10.31341/jios.47.1.1 UDC 004.62:343.3/.7
Open Access Original Scientific Paper

Towards Detecting Influential Members and Critical Topics from Dark Web Forums: A Data Mining Approach

Faris Ali *faris_95208@svuonline.org*
Faculty of Information Technology and Communications
Syrian Virtual University, Syria

Randa Basheer *randa_151689@svuonline.org*
Faculty of Information Technology and Communications
Syrian Virtual University, Syria

Mouhamad Kawas *mouhamad_151733@svuonline.org*
Faculty of Information Technology and Communications
Syrian Virtual University, Syria

Bassel Alkhatib *t_balkhatib@svuonline.org*
Syrian Virtual University, Syria

Abstract

Conventionally, the Internet consists of three parts: Surface, Deep, and Dark Webs. In the last two decades, a massive increase in illicit activities took place on the different platforms of the Dark Web. Moreover, social networks on Dark Web implicate extremism dissemination on a wide scale. In this paper, we propose an approach to generate textual patterns from discussions on Dark Web terrorist forums employing Data Mining techniques. The discovered patterns help identify the influential members and extract critical topics. We describe our system modules that perform data preprocessing, text preprocessing with TF-IDF weighting, outlier detection, clustering evaluation, clustering, and clustering validation, implemented with the RapidMiner tool. We apply K-Means as the Clustering method with different distance metrics, evaluate the clustering process using Elbow and Silhouette methods, and validate it using Davies-Bouldin Index. Furthermore, we investigate the effects of altering the distance metrics for outlier detection on the Clustering results.

Keywords: Dark Web, Forum, Elbow Method, Silhouette Method, Davies-Bouldin Index, Data Mining, Clustering, Text pre-processing, Terrorism.

1. Introduction

Notably, researchers use the terms Deep Web and Dark Web interchangeably, though they do not refer to the same thing. Deep Web refers to websites that standard search engines, such as Google, cannot index. The Deep Web content includes platforms restricted by a paywall or require sign-in credentials. In addition, some administrators set constraints on their websites that block web crawlers, thus making such websites inaccessible by search engines. On the other hand, the Dark Web is a part of the Deep Web that uses specific encryption software and requires particular browsers, such as TOR[1], to be accessed [1].

Websites on the Dark Web look roughly like any other website, though with some essential differences. Of the naming structure, instead of the common extensions, such as .com or .net, websites hosted on the TOR network by TOR Hidden Services have the unique one of .onion. These websites are only accessible via the TOR browser [2] [1]. Websites on Dark Web also use a scrambled naming structure in their URLs to make them impossible to remember. For example, a Dark Web market called Dream Market goes by the incomprehensible address of eajwlvm3z2lcca76.onion.

Many internet users are not thoroughly aware of the criminal activities taking place on the dark side of the web. On the other hand, numerous websites disseminating terrorism and extremism contents use languages other than English, thus making it laborious for investigators to understand the website content [3] [2].

Initially, terrorist groups used open social networks, such as Facebook and Twitter, for thoughts dissemination and recruitment purposes. However, as social media websites employ ever-improved techniques to detect such suspicious behaviors, they immediately disclose and suspend such accounts. Thus, terrorist communities have moved their activities to Dark Web forums [4].

Data Mining and Machine Learning techniques play a vital role in forming the proper intelligence needed to detect and analyze malicious and illicit activities over the Dark Web platforms. In particular, DM and ML help discover new advantageous information about terrorism discussion forums, members, and posts [4].

In this paper, we introduce a methodology based on unsupervised methods, specifically Clustering, to analyze terrorist discussions on the Dark Web, providing a better understanding of the trending and most common topics they discuss and finding influential members. We discuss the importance of evaluating the clustering process to gain accurate results. We start with a brief description of Data Mining, data pre-processing, outlier detection methods, cluster evaluation methods (specifically Elbow and Silhouette methods), and clustering algorithm with several metrics (Cosine Similarity , Euclidean Distance, and Manhattan Distance), and validation metric (the Davies-Bouldin Index). We discuss our results, demonstrating how the outlier detection and clustering validation processes improve the clustering results. Finally, we discuss the utilization of text pre-processing in detecting the relationships between forum members and their posts, and extracting the most critical terms characterizing the terrorism domain, such as "تفجير = explosion", "قتل = kill", and "ارهاب = terrorism". Our approach helps law enforcement and security agencies to investigate terrorist forums and detect influential members and trending topics.

The remainder of this paper is organized as follows: Section 2 discusses related work and the contributions of this paper, Section 3 demonstrates the theoretical and

[1] TOR Project, https://www.torproject.org

mathematical background, Section 4 describes the research methodology and experimental results, and Section 5 is for the conclusion and future work.

2. Related Work

Numerous researches introduced approaches to analyze the Dark Web contents in general and terrorism dissemination in particular, utilizing a variety of Data Mining and Machine Learning techniques. Such studies comprise various approaches of Clustering, Classification, Topic Modeling, and statistical analysis.

The study of Baghel and Yogesh [5] analyzes the increased rate of terrorist incidents from 1970 to 2016 worldwide, intending to detect and curb such incidents in a particular country (India). The study follows a prediction methodology to predict the new areas prone to become the following targets of terrorist activities by determining the pattern of increasing probability of the least significant areas in terms of density and region. The researchers retrieved the global data from online sources to study the specific situation of the selected country and modeled the data using the K-means algorithm. The study suggests that the data contains many broad variables over periods, so assuming values as a whole can lead to erroneous insights. Thus, the number of clusters was verified using the Davies-Bouldin Index to select the best value of k and therefore ensure better efficiency and validate the results.

Similarly, Alkhatib and Basheer [3] introduced an approach that utilizes Data Mining techniques to infer valuable patterns from a Dark Web marketplace content. They demonstrated the system modules that perform several tasks, including crawling, extracting the whole market data, data preprocessing, and Data Mining. They employed K-means algorithm and validated the results using Davies-Bouldin Index. The study discussed the common characteristics among clustered objects, identified Top Vendors, and analyzed products promoted by the latter, in addition to identifying the most viewed and sold items on the market using the RapidMiner tool.

Goel, Sharma, and Gurve [6] proposed a classification approach conducted on the Global Terrorism Dataset (GTD), which comprises statistics on domestic and international terrorist attacks since the 1970s. They developed a model using the Orange Data Mining tool and applied different Machine Learning algorithms, such as Support Vector Machine (SVM), Naïve Bayes, Neural Networks, and K-Nearest Neighbor (KNN). The model employs statistical analysis by classifying five decades of real-time global terrorism datasets to predict future attacks.

Atsa'am, Wario, and Okpo [7] employed the Agglomerative Hierarchical Clustering technique on the Global Terrorism Database (GTD) to classify terrorism into four newly suggested groups based on losses and consequences. The proposed system comprises several processes, including data preprocessing, attribute selection, missing values handling, and data normalization. The approach utilized internal validation methods to estimate the compactness, connectedness, and separation of the generated clusters. Furthermore, it evaluated the optimal number of clusters using the Elbow Method.

Similarly, Kumar et al. [8] proposed an approach that employs classification techniques to observe trends in terrorist attacks worldwide in terms of casualties, attacks frequencies, and most vulnerable locations in the world. The dataset created for experimental purposes was collected from various public and open sources during 1970-2015, consisting of reported attacks, which caused massive losses of life and property. The research focuses on the importance of data preprocessing in increasing the accuracy of the results and analyzes the performance of several classifiers. The classification criteria are based on attacks responsibility by taking the party responsible for the attack as a classification category to predict the organizations involved in the reported attacks.

Alguliyev, Aliguliyev, and Niftaliyeva [9] proposed a method for detecting terrorism-related activities in the e-government environment based on the similarity between user opinions and a vocabulary database. They generated the vocabulary database from terrorism-related words for automatic analysis and identification of potentially dangerous statements using Naive Bayes classification to prevent terrorist threats, identify suspects, and monitor their activities. The approach identifies the positive, negative, and neutral comments, selects only the negative ones for analysis, and calculates the proximity between the comments' words and the terrorism-related dictionary database. If the proximity exceeds the specified limit, the system enters the related statement in the list of suspicious comments. The approach includes several processes, including data preprocessing, determination of comments polarity, and selection of terrorism-related comments.

Onyekachi, Norbert, and Uzoamaka [10] discussed the importance of employing Deep Learning to analyze terrorist events. The research introduced improvements to a previous work by including extra features in the dataset and proposing a Deep Neural Network (DNN) model to predict the success of terrorist attacks. They used a dataset retrieved from the Global Terrorism Database (GTD). The employed methodology comprised several steps, including data preprocessing and handling the missing values.

Mashechkin et al. [11] proposed a methodology to predict potential risks from online communities by detecting potentially dangerous users without having full access to the content they distribute, such as on private channels and chat rooms. The study aims to identify and monitor users, communities, and web resources that spread terrorist or extremist content on the Internet by classifying users as "dangerous" or "non-dangerous". The approach employed Orthonormal Non-negative Matrix Factorization and words representation as n-grams to handle language variations, misspellings, and typos. Furthermore, it relies on noise filtration and weighing the generated n-grams to fit relevant individual text parts and remove the less important parts of the document; thus, the retained text parts describe all the essential thematic topics of the source text. The study paid particular attention to issues related to text quality, such as addressing the presence of links and hashtags, multiple languages, slang, or intentionally disguised words, in addition to typos and grammatical errors.

Soliman and Abou-El-Enien [12] proposed a hybrid prediction algorithm integrating different Operations Research (OR) and Decision support tools with Data Mining techniques (prediction and classification). The research purpose is to predict terrorist groups responsible for terrorist attacks on Egypt from 1996 to 2017 and thus use the predictions as a warning tool that identifies the networks of terrorist groups and reduce or curb terrorist attacks. The system employed K-Nearest Neighbors (KNN) and Random Forest (RF) algorithms controlled by metaheuristics optimization methods to determine the minimum number of the selected features to achieve a high classification accuracy. The proposed methodology included data preprocessing and utilized evaluation and validation measures.

Tavabi et al. [13] analyzed a large corpus from 80 forums on Deep and Dark Webs during 2016 and 2017. They identified topics using Latent Dirichlet Allocation (LDA) and employed a non-parametric Hidden Markov Model (HMM) to model the evolution of the topics across different forums. The proposed system examines the dynamic patterns of discussions according to their similarities. Eventually, they demonstrated anomalous events identification process in rich, heterogeneous data.

Saini and Bansal [4] proposed an automatic model to detect terrorist recruitment over online social media to generate actionable intelligence. The approach is based on classification techniques applied to messages from five Dark Web forums under the Dark Web Portal Project of University of Arizona. Two experts manually labeled the messages into two classes: "recruitment" and "non-recruitment". They employed five classifiers: Support Vector Machine, Logic Boosting, Random Forest, Generalized Linear Model, and Maximum Entropy-based model.

Our work contributes to the same purposes focusing on the substantial role of Data Mining in countering terrorism. The main contributions of this paper are as follows:

- We Propose a Data Mining approach on Dark Web terrorist forums in the Arabic language. It includes Clustering using K-means, Outlier Detection, Cluster evaluation using Elbow and Silhouette Methods, Cluster validation using Davies-Bouldin Index, and text preprocessing.
- We discuss the outlier detection effects on the Clustering process using K-means, comparing different distance metrics. Consequently, we choose the best distance metric after a validation process using Davies-Bouldin Index.
- We demonstrate the implementation of text preprocessing on the selected datasets in discovering influential members and trending topics around critical terms such as “قتل = kill” and “ارهاب = terrorism”.

3. Theoretical and Mathematical Background

Data Mining is a business process that extracts meaningful patterns and rules from a massive volume of data. Data mining techniques help discover hidden relationships among raw data elements, leading to inferring information and insights about the data for decision-making purposes [14]. Data Mining can be defined as a knowledge retrieval process that discovers useful and comprehensive information from massive volumes of raw data to support decision-making and problem-solving objectives in a specific domain [15].

3.1. Clustering

The main goal of clustering is to group the elements that share certain similarities; thus, the elements within the same cluster are similar to each other and different from those in the other clusters [7] [16]. Clustering, or grouping, is an essential unsupervised Machine Learning method and is widely used in various domains due to its ability to find and group objects of specific characteristics [15] [17]. Clustering algorithms vary into multiple types based on partitioning criteria, density, and model. Partition-based clustering creates clusters of a predetermined number k. One of the most commonly used Partition-based clustering algorithms is K-means [17].

K-means involves two main steps: selecting k objects as clusters' centers, called centroids, and assigning the other objects to their closest centroid [14]. K-Means is more sensitive to outliers than other clustering methods such as K-Medoids. Thus, the accuracy of the K-Means algorithm improves when the outliers are removed [18].

3.2. Outlier Detection

Outlier detection is the process of detecting and subsequently excluding outliers from a given set of data [19]. The traditional outlier detection techniques have six types: clustering-based, deviation-based, density-based, subspace-based, statistical-based, and distance-based [20]. We choose a distance-based approach, which calculates the distances between a data point and its neighbors. Outliers are objects that have long distances from their set [21] [22].

3.3. Optimal Number of Clusters

The number of clusters *k* is the most fundamental hyperparameter in K-Means clustering. Therefore, external and internal clustering validations are crucial for results validity [23]. The main goal of cluster validation is to ensure that clustering results are meaningful and not just artificial effects of the clustering algorithm [7]. In this work, we utilize two validation methods: the Elbow and Silhouette methods.

3.3.1. Elbow Method

The Elbow Method is one of the most popular methods for determining the optimal number of clusters. It starts with calculating the Within-Cluster-Sum of Squared Errors (defined in Equation 1) for different values of k and selects the k value for which the change in Sum of Squared Errors (SSE) first starts to decline. In other words, the point after which the distortion starts decreasing, forming a bend as an angle, or elbow, will indicate the best value of *k* to be selected [14].

$$SSE = \sum_{k=1}^{k} \sum_{x_i \epsilon S_k} \| X_i - C_k \| \, {}^{2}_{2}$$

Equation 1*:* Sum of Squared Errors (SSE)

3.3.2. Silhouette Method

The Silhouette Coefficient integrates separation and cohesion [17]. In this method, a value close to *(-1)* indicates that the clustering of the data points is not optimal. Conversely, a value close to *(1)* refers to an optimal clustering [17]. Equation 2 defines the Silhouette Coefficient.

$$S(i) = \frac{b(i)\text{-}a(i)}{\max\{a(i),b(i)\}}$$

Equation 2: Silhouette Coefficient

Where *b(i)* is the smallest average distance of the point *(i)* to all points in any other cluster and *a(i)* is the average distance of *(i)* from all other points in its cluster.

3.3.3. Davies-Bouldin Index

The Davies Bouldin Index computes the quality of clustering. Similarities between points in the same cluster increase if the distance of points to the cluster centroid is minimum, while similarities between clusters decrease if the distance is maximum [24]. The longest distance among clusters and the shortest distance within the cluster produce the lowest Davies-Bouldin score, indicating optimal results [25]. Equation 3 defines the Davies Bouldin Index.

$$DB = \frac{1}{k}\sum_{i=1}^{k} max_{i\neq j}\, R_{ij}$$

Equation 3: Davies-Bouldin Index

3.4. Distance Metrics

The second fundamental factor in a clustering process is determining the distance metrics. A distance metric defines the distance between two objects. In Clustering, we can divide similarity metrics into two categories: Similarity-Based Metrics and Distance-Based Metrics.

From the Similarity-based Metrics, we choose Cosine Similarity. The Cosine Similarity calculates the cosine of the angle between two objects' vectors, defined by Equation 4.

$$\mathcal{S}(x,y) = \frac{x.y}{\|x\|\|y\|} = \frac{\sum_{i=0}^{n-1}\ x_i y_i}{\sqrt{\sum_{i=0}^{n-1}(\mathcal{X}_i)^2} \times \sqrt{\sum_{i=0}^{n-1}(\mathcal{Y}_i)^2}}$$

Equation 4: Cosine Similarity Measure

As Distance-Based Metrics, we choose the Euclidean Distance and Manhattan Distance. The distance between two points as a straight line is the Euclidean distance [26]. It is suitable for continuous numerical variables and is defined by Equation 5.

$$EUC(x, y) = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2}$$

Equation 5: Euclidean Distance

The Manhattan Distance, also known as city block distance, is the sum of absolute differences between points across all the dimensions, and computed as in Equation 6.

$$Manhattan\ Distance = \sum_{i=1}^{n} |x_i - y_i|$$

Equation 6: Manhattan Distance

3.5. Text Preprocessing

Data preprocessing is a substantial task in Data Mining in general and Text Mining in particular. It is a strategy for extracting meaningful and significant data elements from raw data by converting the latter into formats suitable for the intended research and experimental purposes [8].

In Text Mining, There are three weighting approaches applied when converting documents to vectors of terms: Term Frequency (TF), Document Frequency (DF), and Term Frequency-Inverse Document Frequency (TF-IDF). In TF-IDF, if a term is significant, it is assigned a heavier weight. Conversely, an insignificant term will have a trivial weight [27]. TF-IDF is defined as follows:

$$TF = \frac{Number\ of\ times\ the\ word\ t\ occurs\ in\ the\ text}{Total\ number\ of\ words\ in\ the\ text}$$

$$IDF = \frac{Total\ number\ of\ documents}{Number\ of\ documents\ that\ contain\ the\ word\ t}$$

$$TFIDF = TF\ .IDF$$

4. Research Methodology

In this paper, we have two main objectives:

- We investigate the effects of outlier detection by comparing distance metrics in Clustering experiments using the K-Means method. Consequently, we choose the best distance metric by applying the cluster validation method Davies-Bouldin Index.
- We demonstrate the implementation of text preprocessing on the selected datasets to detect members and messages around critical words such as “قتل = kill” and “ارهاب = terrorism”.

After filtering the data, we saved it in an Excel file. We implemented the proposed processes using RapidMiner 9.6. In the following subsections, we describe the datasets we used in experimenting with the designed processes. We demonstrate in detail the modeling steps, processes architecture, and results.

4.1. Dataset

We used three datasets in the Arabic language, namely *AlFaloja*, *Alqimmah*, and *AsAnsar*, retrieved from *Azsecure-data.org*[2]. The datasets include posts of worldwide Jihadi social media from 2005 to 2012. Since the datasets contain missing and null values in addition to duplicates, we removed records with such entries from the dataset to clean the data for further analysis. To experiment the proposed system, we considered approximately 9000 records for analysis.

4.2. Modeling Steps

Our proposed approach consists of four main steps described in the following subsections. Figure 1 illustrates the process design.

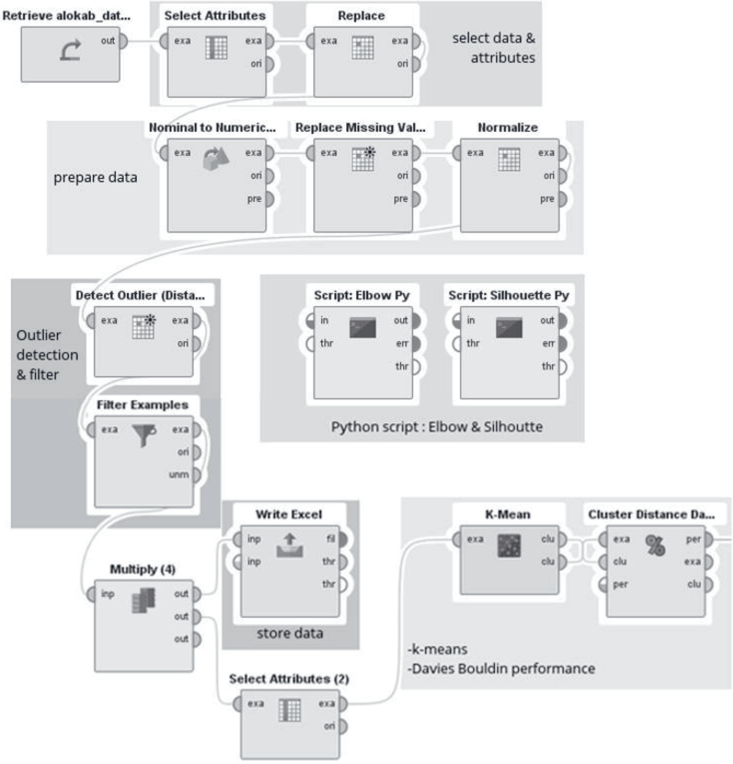


Figure 1. The Process of the Proposed Modeling Method

[2] https://www.azsecure-data.org/dark-web-forums.html

4.2.1. Step 1: Retrieve and Prepare Data

We saved the downloaded datasets in Excel files. On RapidMiner, we retrieve the dataset and select the attributes: *ThreadName*, *MemberName*, *Message*, *P_Year*, *P_Month*, and *P_Day*. Subsequently, we replace special characters - such as "@ #$&…." - with spaces. We convert the results to numerical values, replace the missing values, and normalize them by Z-transformation formula.

4.2.2. Step 2: Outlier Detection

We utilize a distance-based outlier detection tool to discover the outliers from the retrieved data and filter them out. Then, we save the produced dataset in a new Excel file.

4.2.3. Step 3: Validate Clustering

We apply cluster validation on the Excel file data achieved from the previous step with Python code. We employ the Elbow and Silhouette methods, as illustrated in Figures 2 and 3, respectively. As shown in Figure 2, there is a plain bend at three. Thus, the optimal number of clusters with the Elbow method is *k=3* for *AlQimmah* dataset.

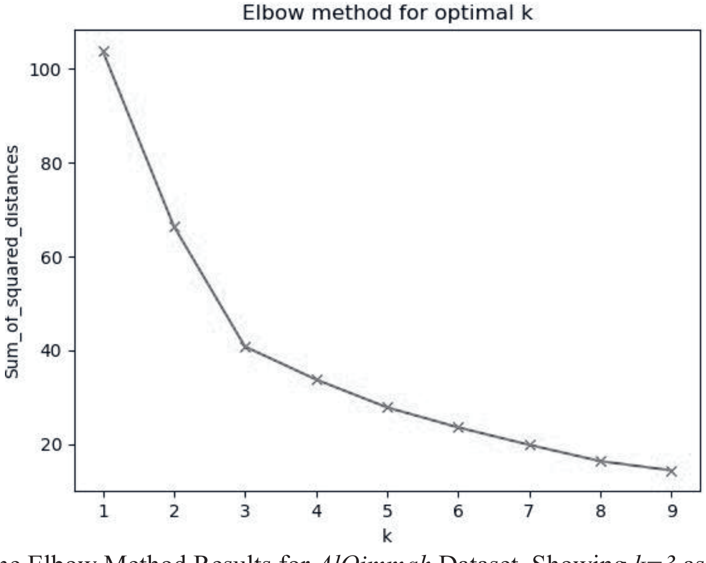


Figure 2. The Elbow Method Results for *AlQimmah* Dataset, Showing *k=3* as the Optimal Value

Similarly, the result of the Silhouette method, as shown in Figure 3, is three for *AlQimmah* dataset, having the highest score in the illustrated chart.

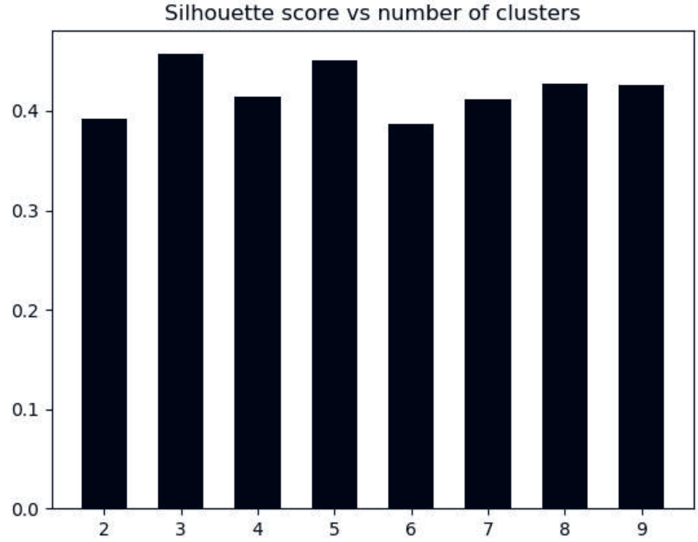
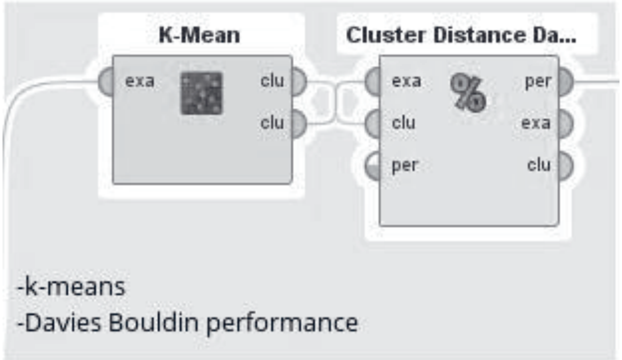


Figure 3. The Silhouette Method Results for *AlQimmah* Dataset, Showing *k=3* as the Optimal Value

4.2.4. Step 4: Applying K-Means and Validation Performance

K-Means clustering algorithm groups the data objects according to their proximity to each other. To compute the closeness of the data points, we test Euclidean Distance, Cosine Similarity, and Manhattan Distance measures, illustrated in the sub-process in Figure 4.



-k-means
-Davies Bouldin performance

Figure 4. K-Means Algorithm and Davies Bouldin Index Performance Measuring

4.2.5. Results

Table 1 illustrates the results of the described process applied to the three datasets (*AlFaloja*, *AlQimmah*, *AsAnsar*).

Dataset	**Elbow & Silhouette (*k*)**	**K-Means**		
		Metrics / Distance		
		Euclidean Distance	**Cosine Similarity**	**Manhattan Distance**
		Davies-Bouldin Index		
AlFaloja	4	0.220	0.226	0.227
AlQimmah	3	0.143	0143	0.141
AsAnsar	4	0.177	0.195	0.178

Table 1. The Value of k Based on Elbow, Silhouette, and Davies Bouldin with Different Metrics (Euclidean, Cosine, Manhattan)

As the table shows, the minimum values of the Davies-Bouldin Index for *AlFaloja* and *AsAnsar* datasets are achieved when using the Euclidean Distance (0.220 and 0.177, respectively), as the best number of clusters is where the average value of Davies-Bouldin Index is minimized [24].

For *AlQimmah* dataset, the scores of the Davies-Bouldin Index with Euclidean Distance and Cosine Similarity are the same, achieving the minimum value (0.141) when using the Manhattan Distance.

4.3. Text Preprocessing

We use a specialized component, Process Document from Data, to perform data preprocessing. This function transforms the textual data into word vectors by calculating the TF-IDF weights of the word set. Before implementing the TF-IDF function, we retrieve the dataset and convert the entries from Nominal to Text, as shown in Figure 5. After retrieving and converting the data (Nominal to Text), we select two attributes (one for *ThreadName*, refers to the subject or title, and one for *Message*, refers to the full topic) as we will discuss in the following sections. The “Replace” component is used to replace outliers or missing values, and the “Process Document from Data” component is used to perform data preprocessing, as illustrated in Figure 6.

In this section, we discuss the text preprocessing steps, illustrated in Figure 6, which include the following:

- Stop words removal
- Tokenization
- Filter tokens by length
- Filter tokens by content

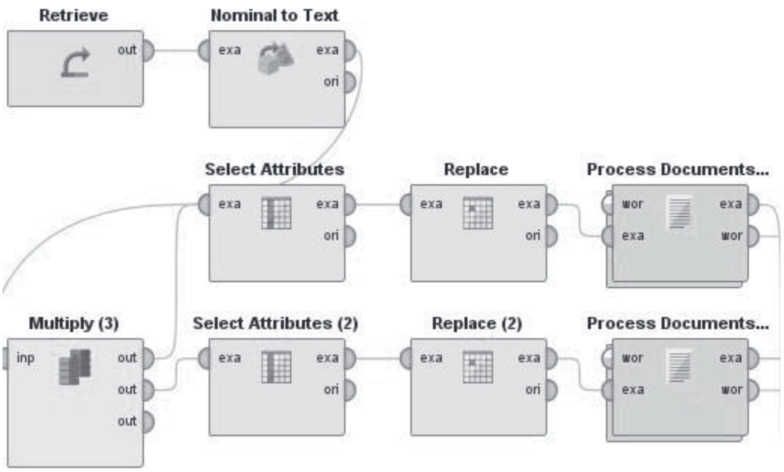


Figure 5. Implementation of Process Documents from Data, Selecting the ThreadName and Message Attributes

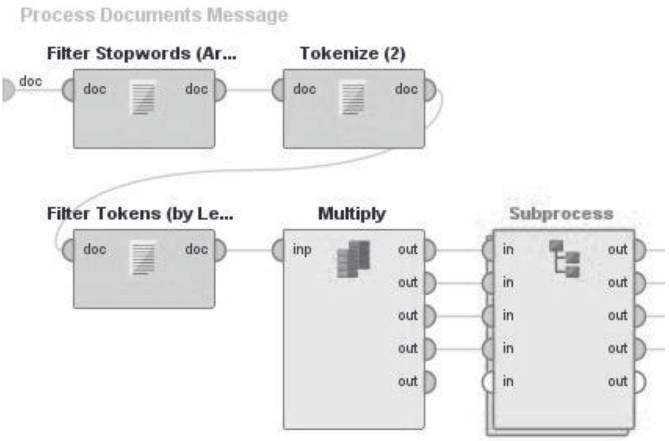


Figure 6. Process Documents for the *Message* Attribute

4.3.1. Stop Words Removal

This process (called *Filter Stopwords (Arabic)* in Figure 6) removes the common words or the words that have no significance in producing intelligence patterns or valuable information. Moreover, such terms may negatively affect the accuracy of the results. For example, the Arabic words "كم" ,"أما" ,"كل" are stop words. On RapidMiner, we can apply various functions, such as Filter Stopwords, based on the used language dictionary. We employ the Filter Stopwords process for the Arabic language.

4.3.2. Tokenization

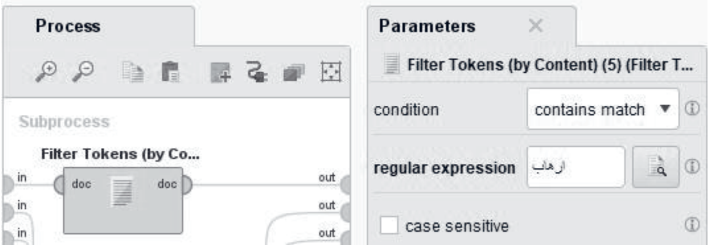
This process (called *Tokenize* in Figure 6) splits the sequence of strings into words [28]. It removes all the punctuations from the textual data and extracts words of the text, called tokens. RapidMiner tool provides three splitting methods; the default and commonly used one is the Non-Letter Character, which splits words based on the non-letters such as spaces, commas, full stops, and others[3].

4.3.3. Filter Tokens by Length

Through this function, we filter tokens of specific lengths. RapidMiner provides two parameters to restrict the lengths of the tokens, the maximum number of characters max chars and the minimum number of characters min chars [27]. Subsequently, the *Multiply* operator is used to make one input available for several outputs after processing. The *Subprocess* in Figure 6 executes multiple filters, which has the same behavior in one package that is divided into five functions (Filter Tokens by Length).

4.3.4. Filter Tokens by Content

This operator (Filter Tokens by Content) filters tokens based on their contents; in other words, it keeps a token according to defined criteria, such as equals, contains, or does not contain a given value. In our approach, we define a regular expression (condition =contains match) to match critical words such as "قتل = kill", "ارهاب = terrorism", or "تفجير = explosion". Figure 7 illustrates this process and its parameters.



Filter Tokens (by Content) (5) (Filter T...

condition: contains match

regular expression: ارهاب

☐ case sensitive

Figure 7. Filter Tokens by Content and its Parameters

4.3.5. Results

Figure 8 shows the counts of the critical words "الاسلحة = weapons", "التفجيرات = explosions" and "الارهاب = terrorism", represented by the *ThreadName* attribute,

[3] In our research, we noticed that Stemming might change the form of the word and make it ambiguous or meaningless, and it is critical in Dark Web analysis to understand the specific meaning of the word; thus, we preferred to keep the forms of the words as they are without Stemming, knowing that RapidMiner does not perform Lemmatization:

https://community.rapidminer.com/discussion/11285/lemmatization.

against members' usernames (from the *MemberName* attribute) in *AsAnsar* forum. The member "عبد الله الوزير" shared the highest number of these words.

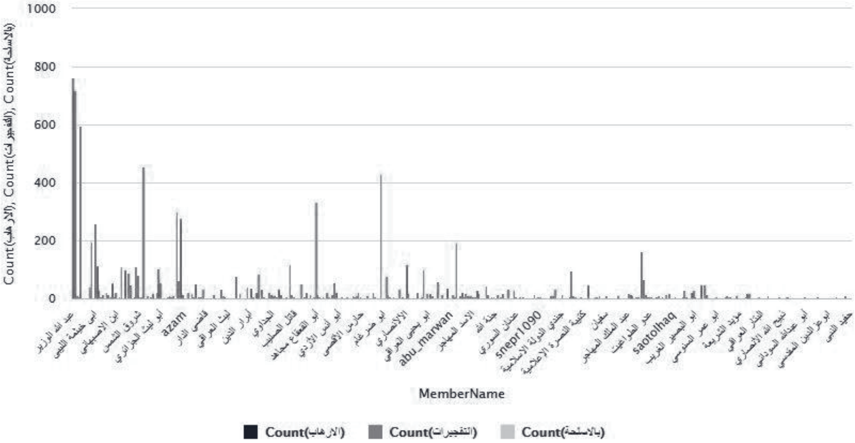


Figure 8. The Frequency of Occurrences of the Critical Words in *AsAnsar* Dataset after Selecting the *ThreadName* Attribute according to the Member Username (*MemberName*)

We sort the example set (selecting *ThreadName* attribute) by the month attribute (*P_Month*). Figure 9 shows that *May* witnessed the highest occurrences of the aforementioned critical words.

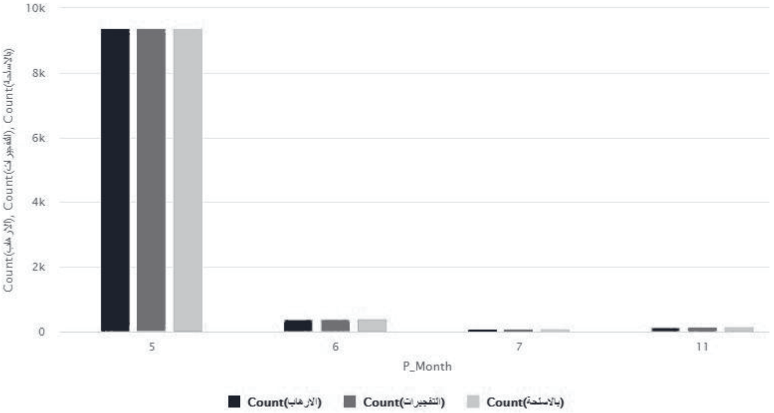


Figure 9. The Frequency of Occurrences of the Critical Words in *AsAnsar* Dataset after Selecting the *ThreadName* Attribute according to the Month of the Year (*P_Month*)

Similarly, Figure 10 also shows that the member "عبد الله الوزير" was the member who shared the highest number of the aforementioned critical words according to the Message attribute.

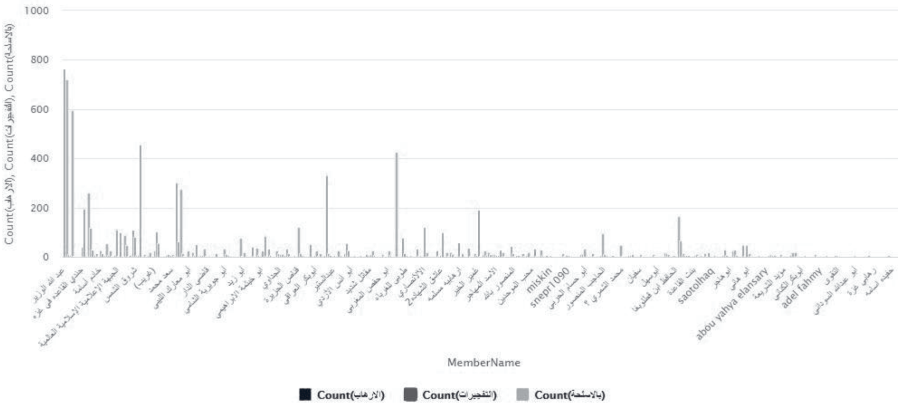


Figure 10. The Frequency of Occurrences of the Critical Words in *AsAnsar* Dataset after Selecting the *Message* Attribute according to the Member Username (*MemberName*)

In Figure 11, we sorted the example set by the year attribute (*P_Year*); thus, it shows that the aforementioned critical words converged in 2012 with excessive usage.

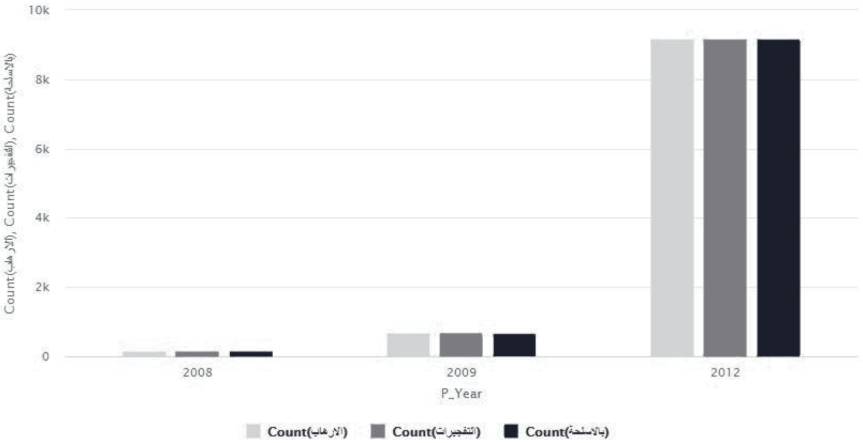


Figure 11. The Frequency of Occurrences of the Critical Words in *AsAnsar* Dataset after Selecting the *Message* Attribute according to the Year (*P_Year*)

5. Conclusion and Future Work

In this paper, we presented a Data Mining approach to analyze the contents of Dark Web terrorism forums, providing valuable insights that help investigators and law enforcement agencies to understand the nature of terrorist discussions and detect influential members.

Using the RapidMiner tool, we applied the K-means clustering method with different distance and similarity metrics (Euclidean Distance, Cosine Similarity, and Manhattan Distance). Our approach comprises several processes, including Outlier Detection (distance-based), Evaluation using Elbow and Silhouette methods to find the optimal number of clusters, and Validation using the Davies-Bouldin Index, as shown in Table 1.

Furthermore, we discussed the importance of text preprocessing and weighting, using TF-IDF, to extract meaningful information from the raw data. We analyzed forums posts according to two main attributes: the thread name (*ThreadName*) and the corresponding message (*Message*), and computed the frequency of occurrences of critical words related to terrorism according to The posting members (*MemberName*), month (*P_Month*), and year (*P_Year*), as shown in Figures 8-11.

Our approach helps investigators detect the most influential members and popular topics from terrorist forums on the Dark Web. In its future directions, we aim to develop our model to analyze larger volumes of data and compare other outlier detection methods. Furthermore, we consider extending the list of critical words and employing Association Rules to extract patterns and relationships among these words.

References

[1] R. Liggett, J. R. Lee, A. L. Roddy and M. A. Wallin, "The Dark Web as a Platform for Crime: An Exploration of Illicit Drug, Firearm, CSAM, and Cybercrime Markets," in The Palgrave Handbook of International Cybercrime and Cyberdeviance, T. J. Holt and A. M. Bossler, Eds., Palgrave Macmillan, Cham, 2020, pp. 91-116. doi:10.1007/978-3-319-78440-3_17.

[2] G. Alrasheed and B. Rigato, "Exploring the Dark Web: Where Terrorists Hide?," 5 February 2019. [Online]. Available: https://carleton.ca/align/2019/illuminate-exploring-the-dark-web-where-terrorists-hide/. [Accessed 28 December 2021]. doi:10.26650/JPLC2020-813328.

[3] B. Alkhatib and R. S. Basheer, "Mining the Dark Web: A Novel Approach for Placing a Dark Website under Investigation," International Journal of Modern Education and Computer Science (IJMECS), vol. 11, no. 10, pp. 1-13, 2019. doi:10.5815/ijmecs.2019.10.01.

[4] J. K. Saini and D. Bansal, "Detecting online recruitment of terrorists: towards smarter solutions to counter terrorism," International Journal of Information Technology, vol. 13, no. 2, pp. 697-702, 2021.

[5] S. Baghel and Yogesh, "Detecting Future Terrorism Trend in India Using Clustering Analysis," in 2018 7th International Conference on Reliability, Infocom Technologies and Optimization (Trends and Future Directions) (ICRITO), Noida, India, 2018. doi:10.1109/ICRITO.2018.8748567.

[6] M. Goel, N. Sharma and M. K. Gurve, "Analysis of Global Terrorism Dataset using Open Source Data Mining Tools," in 2019 International Conference on Computing, Power and Communication Technologies (GUCON), New Delhi, India, 2019.

[7] D. D. Atsa'am, R. Wario and F. E. Okpo, "A New Terrorism Categorization Based on Casualties and Consequences Using Hierarchical Clustering," Journal of Applied Security Research, vol. 15, no. 3, pp. 369-384, 2020. doi:10.1080/19361610.2020.1769461.

[8] V. Kumar, M. Mazzara, A. Messina and J. Lee, "A Conjoint Application of Data Mining Techniques for Analysis of Global Terrorist Attacks – Prevention and Prediction for Combating Terrorism," in Proceedings of 6th International Conference in Software Engineering for Defence Applications, SEDA 2018. Advances in Intelligent Systems and Computing, vol 925, Rome, Italy, 2020. doi:10.1007/978-3-030-14687-0_13.

[9] R. M. Alguliyev, R. M. Aliguliyev and G. Y. Niftaliyeva, "Filtration of Terrorism-Related Texts in the E-Government Environment," International Journal of Cyber Warfare and Terrorism, vol. 8, no. 4, pp. 35-48, 2018. doi:10.4018/IJCWT.2018100103.

[10] U. S. Onyekachi, T. Norbert and E. D. Uzoamaka, "Data Mining Approach to Counterterrorism," Computing, Information Systems, Development Informatics & Allied Research Journal, vol. 11, no. 2, pp. 77-90, 2020. doi:10.1007/s41870-021-00620-2.

[11] I. V. Mashechkin, M. I. Petrovskiy, D. V. Tsarev and M. N. Chikunov, "Machine Learning Methods for Detecting and Monitoring Extremist Information on the Internet," Programming and Computer Software, vol. 45, no. 3, pp. 99-115, 2019. doi:10.1134/S0361768819030058.

[12] G. M. Soliman and T. H. Abou-El-Enien, "Terrorism Prediction Using Artificial Neural Network," Revue d'Intelligence Artificielle, vol. 33, no. 2, pp. 81-87, 2019. doi:10.18280/ria.330201.

[13] N. Tavabi, N. Bartley, A. Abeliuk, S. Soni, E. Ferrara and K. Lerman, "Characterizing Activity on the Deep and Dark Web," in WWW '19: Companion Proceedings of The 2019 World Wide Web Conference, San Francisco, USA, 2019. doi:10.1145/3308560.3316502.

[14] A. Et-taleby, M. Boussetta and M. Benslimane, "Faults Detection for Photovoltaic Field Based on K-Means, Elbow, and Average Silhouette Techniques through the Segmentation of a Thermal Image," International Journal of Photoenergy, 2020. doi:10.1155/2020/6617597.

[15] E. C. ATEŞ, E. BOSTANCI and M. S. GÜZEL, "Big Data, Data Mining, Machine Learning, and Deep Learning Concepts in Crime Data," Ceza Hukuku ve Kriminoloji Dergisi-Journal of Penal Law and Criminology, vol. 8, no. 2, pp. 293-319, 2020.

[16] J. Cheng and L. Zhang, "Jaccard Coefficient-Based Bi-clustering and Fusion Recommender System for Solving Data Sparsity," in Advances in

Knowledge Discovery and Data Mining. PAKDD 2019. Lecture Notes in Computer Science, Macau, China, 2019. doi:10.1007/978-3-030-16145-3_29.

[17] D. M. Saputra, D. Saputra and L. D. Oswari, "Effect of Distance Metrics in Determining K-Value in K-Means Clustering Using Elbow and Silhouette Method," in Sriwijaya International Conference on Information Technology and Its Applications (SICONIAN), Indonesia, 2020.

[18] Kanika, K. Rani, Sangeeta and Preeti, "Visual Analytics for Comparing the Impact of Outliers in k-Means and k-Medoids Algorithm," in Amity International Conference on Artificial Intelligence (AICAI), Dubai, United Arab Emirates, 2019. doi:10.1109/AICAI.2019.8701355.

[19] Y. Zhao, Z. Nasrullah and Z. Li, "PyOD: A Python Toolbox for Scalable Outlier Detection," Journal of Machine Learning Research, vol. 20, pp. 1-7, 2019.

[20] X. Xu, H. Liu, L. Li and M. Yao, "A Comparison of Outlier Detection Techniques for High-Dimensional Data," International Journal of Computational Intelligence Systems, vol. 11, no. 1, pp. 652-662, 2018. doi:10.2991/ijcis.11.1.50.

[21] H. C. Mandhare and S. R. Idate, "A comparative study of cluster based outlier detection, distance based outlier detection and density based outlier detection techniques," in 2017 International Conference on Intelligent Computing and Control Systems (ICICCS), Madurai, India, 2017. doi:10.1109/ICCONS.2017.8250601.

[22] A. Smiti, "A critical overview of outlier detection methods," Computer Science Review, vol. 38, p. 100306, 2020. doi: 10.1016/j.cosrev.2020.100306.

[23] A. C. Benabdellah, A. Benghabrit and I. Bouhaddou, "A survey of clustering algorithms for an industrial context," Procedia Computer Science, vol. 148, pp. 291-302, 2019. doi:10.1016/j.procs.2019.01.022.

[24] M. Mughnyanti, S. Efendi and M. Zarlis, "Analysis of determining centroid clustering x-means algorithm with davies-bouldin index evaluation," in 3rd Nommensen International Conference on Technology and Engineering 2019 (3rd NICTE), Indonesia, 2020. doi:10.1088/1757-899X/725/1/012128.

[25] I. U. Sari, D. Sergi and B. Ozkan, "Customer Segmentation Using RFM Analysis: Real Case Application on a Fuel Company," in Application of Big Data and Business Analytics, S. Kumari, K. K. Tripathy and V. Kumbhar, Eds., Emerald Publishing Limited, Bingley, 2020, pp. 139-158. doi: 10.1108/978-1-80043-884-220211009.

[26] L. He, B. Agard and M. Trépanier, " A classification of public transit users with smart card data based on time series distance metrics and a hierarchical clustering method," Transportmetrica A: Transport Science, vol. 16, no. 1, pp. 56-75, 2020. doi:10.1080/23249935.2018.1479722.

[27] V. Kalra and R. Aggarwal, "Importance of Text Data Preprocessing & Implementation in RapidMiner," in Proceedings of the First International Conference on Information Technology and Knowledge Management (ICITKM), New Delhi, India, 2017. doi:10.15439/2018KM46.

[28] S. A. Salihu, I. P. Onyekwere, M. A. Mabayoje and H. A. Mojeed, "Performance Evaluation Of Manhattan And Euclidean Distance Measures For Clustering Based Automatic Text Summarization," Journal of Engineering and Technology, vol. 4, no. 1, pp. 135-139, 2019.