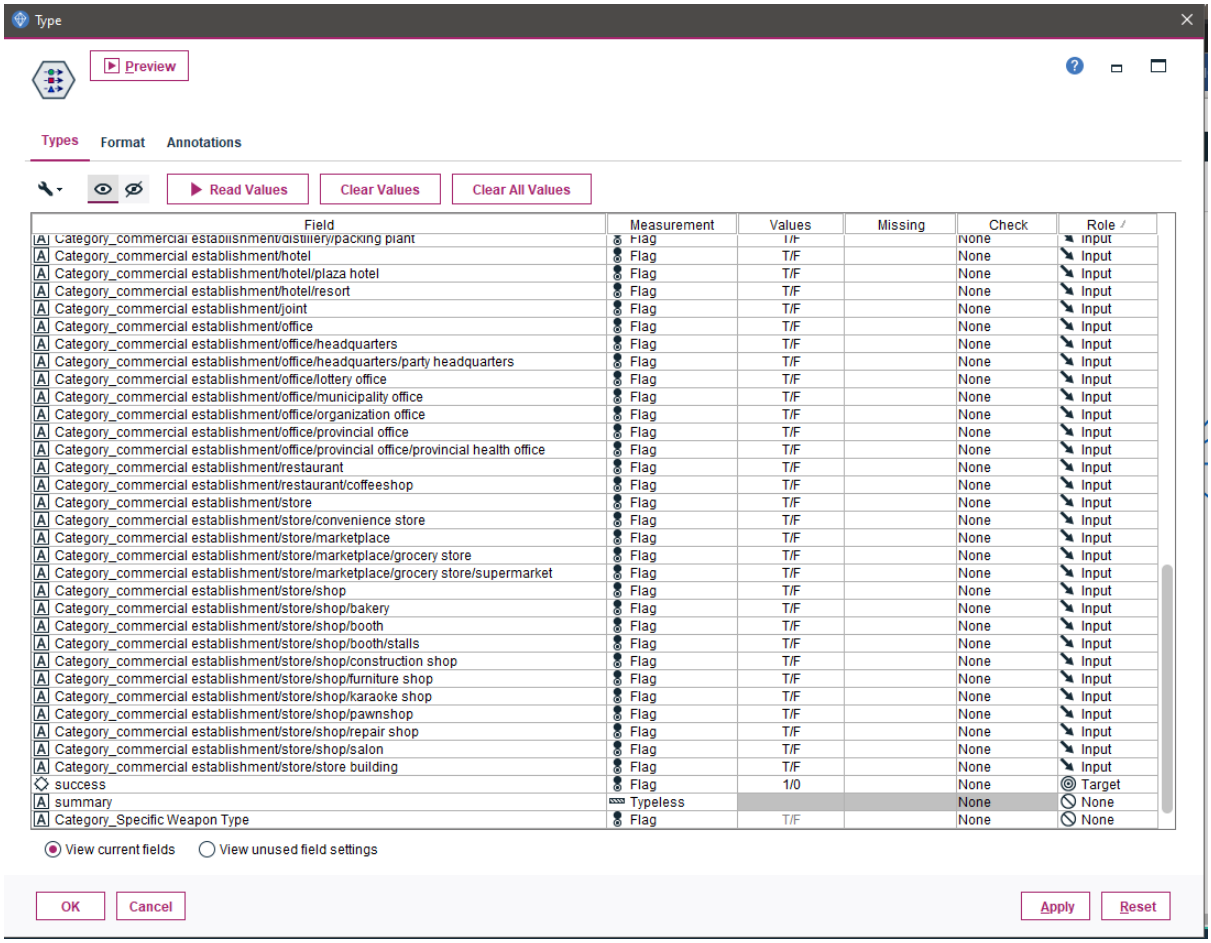


Figure 26: Extended Fields for Predictive Modelling (2nd Half)

3.5.3 Predictive Modelling

After partitioning, three types of supervised classification model will be used to predict the success rate of a terrorist attack. They are the C5.0 Classification model (Pang & Gong, 2009)0, Chi-square Automatic Interaction Detector (CHAID) (Kass, 1980) and the Classification And Regression Tree (CART) (Breiman, Friedman, Stone, & Olshen, 1984) will be used. Each of these methods creates a Decision Tree which will highlight the probability of a factor that will cause the success or non-success of a terrorist attacks (Kingsford & Salzberg, 2008)

The C5.0 Node, CART Node and CHAID Node will be used after the Partition Node to create the predictive model. An Analysis node will be used to determine the predictive ability of each model created. The resulting SPSS stream is shown in Figure 27 below.

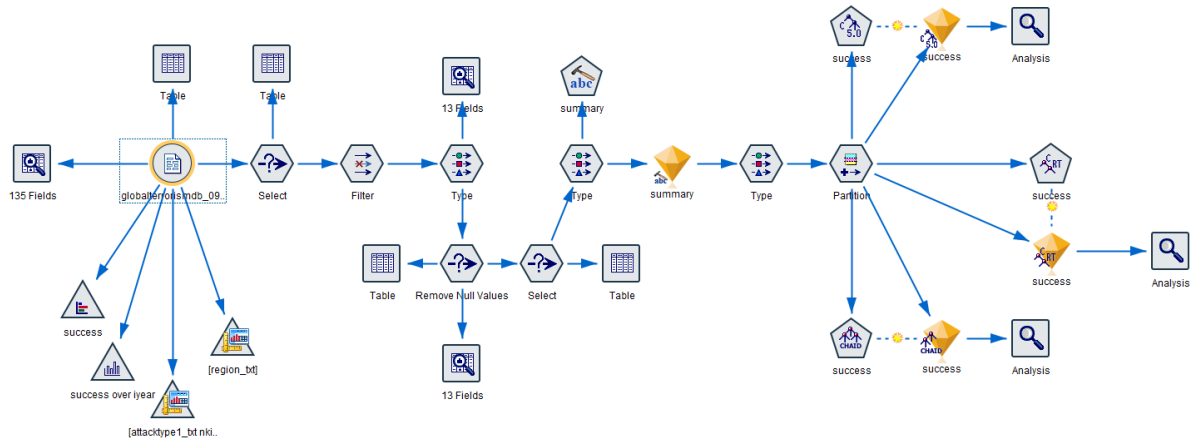


Figure 27: Final SPSS Stream

Results of model created

C5.0

Comparing $C-success with success

'Partition'	1_Training		2_Testing	
Correct	4,273	93.11%	1,852	92.69%
Wrong	316	6.89%	146	7.31%
Total	4,589		1,998	

Coincidence Matrix for $C-success (rows show actuals)

'Partition' = 1_Training	0	1
0	574	144
1	172	3,699

'Partition' = 2_Testing	0	1
0	232	62
1	84	1,620

Figure 28: Results of C5.0 Model

CHAID

Comparing $R-success with success

'Partition'	1_Training		2_Testing	
Correct	4,201	91.54%	1,813	90.74%
Wrong	388	8.46%	185	9.26%
Total	4,589		1,998	

Coincidence Matrix for $R-success (rows show actuals)

'Partition' = 1_Training	0	1
0	550	168
1	220	3,651

'Partition' = 2_Testing	0	1
0	220	74
1	111	1,593

Figure 29: Results of CHAID Model

CART

Comparing $R-success with success

'Partition'	1_Training		2_Testing	
Correct	4,137	90.15%	1,818	90.99%
Wrong	452	9.85%	180	9.01%
Total	4,589		1,998	

Coincidence Matrix for $R-success (rows show actuals)

'Partition' = 1_Training	0	1
0	588	130
1	322	3,549

'Partition' = 2_Testing	0	1
0	243	51
1	129	1,575

Figure 30: Results of CART Model

3.6 Evaluation

3.6.1 Selecting the Champion Model

Table 4 below summarise the results of each model. The accuracy and hit rate are calculated based on the confusion matrix of each model's testing dataset,

Table 4

Results of Models

	Testing Dataset		
	C5.0	**CHAID**	**CART**
Overall Accuracy (%)	92.69%	90.74%	90.99%
Accuracy for Not Successful Attack (%)	78.91%	74.83%	82.65%
Accuracy for Successful Attack (%)	95.07%	93.49%	92.43%
Hit Rate for Not Successful Attack (%)	73.42%	66.47%	65.32%
Hit Rate for Successful Attack (%)	96.31%	95.56%	96.86%

Based on the summary given in Table 4, the champion model to be selected is the C5.0 model. C5.0 has the highest overall accuracy, highest accuracy for successful attacks amongst the three models. Although CART has a higher accuracy rate for not successful attack and higher hit rate for successful attack, the objective of the study emphasises on being able to predict successful attacks rather than non-successful attacks. The hit rate of C5.0 Model is only slightly lesser

than CART (<1%) while the accuracy for successful attack is higher (>2%). Thus, the C5.0 is chosen as the champion model

3.6.2 Decision Tree for CART Model

Figure 31 shows the predictor importance of each input for the model and Figure 32 below shows the Decision Tree generated for the CART model.

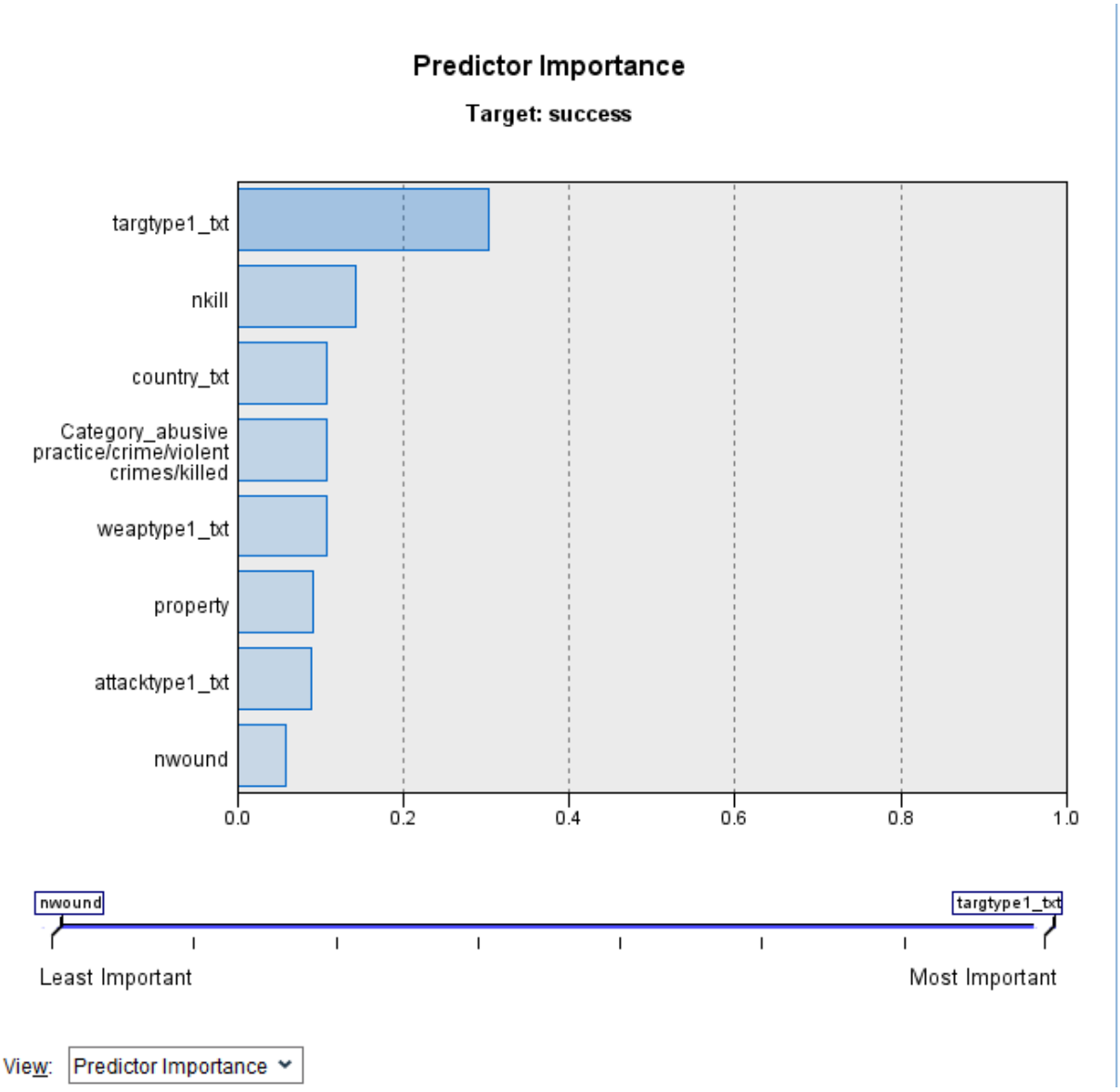


Figure 31: Predictor Importance of Model Inputs

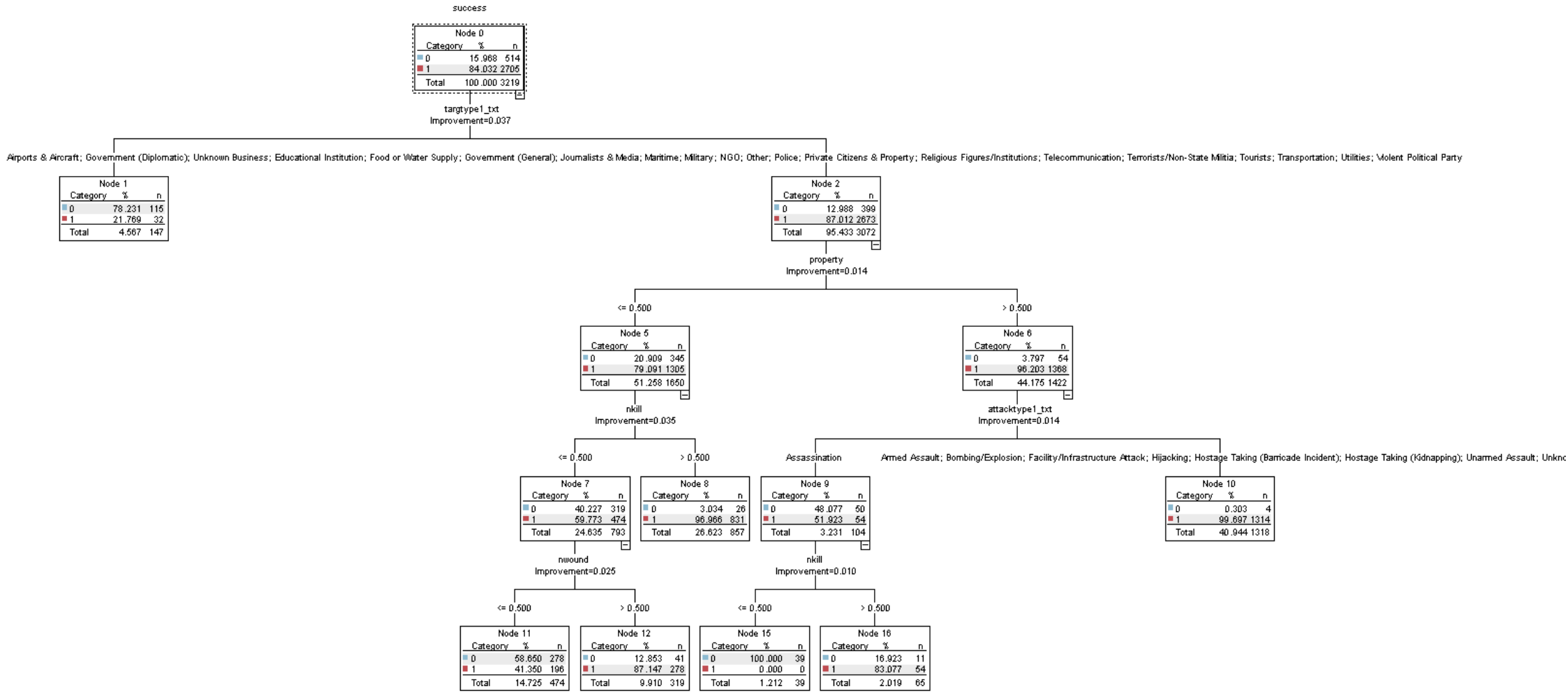


Figure 32: Decision Tree Created

3.6.3 Insights from Model

From Figure 31, it can be illustrated that the top three important predictors for the success of attack are the target type, number of kill and the third being tied between country, killing and weapon type.

From the decision tree in Figure 32, the following observations were made:

From Node 1, if the target type is Airport & Aircraft or Government (Diplomatic), the attack is likely to be not successful (78.231%)

From Node 6, if the target type is not Airport or Government (Diplomatic) and there is property damage present, the attack is likely to be successful (96.2%)

From Node 10, if the target type is not Airport or Government (Diplomatic) and the attack is not an assassination but there is property damage, the attack is most likely to be successful (99.7%).

From Node 8, if the target type is not Airport or Government (Diplomatic) and there is no property damage but there are fatalities, the attack is likely to be successful (96.97%).

From Node 12, if the target type is not Airport or Government (Diplomatic) and there is no property damage, and no fatalities but there are wounded individuals, the attack is likely to be successful (87.15%).

However, the model can be further improved by using flag values of target types and attack types instead of categorical values. Due to the usage of categorical values, the decision tree classified target types into two broad spectrums, with “Airport & Aircraft/Government (Diplomatic)” as one half and the others as the other half. Using flag value and separating each of these target type into a single flag category will allow the model to produce more meaningful insights.

3.7 Deployment of Model

3.7.1 Text Mining Model

The text mining model can be deployed on other databases to provide more insights for the predictive model. However, the text mining model requires constant monitoring to determine the type of text which will be mined. For example, if the database contains new synonyms

and words, there are to be included into the model and properly categorised. The maintenance of this model will be needed to further the enhance the accuracy and effectiveness of this model.

3.7.2 Predictive Model

The predictive model will be deployed to provide additional insights on counterterrorism. Insights, such as target types are of higher risk, will be considered and possibly more protection will be given to said target types. Another insight based on Node 10, states that attacks which are not assassinations but caused property damage are 99.7% likely to be considered a successful attack. This insight will provide authorities with more incentive to prevent property damage in a suspected terror attack to prevent the attack from being successful.

This model will be constantly scored as more newer data is tested. This score should be constantly monitored, and the model performance must be maintained at an accepted score value to ensure the operational effectiveness of the model. More variables can be added to provide more insights on future counterterrorism tactics.

4 Summary and Future Work

A predictive classification method using structured data combined with text categorisation outputs is introduced in this paper to predict the success of terrorist attacks in Southeast Asia from 2010 to 2018. This is achieved by using data from the Global Terrorism Database (GTD), created and maintained by the National Consortium for the Study of Terrorism And Responses of Terrorism (SMART). Using text categorisation, additional fields are created and used as input in the predictive model. The predictive model used is a supervised decision tree, with the three types of algorithm used, namely the C5.0, CHAID and CART. These models are evaluated and from the comparisons of the model's accuracy and hit rates, the C5.0 algorithm was selected as the champion model. The model can predict successful attacks from a testing dataset with an accuracy rate of 95.07%.

Using the inputs from GTD and text categorisation, the decision tree produces multiple meaningful insights. These insights from the model can be then further studied and deployed into real-life counterterrorism tactics.

One benefit of this model is that it can produce structured data to be used in the predictive model from unstructured free-form text. One downside of this model is its reliance on historical data, which means that it is unable to predict any new types of terrorist attack form, such as cyberattacks. Considering this drawback, future work on this topic should consider the use of live web content such as Social Media or discussion forums as their dataset, using web scrapers or social media application programming interface (API), as the data is more current.

References

Breiman, L., Friedman, J., Stone, C. J., & Olshen, R. (1984). *Classification and Regression Trees (Wadsworth Statistics/Probability).* Belmont: Wadsworth Publishing.

Feldman, R., & Dagan, I. (1995, August). Knowledge Discovery in Textual Databases (KDT). *KDD'95: Proceedings of the First International Conference on Knowledge Discovery and Data Mining*, 112-117.

Feldman, R., Fresko, M., Hirsh, H., Aumann, Y., Liphstat, O., Schler, Y., & Rajman, M. (1998). Knowledge Management: A Text Mining Approach. *Proceedings of the 2nd International Conference on Practical Aspect of Knowledge Management (PAKM98), 1998.*

Gupta, V., & Lehal, G. S. (2009, August). A Survey of Text Mining Techniques and Applications. *Journal of Emerging Technologies in Web Intelligence, Vol. 1, No. 1*, 60-76.

Hannover Re. (2020). *Political Violence / Terrorism.* Hanover: Hannover Re.

Kass, G. V. (1980). An Exploratory Technique for Investigating Large Quantities of Categorical Data. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, 119-127.

Kingsford, C., & Salzberg, S. L. (2008). What are decision trees? *Nature Biotechnology*, 1011-1013.

Pang, S.-l., & Gong, J.-z. (2009). C5.0 Classification Algorithm and Application on Individual Credit Evaluation of Banks. *Systems Engineering - Theory & Practice*, 94-104.

Patra, S. (2019). Economic Consequences of Terrorism in South Asia. In R. C. Das, *The Impact of Global Terrorism on Economic and Political Development* (pp. 179-190). Emerald Publishing Limited.

Ritchie, H., Hasell, J., Appel, C., & Roser, M. (2013, July). *Terrorism*. Retrieved from Our World in Data: https://ourworldindata.org/terrorism

Rubin, G. J., & Wessely, S. (2013). The psychological and psychiatric effects of terrorism: lessons from London. *The Psychiatric clinics of North America, 36(3)*, 339-350.

Ruiz Estrada, M., Park, D., & Khan, A. (2018, December). The impact of terrorism on economic performance: The case of Turkey. *Economic Analysis and Policy, 60*, 78-88.

START. (2019). *Global Terrorism Database (GTD) Codebook: Inclusion Criteria and Variables.* Maryland: CHC Global.

START. (2020, June). *Global Terrorism Database (GTD) | START.umd.edu*. Retrieved from National Consortium for the Study of Terrorism And Responses to Terrorism (START): https://www.start.umd.edu/research-projects/global-terrorism-database-gtd

Thuraisingham, B. (2003). *Web Data Mining Technologies and Their Applications in Business Intelligence and Counter-terrorism.* Boca Raton: CRC Press.

Wirth, R., & Hipp, J. (2000, April). CRISP-DM: Towards a standard process model for data mining. *Proceedings of the 4th international conference on the practical applications of knowledge discovery and data mining*, pp. 29-39.

Zamin, N., & Oxley, A. (2012). Unapparent Information Revelation for Counterterrorism: Visualizing Associations using a Hybrid Graph-based Approach. *Knowledge Management International Conference (KMICe) 2012, Johor Bahru, Malaysia, 4 – 6 July 2012*, (pp. 30-37).

Zanasi, A. (2007). *Text Mining and its Applications to Intelligence, CRM and Knowledge Management.* WIT Press.

Zanasi, A. (2009). Virtual weapons for real wars: Text mining for national security. *Proceedings of the International Workshop on Computational Intelligence in Security for Information Systems CISIS'08, Advances in Soft Computing*, (pp. 53-60).