Modeling The Causes Of Terrorism From Media News: An Innovative Framework Connecting Impactful Events With Terror Incidents

Truong Son Pham
Le Quy Don Technical University
Faculty of Information Technology
Hanoi, Vietnam
sonpham.mta@gmail.com

Tuan-Hao Hoang
Le Quy Don Technical University
Faculty of Information Technology
Hanoi, Vietnam
haohth@lqdtu.edu.vn

Abstract— **Terrorism has become an increasingly relevant issue accounting for significant social, economic and political impact. Due to powerful media coverage on the subject, a lot of information is now publicly available, although normally found in an unstructured form. This research aims to better understand the connection between a collection of impactful events, such as external or internal conflicts and military operations, with terror events and its motivations. To this end, a framework was devised, starting with an online news scraper, coupled with machine learning and natural language processing techniques, capable of clustering keywords into the main topics found in the news. The results of these algorithms, in the form of structured data, were later fed to a modeling technique capable of finding, to a certain degree, the connections between topics and terror events.**

The approach presented in this work adopts a perspective that, to the best of our knowledge, has not been previously seen in specialized literature. Furthermore, this methodology constitutes the groundwork for open source intelligence, capable of being applied to various similar domains like the prediction of political risk index or economic risk index.

Keywords: Machine Learning, Modeling, Terrorism, Open Source Intelligence.

I. INTRODUCTION

During the last years, terrorism has become an increasing issue accounting for significant social, economic and political impact. The explanation of the causes of terrorism has gained considered interest.

Crenshaw organized the study of terrorism around three questions about the reason of terrorist attacks, the process of terrorism works and its social and political effects [1]. In her publication, she clarified that "terrorism is neither an automatic reaction to conditions nor a purely calculated strategy" [1]. She also concluded that terrorism represents more often the disaffection of a fragment of the elite and inconsistent reactions of government seem to encourage the terrorism [1]. Krueger confirmed that not the economic deprivation and the lack of education caused people turn to terrorism but the influence of political outcomes [2].

Although media has shown a powerful coverage, the publicly available information on the subject could be normally found in an unstructured form. Therefore, the usage of this information is still limited.

In this study, we aim to develop a dynamic framework for understanding the connection between a collection of impactful events with terror events and its motivations. To our knowledge, this is the first framework in open source intelligence to model the causes for terrorism. Section 2 explains the methodology of the framework. Section 3 continues with the description of the results and in section 4 we will discuss the result and plan our future works.

II. RELATED WORK

Our framework starts with an intelligent new scraper, coupled with machine learning and natural language processing techniques, capable of clustering keywords into the main topics found in the news. This section describes these technologies.

A. Counter Terrorist

Predicting the terrorist behavior is an important task of governments and has gained many attentions from researchers. F. Olajide et al. used social network analysis (SNA) to predict key player of terrorist network [3]. They showed that SNA is quite efficient and effective to understand the terrorist network [3].

In a different approach, S. Tutun et al. at Binghamton University developed a framework to predict the future terrorist behaviors using history data. They recognized the patterns in past attacks to detect terrorism attacks with high accuracy [4].

B. Web Mining

Web mining can be described as an application of data mining techniques to discover useful information from large datasets including web content, structure, and usage automatically [5].

Web content mining contains information extraction, topic tracking, summarization, categorization, clustering and information visualization [6]. This research is close to text mining and information retrieval, which nowadays mainly based on machine learning techniques [7].

Web scraping is a process of collecting information automatically from websites [6]. Web scraping has been widely used in different domains such as focused searching [8] or extraction of big data from the internet [9].

Although a lot of research have been done in Web mining, there is a need to create methodologies to:

978-1-5386-6113-0/18/$31.00 ©2018 IEEE

- Track and detect the changes of Web content or structure automatically,
- Remove irrelevant, duplicated information even in different form, detect fake news, spam and
- Protect personal information on web efficiently.

C. Machine Learning And Natural Language Processing

Machine learning plays a very important role in our framework. It automates the data collecting process using web scraper and builds an information analytic model based on these data. Machine Learning is one of the main branches of artificial intelligence. It concerns the construction and study of systems that can learn from data [9]. Based on the type of dataset, machine learning algorithms can be organized into three categories. The first category is supervised learning, where all the samples in training set are labeled. The second category contents semi-supervised learning algorithms, where just a part of the samples in training set is labeled. In the last category are unsupervised learning algorithms, where these algorithms operate on unlabeled samples [10].

In this study, we used the semi-automated to collect and label the data. That means, supervised machine learning algorithms were applied to learn from this data and make the scraper functional by an unknown website. The web elements should be classified into multi-classes such as text, date and image. For this reason, we applied multiclass algorithms to predict if a web element contains this information or it is not relevant for the text analysis processes.

Basically, there are two well-known techniques for text feature extractions, namely "term frequency" and "term frequency-inverse document frequency". While a term frequency weight (TF) simply represents the weight of a term that occurs in a document, a term frequency-inverse document frequency weight (TF_IDF) is the product of term frequency weight and inverse document frequency (IDF) weight of this term in a collection of documents, which can be computed as below [11]:

$$TF(t) = \frac{number\ of\ times\ term\ t\ appers\ in\ a\ document}{total\ number\ of\ terms\ in\ document}$$

$$IDF(t) = log_e \frac{total\ numbers\ of\ documents}{number\ of\ documents\ with\ term\ t\ in}$$

$$TFIDF(t) = TF(t) * IDF(t)$$

III. METHODOLOGY AND RESULTS

In this study, we provided a framework to better understand the connection between a collection of impactful events with terror events and its motivations. This framework could automatically collect the data from open sources using an intelligent web scraper trained by machine learning algorithms and model the causes for terrorism based on these data. Figure 1 shows the complete structure of this framework.

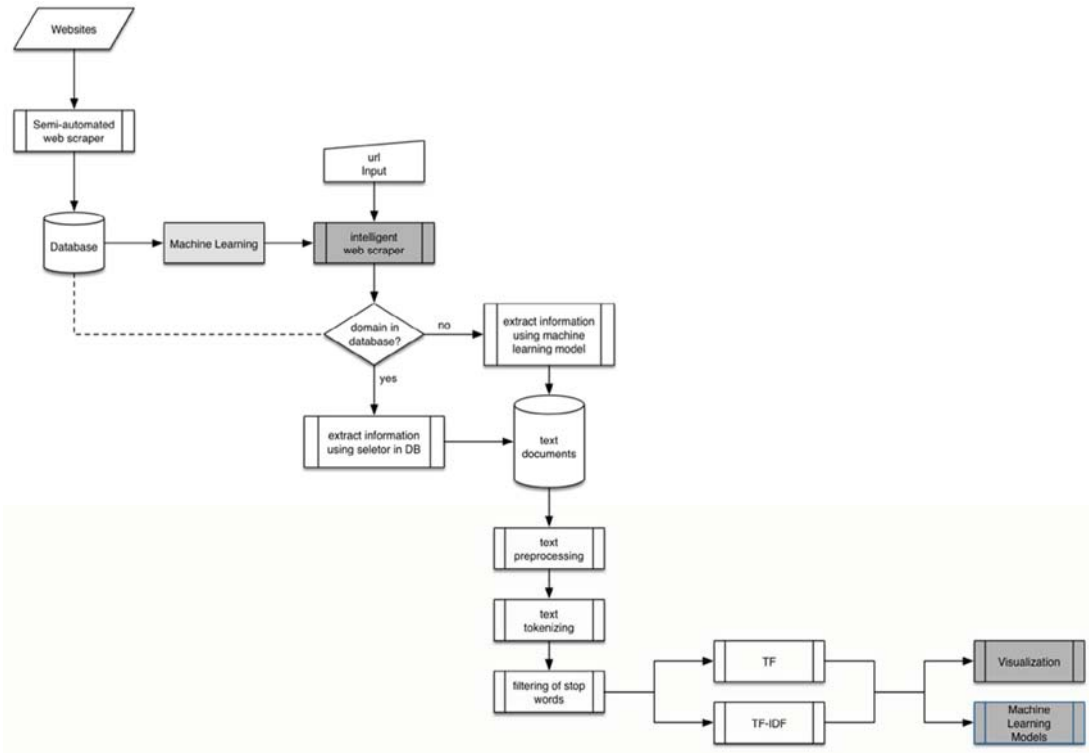

Figure 1 Framework for modeling the causes of terrorism

Normally, a web scraper must be written for each web to collect the data and it will work only on that web. That means it mays cost a lot of time if we want to collect data from many websites. Therefore, in this study an intelligent web scraper was developed. The term "intelligent web scraper" is understood to mean a fully automatic web scraper using advanced machine learning techniques, which could extract and analyze information from any website. Figure 2 shows the structure of an intelligent web scraper.

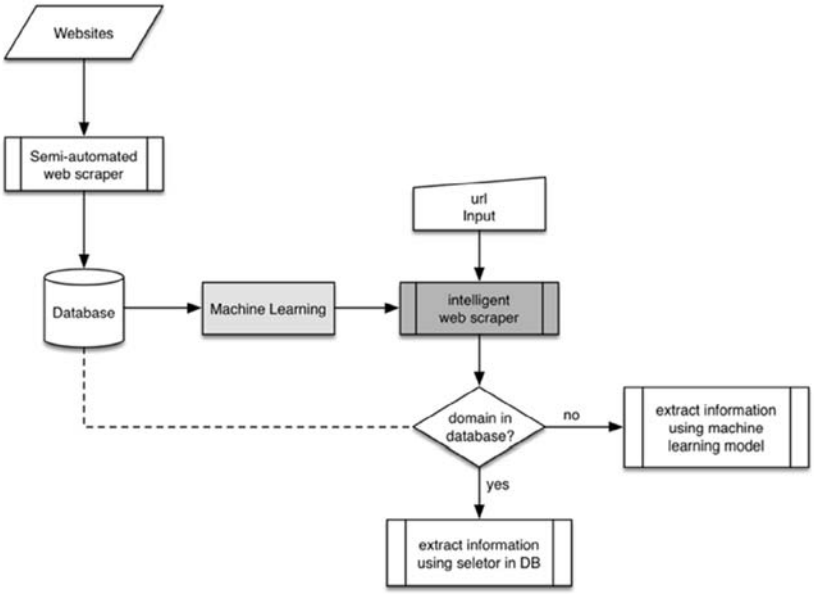

Figure 2 Intelligent web scraper structure

We needed a lot of data to train the machine learning algorithms. For this task, we created a semi-automated data collecting tool to collect data from different websites for the training process. This tool is a CSS selector. It can extract the features of web element, such as title, image, text and date automatically using mouse clicks. Figure 3 shows the user interface of this tool.

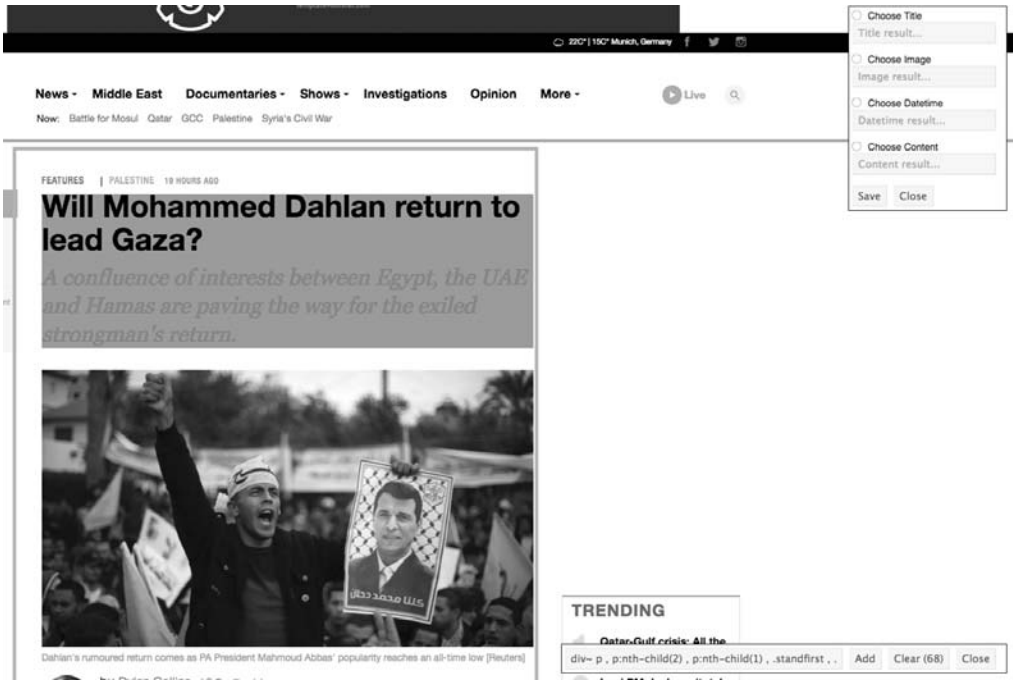

Figure 3 Semi-automated web scraper UI

In the dataset, the name of a web element is the label of a sample and the features of web elements such as text size, color, text length and image size are the features of the samples. The machine learning techniques will be trained with this dataset to make our web scraper functional on an unknown web. Once a web URL is given as an input, the scraper will distinguish if the domain of this website already exists in the database. If it exists, the web scraper will use the selectors in the database to extract information from the web, otherwise, trained machine learning models will be used for this task. Table 1 shows the data structure of the database for machine learning algorithms.

TABLE 1 DATASET STRUCTURE FOR WEBSCRAPER

domain	Selector	Text size	top	Color	Text length	Image size	Label
dailymail.co.uk	h1	19	27	(17,17,17)	35	None	Title
dailymail.co.uk	div span	12	35	(0,102,192)	10	None	Date
dailymail.co.uk	div iframe	None	109	None	None	(87,130)	Image
	…	…	…	…	…	…	…

With the intelligent web scraper, we could collect more than three thousand articles about the Middle East from different websites automatically.

After collecting the text data from websites, we applied modern topic modeling techniques and machine learning algorithms to analyze this collection for understanding the connection between the publicly available news and the terror events and its motivations. In order to perform topic modeling and machine learning algorithms, we have to extract features from these text documents. This feature extraction process starts with text pre-processing, tokenizing and filtering of stop words, capable of calculating a statistical measure of each word to evaluate how important it is to a document. Figure 4 explains the structure of a text analysis process from a collection of text documents.

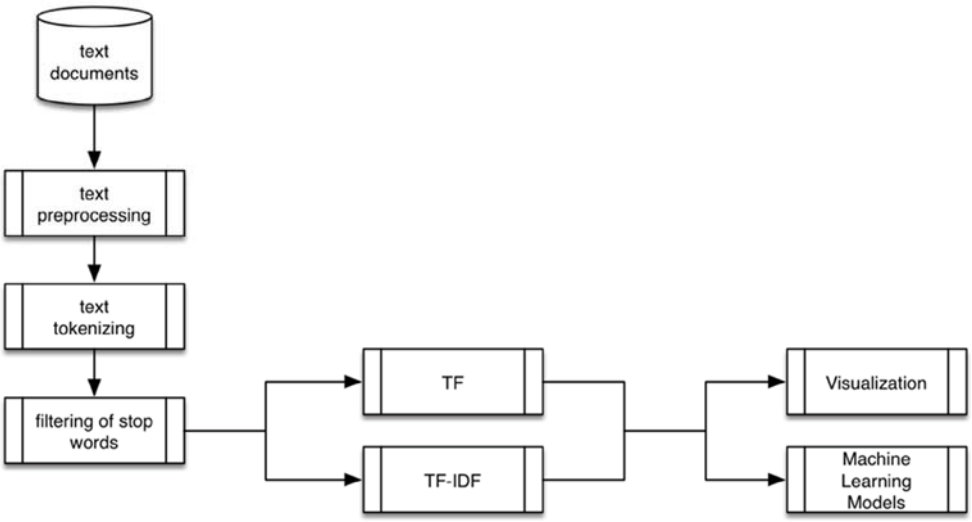

Figure 4 Text analyze process

We also applied the feature extracting process based on TF and TFIDF to evaluate the importance of words in the text documents and built a words cloud based on the weights of each word. At the end, we modeled the topics of this text and visualized the most relevant of them in combination with the top important words. While Figure 4 shows the word cloud and topics modeled with TF, figure 5 represents this combination modeled with TFIDF.

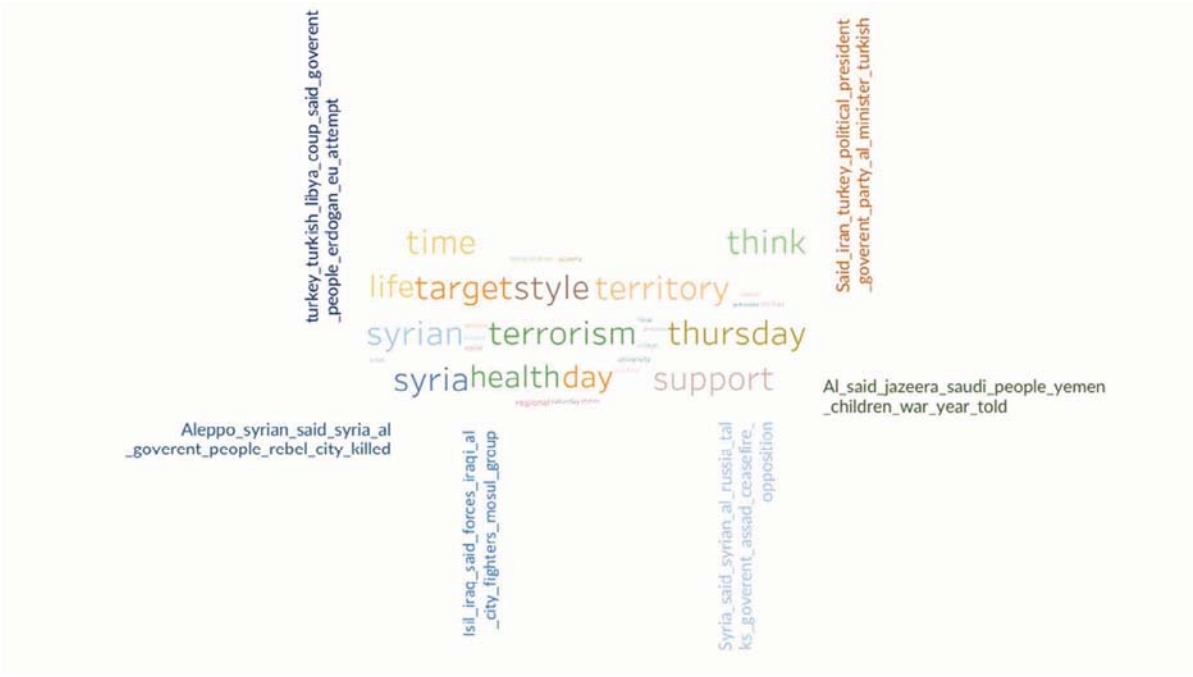

Figure 5 Word cloud and most discussed topics modeled by TF

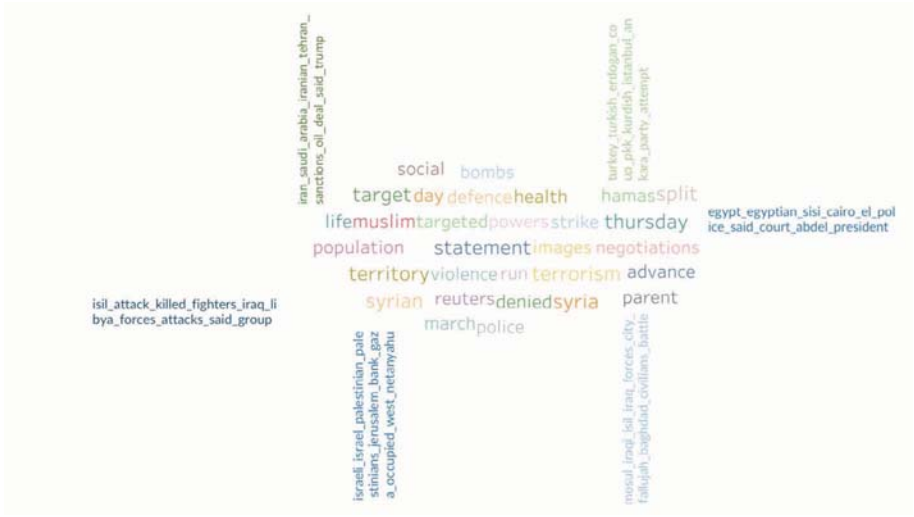

Figure 6 Word cloud and most discussed topics modeled by TF-IDF

The result showed that our framework is able to extract important information such as Thursday's attack happened in Paris or the most discussed topics right now such as Turkey, Egypt or ISIL attacks.

IV. CONCLUSION AND FUTURE WORK

In this study, we collected news about the Middle East using an intelligent web scraper powered by machine learning. Further, we could build a text analyzing the process to model the most discussed topic in the field of terrorism and visualized the most important words in combination with these relevant topics. We had a problem by collecting much more data for the machine learning algorithms to build a predictive model. These algorithms could only perform well with a lot of training data. We are planning to improve the web scraper, to turn it into a focused search engine so that we can search horizontally for more information on the internet to get hundreds of thousands of articles downloaded and use them as the training data for machine learning algorithms. We are also planning to improve the machine learning performance by applying some feature selections and parameter tuning algorithms to get a better result by predicting the terrorist attacks based on open sources information.

REFERENCE

[1]. Crenshaw, Martha. "The causes of terrorism." Terrorism studies: A reader (2012): 99-114.

[2]. Krueger, Alan B. "What makes a terrorist." Economics and the Roots of Terrorism 6 (2007).

[3]. Olajide, F., & Adeshakin, K. (2016). Towards the investigation of using social network analysis for counter terrorism in West Africa: case study of Boko Haram in Nigeria. J. Eng. Sci. Technol, 11(11), 1629-1638.

[4]. Tutun, S., Khasawneh, M. T., & Zhuang, J. (2017). New framework that uses patterns and relations to understand terrorist behaviors. Expert Systems with Applications, 78, 358-375.

[5]. Mohata, P., and Sheetal Dhande. "Web Data Mining Techniques and Implementation for Handling Big Data." Computer Science and Mobile Computing 4.4 (2015): 330-334. APA

[6]. Johnson, F., & Gupta, S. K. (2012). Web content mining techniques: a survey. International Journal of Computer Applications, 47(11). Chicago

[7]. Malik, Sanjay Kumar, and S. A. M. Rizvi. "Information extraction using web usage mining, web scrapping and semantic annotation." Computational Intelligence and Communication Networks (CICN), 2011 International Conference on. IEEE, 2011.

[8]. Haddaway, Neal R. "The use of web-scraping software in searching for grey literature." Grey J 11.3 (2015): 186-90.

[9]. Landers, Richard N., et al. "A primer on theory-driven web scraping: Automatic extraction of big data from the Internet for use in psychological research." Psychological methods 21.4 (2016): 475.

[10]. Pham, Truong Son, Quang Uy Nguyen, and Xuan Hoai Nguyen. "Generating artificial attack data for intrusion detection using machine learning." Proceedings of the Fifth Symposium on Information and Communication Technology. ACM, 2014.

[11]. Luhn, Hans Peter. "A statistical approach to mechanized encoding and searching of literary information." *IBM Journal of research and development*1.4 (1957): 309-317