Decoding the Black Box: Interpretable Methods for Post-Incident Counter-terrorism Investigations

Ankit Tewari
Universitat Politecnica de Cataluña (UPC), Universitat of Barcelona (UB)
Barcelona, Spain
mail.ankit.tewari@gmail.com

Figure 1: National Security Guards (NSG) commandos during a mock drill at the Delhi Metro Railway Station

ABSTRACT

Predicting perpetrators in incidents of terrorism by detecting patterns in terror attacks is an extremely complex problem in the domain of homeland security. It forms a crucial part pf post-incident investigations which relies significantly on OSINT data sources and can be considerably simplified with the involvement of machine learning algorithms. However, one of the major challenges that needs to be addressed is our ability to explain such patterns in a human friendly manner and not just the decisions based on such patterns. This not only allows for the verification of the captured patterns with domain experts but also makes the derived intelligence actionable by providing unique pattern in attacks carried out by a terrorist outfit. We present a case study where we first attempt to predict the perpetrators in incidents of terrorism and then use a variety of interpretability mechanisms to present those patterns in a human understandable manner.

WebSci '20 Companion, July 6–10, 2020, Southampton, United Kingdom

ACM ISBN 978-1-4503-7994-6/20/07…$15.00
https://doi.org/10.1145/3394332.3402839

KEYWORDS

homeland security, open source intelligence, artificial intelligence, interpretability, transparency, trustworthiness

ACM Reference Format:
Ankit Tewari. 2020. Decoding the Black Box: Interpretable Methods for Post-Incident Counter-terrorism Investigations. In *12th ACM Conference on Web Science (WebSci '20 Companion), July 6–10, 2020, Southampton, United Kingdom.* ACM, New York, NY, USA, 6 pages. https://doi.org/10.1145/3394332.3402839

1 INTRODUCTION

In the contemporary world, democracies across the globe are facing a number of emerging threats. Terrorism and the terrorist threat is one of those emerging threats that not only affects basic fundamental rights but also undermines the credibility of democratic governments.

Terrorism challenges a democracy in both direct and indirect ways. While a direct challenge is described as the one in which a terrorist outfit challenges the safety and security of citizens in a democracy which is considered a fundamental right. The indirect challenge to democracy occurs from the state's response to the terrorism by trying to deliver themselves security by the deeply misguided means of tampering with their own civil liberties. They do it in order to make the job of the security services easier, by assuming greater powers. These harms to the fabric of a free society have longer effects, and in the long run worse effects, than terrorists' bombs.

The response of a state against terrorism is generally studied in terms of pre-incident measures and post-incident counter-measures. While there is a lot of technological progress in the context of pre-incident measures, there is still need to explore the possibility of technology driven options for post-incident measures. In fact, data driven post-incident analysis of acts of terrorism has the potential to learn about ideology, strategy and operational tactics of the corresponding terrorist outfit. Such an analysis becomes crucial while planning for an effective policy and deployment of counter-measures in order to suppress the growth and influence of the terrorist outfits operating in that environment.

Since, every counter terrorism strategy is improvised based on the lessons learnt from the past. It suggests the need of a system that can develop an understanding about the adversary. The initial idea of such a counter-terrorism decision support system is to analyze large chunks of data about conflicts that happened in the past and create a behavioural model for the terrorist outfits involved in those conflicts. The goal of such a system is two fold primarily. On the one hand, it helps to identify the ideology, goals, targets, tactics etc. of those terrorist outfits which in turn helps in assessing its capabilities (which eventually allows to carry out a variety of tasks such as predicting future vulnerable places, communities etc). On the other hand, it also allows to associate a terrorist outfit with an event that may happen in future by comparing the profile of the extremist outfit with the measured attributes of the incident based on open source intelligence data primarily. In other words, the system tries to generate profiles of such terrorist outfits using machine learning.

2 SUPERVISED PROFILING AND INCIDENT EXPLANATION SYSTEM (SPIES)

In this work, we will describe an ongoing project called the Supervised Profiling and Incident Explanation System (SPIES). This system can be considered to act as a situational awareness system for both incidents happened in the past as well as the ones anticipated or will occur in the future. The goal of the system is to perform data fusion between manually extracted data describing attributes of a conflict and news reports describing that conflict. This data is then used to create predictive behavioural models to estimate the probability of involvement of a particular terrorist group under consideration in a future attack. In the context of the present work, we will begin with an open source database called Global Terrorism database (GTD) [2]. The idea of the work will be to use this open source dataset to build machine learning models that can capture the behavioural trends of the operations of the extremist groups under our study and then use these models to determine the perpetrators based on information available through Open Source Intelligence (OSINT) sources in the future.

Although, it seems to be a feasible task depending on the open source data and availability of a number of machine learning algorithms with accuracy and speed observed as never before, however, as widely discussed and criticised, these highly accurate but complex algorithms come at the cost of interpretability making the predictions risky to believe. Therefore, we will make attempts to explain the behavior of our models using different ideas of interpretability. In order to begin with and for the sake of simplicity, we had begun with only two terrorist outfits "Lashkar-e-taiba (LET)" and "Jaish-e-Mohammad (JEM)", both based in South Asia and our dataset specifically focuses on their acts of terror in the Jammu and Kashmir state of India. Based on a simple model built built using Random Forest Algorithm using a set of 18 features and the name of the extremist outfit as the label to train the model, our model achieved an accuracy of 85.45% in the first attempt and we present a classification report below in the figure 2

	precision	recall	f1-score	support
0	0.25	1.00	0.40	2
1	1.00	0.89	0.94	53
accuracy			0.89	55
macro avg	0.62	0.94	0.67	55
weighted avg	0.97	0.89	0.92	55

Figure 2: Classification Report: Based on Random Forest Model

Since, in this work, our focus is mostly directed on explaining the decision making of the model we trained above, we do not make further attempts to tune the parameters to obtain higher accuracy. It allows us understand the behavioural attributes of the extremist outfits under our consideration captured by the baseline model which is extremely fundamental in nature. The features used for modelling along with their names and description can be found in the Global Terrorism Codebook [1].

Also, note that in the context of all the figures corresponding to Shapley values, the Class 0 will correspond to 'Jaish-e-Mohammad (JeM)' and Class 1 will correspond to 'Lashkar-e-Taiba (LeT)'.

3 CASE STUDY: EXPLAINING ATTRIBUTES WITH MULTIPLE FEATURES

3.1 Global Explanations Based on Shapley Values

The idea of using the Shapley values, also known as SHAP (SHapley Additive exPlanations) for explaining the predictions of complex blackbox machine learning models is already proven [3]. The SHAP is an extremely useful feature to explain the underlying machine learning at the global as well as the local. In this section, we will compare the global intepretability provided by SHAP with the standard feature importance of random forest model and attempt to make conclusions about the overall understanding of the model.

SHAP assigns each feature an importance value for a particular prediction. Its novel components include: (1) the identification of a new class of additive feature importance measures, and (2) theoretical results showing there is a unique solution in this class with a set of desirable properties. The new class unifies six existing methods, notable because several recent methods in the class lack the proposed desirable properties. In order to explain the decision boundary of our model at a global level, we can exploit the idea of feature importance in order to determine how significant a particular feature is while predicting the perpetrator. For the sake of simplicity, we will use the idea that features with large

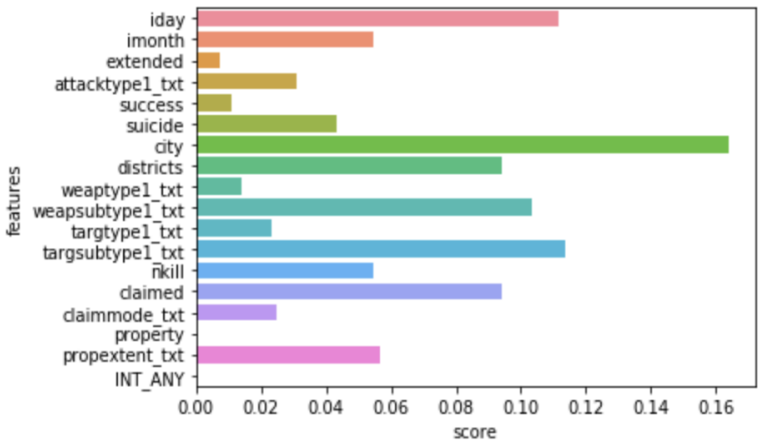

Figure 3: Feature Importance: Based on the Random Forest Model

absolute Shapley values are more important than the ones which have smaller absolute Shapley values. Since we want the global importance, we average the absolute Shapley values per feature across the data:

$$I_j = \sum_{i=1}^{n} |\phi_j^{(i)}| \tag{1}$$

Once we have averaged the absolute Shapley values, we sort the features in a decreasing order and plot the values. Such a plot describes the features which has most significant impact in the decision making of the model. The figure 4 describes the feature significance for the random forest model we have created for predicting perpetrators.

We can clearly observe from the figure 4 that "city" as a feature plays the most important role followed by "claimed" which means whether or not the group claimed its incident which is followed by "day", "districts" and "target_subtype". Here, the idea of target_subtype maybe understood in the sense that different extremist outfits have different ideology which leads to their different target selections to seek their objectives. The observations are aligned with the feature importance plot based on random forest based model described in figure 3 where "city" is the most significant observation. It also relates to the fact the different extremist outfits maintain their individual networks separately while being in contact with other extremist outfits in the same region. However, in the specific context of Jammu and Kashmir, it may be noted that some of the cities are particularly known for harboring some of the specific the extremist outfits which depends on the networks of that extremist outfit with the ideological and political supporters found in that city.

3.2 Local Explanations based on Local Interpretable Model-Agnostic Explanations (LIME)

The idea of interpreting a machine learning model locally is based on the requirements to understand why a particular data point is classified into a particular class if the problem is about classification. Now, the ability of local interpretable model-agnostic explanations is proven and tested in achieving one of the best standards in explaining the observations of a complex machine learning model at the instance level. It is a type of surrogate model that is deployed for explaining the prediction of a single instance made by a black-box model for e.g. a random forest model in our case consisting of some 1000 trees. Surrogate models are trained to approximate the predictions of the underlying black box model [4]. While a global surrogate model attempts to explain the black-box model as a whole, a local surrogate model attempts to explain the individual predictions made by the underlying black-box model.

Mathematically, we can say, an explanation model for an instance x can be defined as,

$$\text{explanation}(x) = \arg\min_{g \in G} L(f, g, \pi_x) + \Omega(g) \tag{2}$$

where,
the explanation model is the model that minimizes the loss function L (for e.g. mean squared error) which measures how close the explanation is to the prediction of the original model f (e.g. a random forest model) and $\Omega(G)$ is the model complexity which has to be kept low. The results based on the application of this method on our random forest model is described in the figure 5 and 6. Also, the names of variable in the model trained using LIME are different as compared to SHAP based explanations for which we had simplified the names of variables. The figure **??** may be used to compare between the two models.

Let us begin with decoding the figure 5 in which we have correctly classified the actual perpetrator as "Lashkar-e-Taiba (LET)". The plot at the top in this figure represents the probability for classification into the LET class as being 0.75 which is also the correct class. The green bars support the decision made by the model while the red bars oppose the decision. The model clearly gives significant emphasis to the feature number of deaths denoted by "kill_count". This can be understood from the value of this feature which is greater than 4. In general, most of the incidents carried out by this outfit involved the death of fewer persons. This claim can

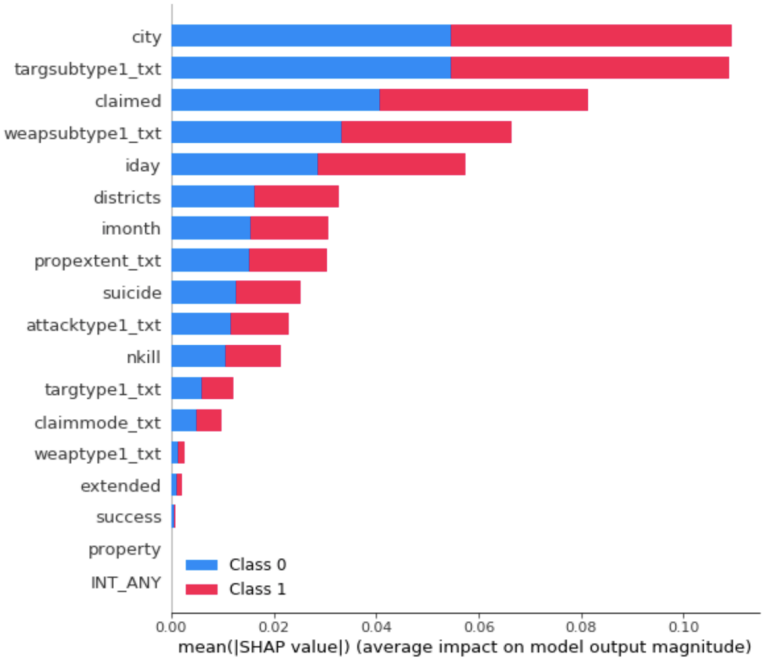

Figure 4: Mean(|SHAP Value|): Average Impact on Model Output magnitude

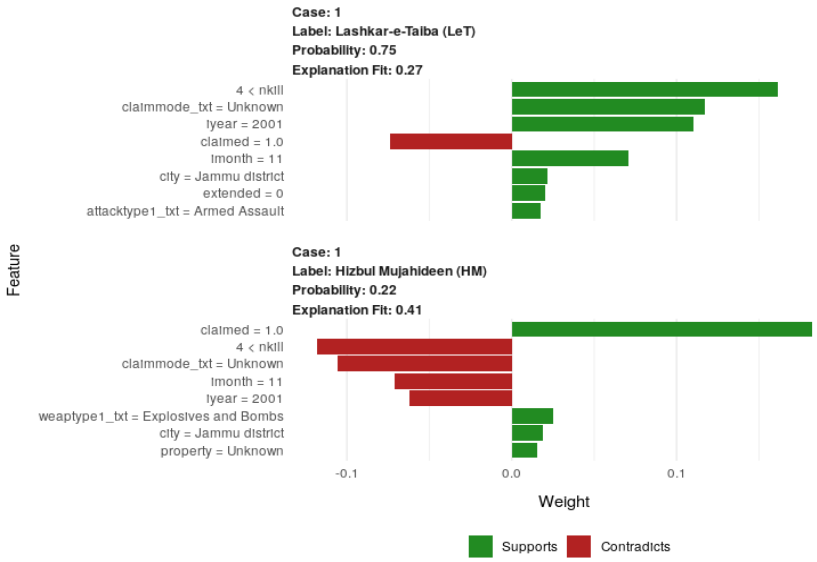

Figure 5: Local Interpretable Model Agnostic Explanations- Explaining the Attributes of Lashkar-e-Taiba (LET)

be verified through the figure 7. The second most important feature that supported this prediction was the claim mode but since the model is unable to distinguish based on that, we consider year of the incident which is 2001 as the second most important feature. This makes sense because it was one of the most active outfits that year. Similarly, the attribute denoting the month of attack "month" also contributed to supporting the prediction. This fact can be verified from the figure 8.

Similarly, the figure 6 explains an incident correctly classified as performed by "Jaish-e-Mohammad (JEM)". The first figure on top represents the probability of the classification into JEM class as 0.44 which is almost similar to the one of LET class being 0.40.

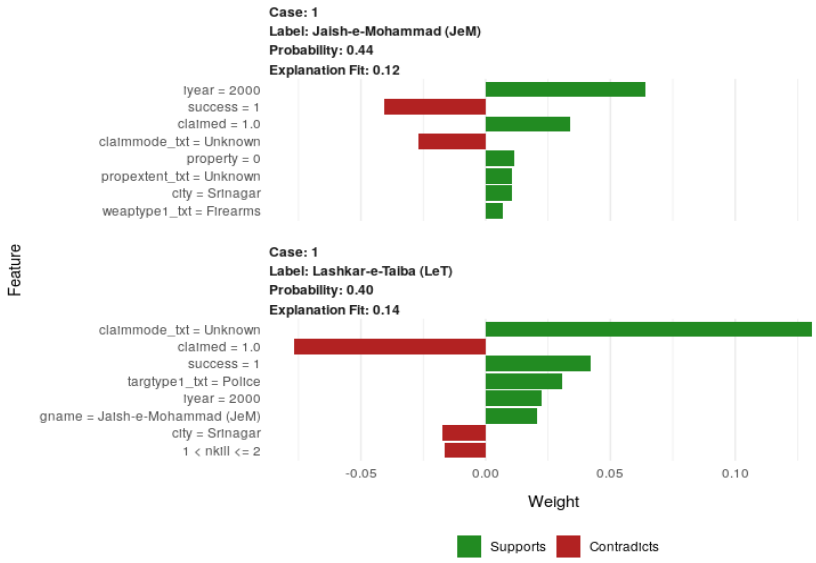

Figure 6: Local Interpretable Model Agnostic Explanations- Explaining the Attributes of Jaish-e-Mohammad (JEM)

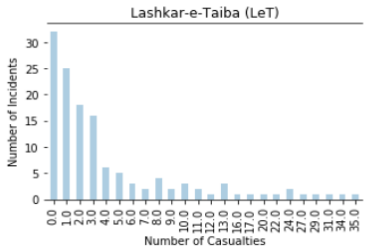

Figure 7: Incident Frequency Distribution: Lashkar-e-Taiba (LET)

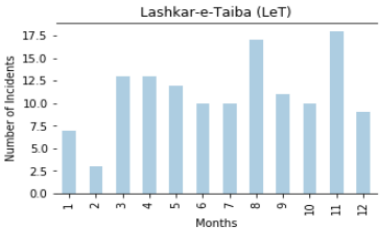

Figure 8: Casualties Monthwise: Lashkar-e-Taiba (LET)

However, some interesting observations that maybe noted down from here can be the presence of similar features as the ones in the plot LET being the most significant but with different values. One particular which played a crucial role as we can observe is the attribute "claimed" being 1.0 in the case of JEM which means True and it supports the decision while exactly same feature was valued 0.0 in the case of LET in figure 5 and it also supported the decision. Therefore, we can observe that most of the times, the incidents performed by LET go unclaimed while the ones performed by JEM are claimed by the extremist outfit. It also makes an attempt to explain the ideology of the the two outfits in which the JEM is more vocal than the LET.

4 CONCLUSIONS

There are many different objectives for pursuing this project, however, due to scarcity of time and limited context of this workshop, we are not able to demonstrate all the objectives. We are working to build a complete predictive intelligence system and this project is a sub component of the assignment. While in this work, we have discussed some of the major challenges related to interpretability in implementing the machine learning based models in the context of defense and homeland security, the work still remains incomplete as the models were not tuned to their full potential as well as a completely open source dataset was used to build the model.

The ideas of having applications of modern technological advances in artificial intelligence pose a great hope for making sustainable use of this technology in the context of homeland security and defense not only for India but for the world as a whole. However, in order to exploit most from this available opportunity, we need to have institutional frameworks along with government interventions. In other words, for developing algorithms for smart decision making, either based on rules or driven by data, we need either well established rules for decision making or extensive amount of data for training models which can complement the open source data to be even more agile.

In the context of India, the Government of India has adopted to institutional frameworks which allow for structured formats for data collection across a wide spectrum. Such interventions

have led to the creation of National Intelligence Grid (NATGRID) which maybe acting as the sole data warehouse concerned with the warehousing of all relevant intelligence data for homeland security and defense. Further, the adoption of National Strategy on Artificial intelligence by the National Institution for Transforming India (NITI Aayog) is another welcome step in this regard.

However, despite having such huge data warehouses, the procedure for applying artificial intelligence is not as simple as it appears. There are inherent challenges associated with artificial intelligence algorithms which must be addressed. Some of them to mention are described here-

(1) Privacy and security
(2) Fairness and Bias Correction
(3) Interpretability and Transparency
(4) Governance, Accountability and Legal Frameworks
(5) Human Resource and Finance

Thus, while building robust artificial intelligence based products is the need of the hour, it is equally important to put additional emphasis on the various challenges mentioned above and ensure that while technology remains an aid to the advancement of sustainable development goals, it must not violate them.

ACKNOWLEDGMENTS

The author, Ankit Tewari would like to extend sincere gratitude towards Shri Om Prakash Tewari, an eminent educationist and visionary for his continuous motivation, guidance and ideological and material support.

REFERENCES

[1] GTD Codebook. 2015. Global Terrorism Database. Codebook: Inclusion Criteria and Variables.
[2] Global Terrorism Database (GTD). 2017. *Study and Response to Terrorism (START).* https://www.start.umd.edu/gtd/
[3] Scott M Lundberg and Su-In Lee. 2017. A unified approach to interpreting model predictions. In *Advances in neural information processing systems.* 4765–4774.
[4] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. " Why should i trust you?" Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining.* 1135–1144.