Detection of Arabic Terrorism Keyword in Websites

Arief Rahmana[1] , Madihah Mohd Saudi [2, *], Widyatama[3] , Hassan Awad Hassan Al-Sukhni[4] , Azuan Ahmad[5]

[1]Widyatama University
[2]CyberSecurity and Systems (CSS) Research Unit, Faculty of Science and Technology (FST), Universiti Sains Islam Malaysia (USIM), 71800 Nilai, Negeri Sembilan, Malaysia
[3]Widyatama University
[4]CyberSecurity and Systems (CSS) Research Unit, Faculty of Science and Technology (FST), Universiti Sains Islam Malaysia (USIM), 71800 Nilai, Negeri Sembilan, Malaysia
[5]CyberSecurity and Systems (CSS) Research Unit, Faculty of Science and Technology (FST), Universiti Sains Islam Malaysia (USIM), 71800 Nilai, Negeri Sembilan, Malaysia
[2]madihah@usim.edu.my

Article History: Received: 10 January 2021; Revised: 12 February 2021; Accepted: 27 March 2021; Published online: 20 April 2021

Abstract: This paper presents a detection technique for cyberterrorist suspected activities in the websites by using Krill Herd and Simulated Annealing algorithms. There were 1,700 dataset used with 10 cross-validation for training and testing from Tawhid, Islamion and Alemarah News websites. Based on the experiment conducted, this proposed detection technique has produced 84.9% accuracy rate, which has outperformed with the benchmark work. This paper could be used as guidance other researchers with the same interest.

Keywords: Krill Herd, Arabic web content classification, cyber terrorists, simulating annealing.

1. Introduction

Nowadays there are so many information could be accessed from the websites. The most worrying would be on the increase of cyber-terrorist activities in websites. Nowadays, information distributed electronically comprising of terrorism content and keywords are considered to be a driving factor in the recruitment and radicalisation of innocent civilians globally. Further, information may be used to influence radicalised behaviour to become actively involved in terrorism activities in other countries and even decide to fight with other terrorism groups overseas. Additionally, the vast amount of online data has made it virtually impossible for authorities to investigate online media for examples Facebook, web pages and messaging possibly related to terrorism activities or containing terrorism propaganda and other content. Based on works by [1,2], this paper defines a terrorist as any action is intended to harm others based on any ideological motivation to justify their crimes. It is called as cyber-terrorist when the terrorist uses the Internet to spread his ideology that threatens and caused violence. Lack of expertise and efficient technique to detect Arabic word in website that is related or suspected with terrorist activities is the urge of this paper. This paper aims to detect terrorism web content by extracting terrorism keywords and to reduce a large number of features by applying krill herd and simulating annealing.

This paper is presented based on the following sections. Section 2 discusses the main related works in the domain of feature selection (FS) and cyber terrorists. Section 3 discusses the methodology, while section 4 presents the finding for this paper. Section 5 concludes this paper together with future work.

2. Related Works

The extraction process in this paper aims to extract the most related words from text based on certain methods, such as by extracting the sense of each word using Arabic WordNet, which is the database of the vocabulary and senses of words [3]. Another example of the extraction process is the extraction of all words following a pre-processing stage called the “Bag of Words”. On the other hand, training features called feature selection is a method used to reduce the number of features (terms or words) which are unimportant or irrelevant to the topic. Moreover, the more popular feature selection is called meta-heuristic feature selection, which is split into two parts: local search and global search optimisation feature selection. As an example of global search optimisation is harmony search (HS) and krill herd (KH). The KH is designed to determine the optimal solution by increasing krill density and achieving the optimised solution of features. After that, the content classifier process can be employed to classify or label the new, unseen word problems associated with ‘terrorism’ or ‘non-terrorism’.

Typically, two methods are used to represent a ‘text’ as a set of features, which are (i) Bag-Of Words (BOW), which involves using single words or phrases as features or n-gram, (ii) Word Level n-gram, which involves using a sequence of words or characters (Character Level n-gram) of length n [4]. However, one of the weaknesses that normally arises from building the text classifier (TC) system is associated with handling the enormous number of features, which can easily reach a magnitude in the tens of thousands [5-7]. Accordingly, to reduce this feature space dimension, many information retrieval techniques have been used, ranging from Stemming, Stop-words Removal and Feature Selection (FS). FS techniques, such as Mutual Information (MI), Chi-Square Statistic (CHI), Information Gain (IG), GSS Coefficient (GSS) and Odds Ratio (OR) are used to reduce the dimensionality of the feature space through eliminating the features that are considered irrelevant for a particular category [8-13]. However, FS is characterised with certain weaknesses in ignoring the relationship between the features [14]. Hence this paper aims to increase the accuracy of the Arabic web content of terrorism keywords extraction and improve the FS is by combining krill herd and simulated annealing.

Currently, there has been an increasing trend related to the web globally, associated with the classification of data and in determining a predefined category of natural language webs (content web classification) [14]. The categorisation of news articles into topics, such as politics and sports is a classic example of content classification. Notwithstanding, the efficiency of a classification system is based on the performance of the extraction process and FS employed [15]. As such, gathering a magnitude of terms for extraction, features can consume much time and be expensive. Therefore, the present study will select examples from pools that are labelled using enhanced content classifiers instead of extracting all terms. In other words, this research extract 'the most related words to terrorism webs' and uses FS for labelling in building the training set [14, 16].

The text classification is part of Machine learning (ML) can be split into two parts: supervised and unsupervised learning [14]. Previous studies have indicated that both learning methods are used for ML in which the former, can be described as the task of generating mapping from a supervised or labelled training data to an output of classes or predictions. The main application of this type of learning is in the classification of tasks, where the goal is to create a function from the input objects to output values called labels or classes. In the classification of tasks using the ML approach, the main notion is to collect a set of examples, including classes by manually labelling some the examples as terrorism or non-terrorism. The set of labelled examples is known as a training set of terms upon which the classifier will be used with the training set of features in order to generate a mapping from the examples that have been labelled [7-13,15-17].

Based on the challenges highlighted and discussed, this paper has proposed krill herd and simulating annealing to detect terrorism web content by extracting terrorism keywords.

3. Methods

The whole steps involved in this paper as displayed in Figure 1 and Figure 2. The experiment is simulated using MATLAB, with the results evaluated based on the accuracy rate.

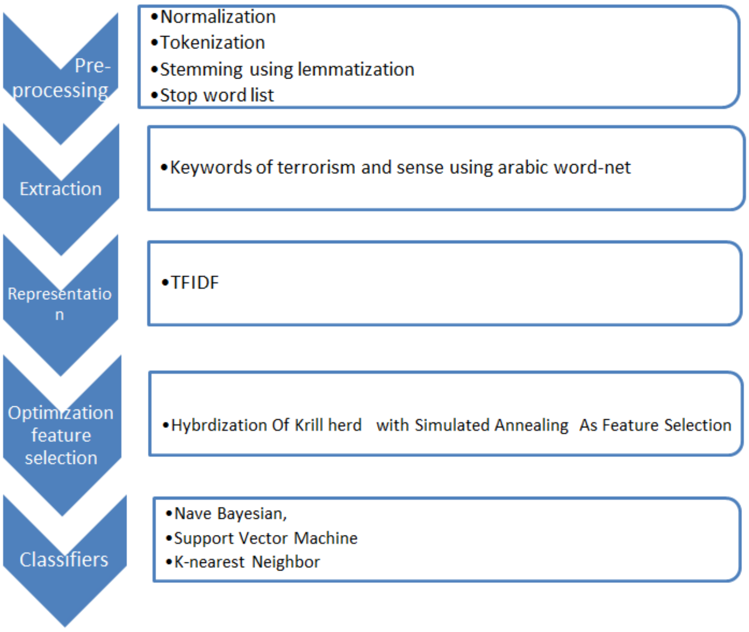
Pre-processing
•Normalization
•Tokenization
•Stemming using lemmatization
•Stop word list
Extraction
•Keywords of terrorism and sense using arabic word-net
Representation
•TFIDF
Optimization feature selection
•Hybrdization Of Krill herd with Simulated Annealing As Feature Selection
Classifiers
•Nave Bayesian,
•Support Vector Machine
•K-nearest Neighbor

Figure 1. Flowchart of the Arabic web content detection

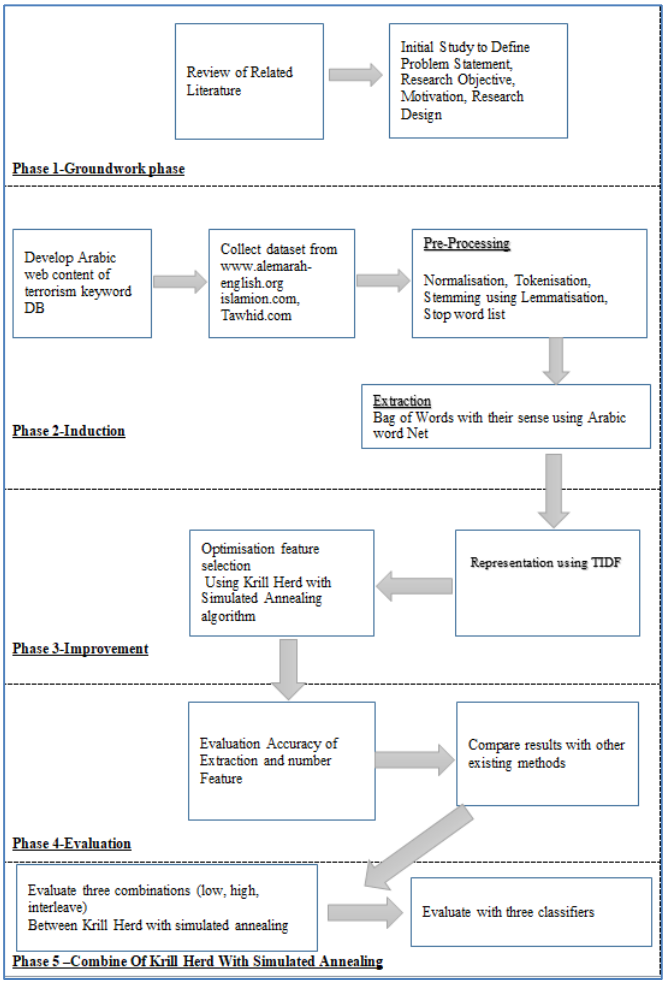
Review of Related Literature
Initial Study to Define Problem Statement, Research Objective, Motivation, Research Design
Phase 1-Groundwork phase
Develop Arabic web content of terrorism keyword DB
Collect dataset from www.alemarah-english.org islamion.com, Tawhid.com
Pre-Processing
Normalisation, Tokenisation, Stemming using Lemmatisation, Stop word list
Extraction
Bag of Words with their sense using Arabic word Net
Phase 2-Induction
Optimisation feature selection Using Krill Herd with Simulated Annealing algorithm
Representation using TIDF
Phase 3-Improvement
Evaluation Accuracy of Extraction and number Feature
Compare results with other existing methods
Phase 4-Evaluation
Evaluate three combinations (low, high, interleave) Between Krill Herd with simulated annealing
Evaluate with three classifiers
Phase 5 –Combine Of Krill Herd With Simulated Annealing

Figure 2. Summarisation of the Methods Involved.

4. Findings

This section assesses the keywords with a sense of BOW that were employed for the extraction, using the krill herd combined with simulated annealing as a feature selection (interleave). However, the experiments and evaluations conducted in this research were based on the actual dataset. The results obtained in this particular study were made available by the average of all findings from executing the algorithm 20 times (in drawing a fair comparison). To make the comparison, the algorithm was applied 1,000 times [iterations] for all the respective runs. The parameter for KH, for each dataset, parameter was set to NK = 6, and the Vf = 0.2, which is the Foraging Motion. Table 1 shows a summary of the results of the new number of features using SVM classifier selections and the proposed extraction of using keywords and sense of BOW. Combining KH with SA and the best results for the number of feature selections compared with other optimization as feature selection is shown.

Table 1 Summary of the New Number of Features using SVM Classifier Selections and Proposed Extraction

		interleave	KH	SA	GA	PSO	HS
Dataset	Original # of features	New # of features	New # of features	New # of features	New # of features	New # of features	New # of features
DS1	**10880**	**5943**	6498	7157	7245	7256	8034
DS2	**12923**	**7136**	9367	8927	9128	9587	7983
DS3	**10855**	6829	**6726**	8469	7483	7106	9438

Table 2 shows a summary of the results for the SVM classifier's performance using feature selections and the proposed extraction method. The results show that the accuracies achieved with interleave acquired the best accuracy of 84.9% for the DS2 dataset. The second-ranking was for the interleave based on the dataset DS1 with an accuracy of 80.1%.

Table 2 The Performance of SVM Classifier using Feature Selections and Proposed Extraction

	interleave	KH	SA	GA	PSO	HS
Dataset	accuracy					
DS1	**80.10**	77.61	60.80	60.88	59.40	59.00
DS2	**84.90**	73.12	73.01	72.34	71.55	72.01
DS3	**75.10**	70.64	71.98	68.41	68.10	66.30

Table 3 shows a summary of the results of the new number of features using the KNN classifier selections and the proposed extraction of using keywords and sense of BOW. The interleave of combining KH with SA and the best number of feature selections is compared with other optimizations.

Table 3 Summary of the New Number of Features using KNN Classifier and Proposed Extraction

		interleave	KH	SA	GA	PSO	HS
Dataset	Original # of features	New # of features					
DS1	10880	**6121**	6301	7291	7302	7228	8671
DS2	12923	**7490**	9201	8990	9098	9560	7871
DS3	10855	**7010**	6809	8450	7208	7202	8560

Table 4 shows a summary of the KNN classifier's performance using feature selections and the proposed extraction method. The results show that the accuracies achieved with interleaving had the best accuracy score of 79.8% for the DS1 dataset. The second-ranking was for the interleave, based on the dataset DS3 with an accuracy of 74.8%. The table shows that the interleave scored in DS2 was 70.21%.

Table 4 The Performance of KNN Classifier using Feature Selections and Proposed Extraction

	interleave	KH	SA	GA	PSO	HS
Dataset	accuracy					
DS1	**79.8**	73.66	58.02	60.1	62.27	61.2
DS2	**70.21**	67.91	59.9	59.29	61	65.39
DS3	**74.8**	71.02	70.9	70.1	70.9	70.9

Table 5 shows a summary of the new number of features using NB classifier selections and the proposed extraction method of using keywords and sense of BOW. The interleave of combining KH with SA was the best in the number of feature selections compared with other optimizations as feature selection.

Table 5 Summary of the New Number of Features using NB Classifier Selections and Proposed Extraction

		interleave	KH	SA	GA	PSO	HS
Dataset	Original # of features	New # of features					

DS1	10880	**6224**	648 7	721 0	730 1	730 7	813 4
DS2	12923	**7321**	941 7	910 4	920 9	961 4	799 7
DS3	10855	**7109**	670 7	870 1	800 9	729 6	956 0

Table 6 displays a summary of the NB classifier's performance using feature selections and the proposed extraction method. The results show that the interleave accuracy achieved with the best accuracy of 80.3% for the DS2 dataset. The second-ranking was for the interleave based on the dataset DS1 having an accuracy of 78.1%. The table shows that the interleave scored in DS3 was 70.9%. The best minimum number of features with accuracy was related to the interleaving, as mentioned in Table 18.

Table 6 The Performance of NB Classifier using Feature Selections and Proposed Extraction

	interleave	KH	SA	GA	PSO	HS
Dataset	accuracy					
DS1	**78.1**	72.34	66.4	63.8	62.3	68.9
DS2	**80.3**	66.8	58.87	59.86	63.4	67.51
DS3	**74.9**	72.89	70.97	71.7	70.1	71.1

Based on Tables 1 to 6, the SVM classifier was better than the other classifiers and interleave as a feature selection better than other optimization as feature selections. On the other hand, the combination between proposed extractions on the proposed feature selection enhanced the content classifier's performance of Arabic web content classification. The major portions covered in this section determined all the near-optimal features and the selection of all the optimal features. Regarding the fitness function, which is given in the criteria, determining all the features' values was available in a specific classifier for the different classes and category types.

5. Conclusions

In considering all the algorithms, SVM had the best performance, and NB was better than KNN. Depending upon the behavior of all classifiers that were obtained, the meta-heuristic, such as HS feature selection, helped determine the problem and resolve the gaps. The HS algorithm HS's reaction was demonstrated in this paper, which was based on the fitness function as the evaluation measure. Here, it was found that the high level combining KH with SA had the worst performance compared with the low level of combining KH with SA and interleaves of combining KH with the SA proposed feature selection. A conclusion can also be made from this particular observation, revealing that content classification can be performed well with all optimal features generated using the interleave. This paper suggests some directions for future research; future studies should try to find out the results for the new stages available in the extraction methods called the pragmatic of 'Bag-of-Narratives,' focusing on conweb-aware and intent-driven. These pragmatic curves play an important role as a key for analyzing tasks such as sentiment analysis, known as a concept in which a negative connotation is generally taken into account. Therefore, future research could find a new way to find the best parameters setting, which is better than a scenario.

Acknowledgment

The authors would like to express their gratitude to Widyatama University, Indonesia, and Universiti Sains Islam Malaysia (USIM) (USIM grant no: PPPI/UGC_0119/FST/051000/13519) for the funding, support, and facilities provided.

References

1. Al Mazari, A., Anjariny, A.H., Habib, S.A. and Nyakwende, E. Cyber terrorism taxonomies: Definition, targets, patterns, risk factors, and mitigation strategies. In Cyber Security and Threats: Concepts, Methodologies, Tools, and Applications (pp. 608-621). IGI Global.2018
2. Holbrook, Donald. The Al-Qaeda Doctrine. London: Bloomsbury Publishing. pp. 30ff, 61ff, 83ff. ISBN 978-1623563141.2014

3. Bsoul, Q., Al-Shammari, E., Mohd, M. and Atwan, J., "Distance Measures and Stemming Impact on Arabic Document Clustering|. In Asia Information Retrieval Symposium. Springer, Cham. (pp. 327-339).2014
4. El-Kourdi, M., A. Bensaid & T. Rachid. Automatic Arabic document categorization based on the Naïve Bayes Algorithm. Proceedings of the Workshop on Computational Approaches to Arabic Script-Based Languages, (CAASL' 04), Stroudsburg, PA, USA, pp. 51-58.2004
5. Al-Harbi, S., A. Almuhareb, A. Al-Thubaity, M.S. Khorsheed & A. Al-Rajeh. Automatic Arabic Text Classification. 9es journées internationales analyse statistique des données textuelles, JADT, 08: 77-83.2008
6. Eyheramendy, S., Lewis, D. & Madigan , D. On the naive bayes model for text categorization. Citrseerex. 2003
7. Dada, E.G., Bassi, J.S., Chiroma, H., Adetunmbi, A.O. and Ajibuwa, O.E.. Machine learning for email spam filtering: review, approaches and open research problems. Heliyon, 5(6), p.e01802.2019
8. Al-Tashi, Q., Kadir, S.J.A., Rais, H.M., Mirjalili, S. and Alhussian, H. Binary Optimization Using Hybrid Grey Wolf Optimization for Feature Selection. IEEE Access, 7, pp.39496-39508.2019
9. Gupta, S., Khattar, A., Gogia, A., Kumaraguru, P., & Chakraborty, T.Collective classification of spam campaigners on Twitter: A hierarchical meta-path based approach. In Proceedings of the 2018 World Wide Web Conference on World Wide Web (pp. 529-538). International World Wide Web Conferences Steering Committee.2018
10. Wang, F., Xu, T., Tang, T., Zhou, M. & Wang, H.Bilevel feature extraction-based text mining for fault diagnosis of railway systems. IEEE Transactions on Intelligent Transportation Systems, 18(1):49-58.2017.
11. Jain, G., Sharma, M., & Agarwal, B. Spam detection on social media using semantic convolutional neural network. International Journal of Knowledge Discovery in Bioinformatics (IJKDB), 8(1), 12-26.2018
12. Trivedi, S. K., & Panigrahi, P. K. Spam classification: a comparative analysis of different boosted decision tree approaches. Journal of Systems and Information Technology, 20(3), 298-105.2018
13. Saha, Soumyabrata, Suparna DasGupta, and Suman Kumar Das. "Spam Mail Detection Using Data Mining: A Comparative Analysis." In Smart Intelligent Computing and Applications, pp. 571-580. Springer, Singapore.2019
14. Mafarja, M.M. & Mirjalili, S. Hybrid Whale Optimization Algorithm with simulated annealing for feature selection. Neurocomputing.2017
15. Ghareb, S., Bakar, A. & Hamdan, R. Hybrid feature selection based on enhanced genetic algorithm for text categorization. Expert Systems with Applications, 49, pp. 31-47. 2016
16. Wang, Y., Liu, Y., Feng, L. & Zhu, X. Novel feature selection method based on harmony search for email classification. Knowledge-Based Systems, pp.311-323.2015
17. Aghdam, H. & Heidari, S. Feature selection using particle swarm optimization in text categorization. Journal of Artificial Intelligence and Soft Computing Research, 5(4): 231-238.2015.