Computers, Materials & Continua

Tech Science Press

DOI: 10.32604/cmc.2023.040856

ARTICLE

Multiclass Classification for Cyber Threats Detection on Twitter

Adnan Hussein[1] **and Abdulwahab Ali Almazroi**[2,*]

[1]College of Computer Science and Engineering, Department of Computer Science, AL-Ahgaff University, Mukalla, Yemen

[2]College of Computing and Information Technology at Khulais, Department of Information Technology, University of Jeddah, Jeddah, Saudi Arabia

*Corresponding Author: Abdulwahab Ali Almazroi. Email: aalmazroi@uj.edu.sa

Received: 01 April 2023 Accepted: 25 July 2023 Published: 26 December 2023

ABSTRACT

The advances in technology increase the number of internet systems usage. As a result, cybersecurity issues have become more common. Cyber threats are one of the main problems in the area of cybersecurity. However, detecting cybersecurity threats is not a trivial task and thus is the center of focus for many researchers due to its importance. This study aims to analyze Twitter data to detect cyber threats using a multiclass classification approach. The data is passed through different tasks to prepare it for the analysis. Term Frequency and Inverse Document Frequency (TFIDF) features are extracted to vectorize the cleaned data and several machine learning algorithms are used to classify the Twitter posts into multiple classes of cyber threats. The results are evaluated using different metrics including precision, recall, F-score, and accuracy. This work contributes to the cyber security research area. The experiments revealed the promised results of the analysis using the Random Forest (RF) algorithm with (F-score = 81%). This result outperformed the existing studies in the field of cyber threat detection and showed the importance of detecting cyber threats in social media posts. There is a need for more investigation in the field of multiclass classification to achieve more accurate results. In the future, this study suggests applying different data representations for the feature extraction other than TF-IDF such as Word2Vec, and adding a new phase for feature selection to select the optimum features subset to achieve higher accuracy of the detection process.

KEYWORDS

Cybersecurity; cyber threat detection; artificial intelligence; machine learning; Twitter

1 Introduction

With the advancement in technology and electronic systems, people's lives have become almost completely dependent on technology because of the ease and speed of information transmission [1]. On the other hand, there are some challenges and issues that have emerged with this advancement in technology. One of these challenges is detecting and mitigating security threats that try to penetrate the cyber systems and manipulate or destroy the system to achieve illegal goals [2,3]. Recently, these challenges have increased, and there is a need to stop these attacks and threats. Cybersecurity threat detection is a field of study that detects cyber threats to stop them and mitigate them [4].

Recently, there is integrity between several kinds of sciences to solve the issues and improve performance [5]. Machine learning is one of the promising fields in artificial intelligence science, which is changed many human life aspects by integrating it into other fields. This integrity creates a notable advancement in many technological aspects of human life. The use of artificial intelligence and machine learning in cybersecurity and the detection of cyber threats is important to be studied and investigated [3]. Some recently published research has indicated that the use of artificial intelligence techniques in cybersecurity has contributed to solving multiple problems and contributed to improving the current security systems [6]. Thus, there is a need to expand the research in this area and investigate more in this field.

Social networks such as Twitter, Facebook, Instagram, and others have contributed significantly in the last few decades to connecting the world and communities. With social networks, dissemination of news and information has become easy, cheap, and very fast, an event occurs in the West at a specific time, and people in the East see it and interact with it at the same moment of occurrence. This is considered a revolution in the world of data transmission [1]. There are many people with bad intentions trying to use these platforms to spread conflict and inconvenience, as well as to destabilize the security of others on multiple levels, including the individual, the community, and the level of companies or countries. Confronting such threats and risks requires reliable studies that detect these threats before they spread, expose the identity of the attackers, and stop them thus limiting the spread of threats.

This study uses artificial intelligence and machine learning techniques to detect cybersecurity threats using social networking data taken from Twitter. The next section includes an explanation of the research literature. After that, the research methodology pipeline is explained in detail for each different practical stage for the data analysis. The following section includes a discussion of the experiment results and finally the conclusion and future work suggestions.

2 Literature Review

Machine learning is a subfield of artificial intelligence which focuses on using data and algorithms to simulate how humans learn. The main purpose of using machine learning in different systems is to increase and refine the outcomes of these systems. There are three main approaches to machine learning algorithms: supervised machine learning, unsupervised machine learning, and reinforcement machine learning [3].

Recently, machine learning has been integrated with other sciences and it has achieved a notable improvement in several fields such as healthcare sciences, military sciences, economic sciences, and so on. Cybersecurity is one of the fields that is concerned with defending against the threats and attacks of the cyber systems, machine learning is used to help in several cybersecurity tasks such as detection, mitigation, and prevention to build strong and accurate cybersecurity systems. In literature, there are a few works have been published to tackle the integration between cybersecurity and machine learning and there is a need to investigate more in this field.

Silvestri et al. in [7] analyzed a healthcare dataset collected from hackers' news websites. This study adopted a machine learning model named Bidirectional Encoder Representations from Transformers (BERT) and eXtreme Gradient Boosting (XGBoost) and extracted information from the available documents on the web to determine the threats and vulnerabilities that impact the healthcare system. The results show an accuracy of 0.99.

On the other hand, the study [8] studied the Network traffic data to detect cybersecurity threats in the Internet of Things (IoT) Environment. The study used the N-BaIoT dataset and applied a supervised machine learning approach using Regularized Extreme Learning Machine (Regularized RELM) and Mafly optimization. The system achieved 98.93, 98.95, and 98.94 precision, recall, and F-measure, respectively.

Shaukat et al. in [9] applied a supervised machine learning approach using Support Vector Machine (SVM), Deep Belief Network (DBN), Decision Tree (DT) models to evaluate the cyber threats detection on several different datasets Knowledge Discovery and Data Mining Tools Competition dataset (KDD CUP 99), Spambase, Twitter dataset, Enron, Neural Structured Learning–Knowledge Discovery in Databases (NSL-KDD), Defense Advanced Research Projects Agency (DARPA), and malware datasets. The results of this study show that using a specific learning technique for a particular cyber threat detection is not recommended. In addition, there is a need to investigate and analyze different learning models for cyber threat detection and build a benchmark dataset for that purpose.

Ke [10] studied some different machine learning methods to detect the threats in cybersecurity based. The study analyzed the NSL-KDD dataset using supervised machine learning models: Logistic Regression (LR) Multi-Layer Perceptron (MLP) and SVM and deep learning models: DBN and Stacked Non-Symmetric Deep Auto-Encode (SNSDAE). The results show that deep learning performs better than shallow learning to detect cybersecurity threats.

Moreover, the study [11] applied shallow supervised machine learning and deep learning models to examine its performance on three different cybersecurity issues: spam detection, intrusion detection, and malware analysis. The analyses used RF model and the Feedforward Deep Neural Network model on DGA datasets. The results show that there is a need for more investigation on using machine learning to deal with cybersecurity sensitive problems and the results still show some shortcomings that limit the applicability of machine learning models on cybersecurity issues. Nevertheless, the study expected that deep learning could have improved to deal with cybersecurity in the future.

Table 1 gives a summary of the most important literary works in the field of cyber threat detection using machine learning algorithms.

Table 1: Cyber threats detection literature works summary

Ref.	Year	Data	Domain	Algorithms	Result
[7]	2023	CS news posts in thehackernews.com	Healthcare	BERT	Precision = 96.62 Recall = 79.95 F1-score = 87.50 Accuracy = 99.75
[8]	2022	N-BaIoT dataset	Network flows	Regularized RELM and Mafly Optimization	Precision = 98.93 Recall = 98.95 F1-score = 98.94
[12]	2022	MAWI, NCCDC, ISOT, ISCX, OIF and Codex	Multidomain	K-Nearest Neighbor (KNN)	Precision: 0.72–0.97 Recall: 0.58–0.97
[5]	2021	NSL-KDD	Spam detection	LR, MLP, SVM, DBN and SNSDAE	Precision = 100.00 Recall = 85.42 F1-score = 87.37

(Continued)

Table 1 (continued)

Ref.	Year	Data	Domain	Algorithms	Result
[13]	2021	KDDCUP99, CICIDS2017 and AAGM	Network flows	SVM	Precision = 99.65 Recall = 42.78 F1-score = 59.86
[14]	2021	Dataset of 15000 legitimate and 15000 phishing websites	Websites	RF, DT, SVM, Gaussian Naïve Bayes (GNB), Stochastic Gradient Descent (SGD) and KNN	Precision = 95.45 Recall = 96.25 F1 score = 95.80 Accuracy = 94.28
[15]	2021	CIDDS-01 dataset	Network flows	Long Short-Term Memory Networks (LSTM)-Convolutional Neural Networks (CNN) and Gated Recurrent Unit (LSTM-GRU)	Precision = 99.85 Recall = 99.85 F1-score = 99.91 Accuracy = 99.92
[16]	2021	KDD99, NSL-KDD, CICIDS2017 and Bot-IoT dataset	Network flows	Enhanced Geometric Synthetic Minority Oversampling Technique (EG-SMOTE)	F1-score = 99.99
[17]	2021	NSL-KDD	Network flows	KNN, Linear SVM, DT, RF and SGD	Accuracy = 82.74
[9]	2020	KDD CUP 99, Spambase, Twitter dataset, Enron, NSL-KDD, DARPA and malware datasets	Spam detection	SVM, DT and DBN	Precision = 98.00 Recall = 98.02 Accuracy = 97.43
[18]	2020	Twitter streaming data	vulnerability	CNN, Recurrent Neural Network (RNN), LSTM and GRU	Precision = 90.30 Recall = 89.30 F1-score = 89.30 Accuracy = 89.30
[19]	2019	UNB-CIC Tor Network Traffic datasets	Network flows	KNN, C4.5, Latent Dirichlet Allocation (LDA), MLP and SVM	Accuracy = 100
[11]	2018	DGA datasets	Network flows	RF and Feedforward Deep Neural Network	Precision = 87.27 Recall = 73.60 F-score = 79.85
[20]	2018	r2 dataset	Activity records	Regression, Neural Network, and SVM	Accuracy = 99.71
[21]	2017	Threats dataset	Malware traffic	RF, Partial Decision Tree Algorithm (PART), Repeated Incremental Pruning to Produce Error Reduction (RIPPER), Ensemble Method	Accuracy = 98.20

In this work, an empirical investigation is conducted to detect cybersecurity threats in Twitter posts by using a multiclass classification machine learning approach to enrich the research in this field and explore the performance of using machine learning in the detection of cybersecurity threats.

3 Proposed Framework

The pipeline of the framework for analyzing data to detect cybersecurity threats is illustrated in Fig. 1. The details for each stage in the proposed pipeline are explained in the next subsections.

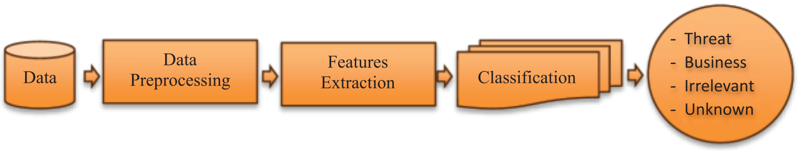


Figure 1: The phases of the study framework pipeline

3.1 Data Acquisition

The data used in this study is secondary data published by [22]. The data is collected from Twitter related to security issues. The dataset has 21368 data rows in four columns (id, text, target, type). The target (label) includes four classes, which are threat, business, irrelevant and unknown. Table 2 illustrates the statistics of each class.

Table 2: Tweets class statistical distribution

Target	No. of tweets
Threat	8280
Business	2331
Irrelevant	6598
Unknown	4159

3.2 Data Preprocessing

Twitter data usually contains many undesired and irrelevant extra data that is considered noise which consumes more processing time and affects the analysis results [23]. The data preprocessing stage removes these unwanted data and cleans it to be ready for further stages of analysis. To achieve this mission, some tasks should be accomplished. These tasks include Data cleaning, Tokenization, Removing stop words, and Stemming [24].

3.2.1 Data Cleaning and Filtering

In this task, the raw tweets passed through some steps to clean and remove unwanted and unrelated symbols and noisy non-word characters from the tweets such as the punctuations, special characters, and non-English words and symbols. These non-word characters affect the analysis process, consume a lot of processing resources, and affect the result accuracy [24]. The output of this task is cleaned tweets that consist of cleaned words without any non-words characters. In addition, the filtering process is taking place to remove the duplicated and empty tweets that consume the processing resources without any benefits [25].

3.2.2 Tokenization

This task includes the process of breaking down the tweets into their basic units (tokens or words) by using the white space as a separator. Tokenization is a very important task where future stages of analysis id depend completely on its output [26]. There are many tokenizing techniques used in literature, word tokenizer technique produced by Natural Language Tool Kit (NLTK) [27] is one of the famous tokenizers and it is used in this study. The expected output of this task is a list of lists (a list of tweets and each tweet is a list of words).

3.2.3 Removing Stopwords

Stopwords are the words that are used to connect sentences and it does not have meaning like determiners, prepositions, and coordinating conjunctions. Removing such words will save more of the analysis resources to focus on the important words instead. Each language has its stopwords. In this study, the stopwords of the English language are used, and the study uses the stopwords list that is published by NLTK [28].

3.2.4 Stemming

The stemming is a process to remove the suffixes and prefixes from the word to return it to the basic original form (stem). This process will help to find the true frequency of one word in one tweet or the whole dataset. In this study, snowball stemmer [29] is used to transform the tweet's words into its basic stem.

3.3 Feature Extraction

The data produced in the preprocessing stage is still textual and it needs to be prepared for the machine learning processes. The machine is a digital platform that only understands numeric data [30]. Machine learning algorithms accept numeric data and these data are called features. Features can be extracted from the textual data to be analyzed by the algorithm for learning.

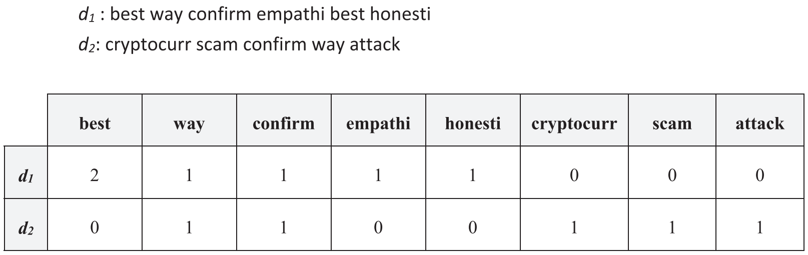
In the feature extraction stage, valuable information will be extracted from the cleaned data as features. Many kinds of features can be extracted. This work utilizes two famous models for data representation used widely in the literature. The first one is called Bag of Words (BOW) [31] and the other model is TF-IDF [32], these models extract numeric features to be used in the machine learning algorithm.

The main idea behind the model BOW is the representation of all tweets in the dataset as a matrix where each row in that matrix represents one tweet while each column in that matrix represents one word of the whole tweets. After that, the frequency of appearance of each word (column) is calculated for each tweet (row) and held in the corresponding cell. Fig. 2 illustrates the BOW representation for two different tweets selected randomly from the dataset.

From the above sample, d1 and d2 represent two different tweets and are considered as rows in the BOW matrix, while each word in those tweets is extracted and represented as a column in the BOW matrix. The frequency of word appearance in each tweet is held in the corresponding cell. For example, the word "best" appears in tweet d1 two times and the word "way" appears in both tweets d1 and d2 one time for each. The word frequency in each tweet is given by Eq. (1), which is called the Term Frequency (TF) equation as follows:

$$tf(\mathrm{w}, \mathrm{t}) = f\mathrm{w,t} \tag{1}$$

where w represents the tweet words, t is the tweet, and fw,t is the frequency of the word w occurring in tweet t [33].



d_1 : best way confirm empathi best honesti

d_2: cryptocurr scam confirm way attack

	best	way	confirm	empathi	honesti	cryptocurr	scam	attack
d_1	2	1	1	1	1	0	0	0
d_2	0	1	1	0	0	1	1	1

Figure 2: The representation of two tweets using the BOW model

However, this way of representation has a drawback, suppose there is an irrelative word that has a high score of word frequency because it appears in one or two tweets with high frequency. Then TF representation will give that word high priority while it is not important for the general topic of the dataset. For that reason, the learning will be low in accuracy. To overcome this issue, the frequency of words is integrated with another concept to give a high score for the most important word in the dataset. This concept is IDF, which calculates the inverse fraction of all words in the dataset to give a correct explanation of the importance of each word related to the whole dataset. Eq. (2) calculates the IDF for the given word:

$$\mathrm{idf}\,(\mathrm{w}, \mathrm{T}) = log10\frac{|\mathrm{T}|}{|[\mathrm{t} \in \mathrm{T}: \mathrm{w} \in \mathrm{t}]|} \tag{2}$$

where t is a tweet, T represents all tweets in the dataset, |T| is the number of tweets, and $|[\mathrm{t} \in \mathrm{T} : \mathrm{w} \in \mathrm{t}]|$ is the number of words w in all tweets [32]. The idf(w, T) is integrated with TF to get the frequency of importance weight for each word in the whole dataset [34]. Eq. (3) combines the two concepts.

$$\mathrm{tf\_idf}(\mathrm{w}, \mathrm{t}, \mathrm{T}) = tf(\mathrm{w}, \mathrm{t}).idf(\mathrm{w}, \mathrm{T}) \tag{3}$$

3.4 Multiclass Classification

The results of the previous stage of feature extraction is a numeric data in form of TF-IDF values. This kind of data is suitable to be used to train and learn the machine. As mentioned earlier, the dataset that is used in this work contains tweets belonging to four different classes (threat, business, irrelevant and unknown). This arise a popular machine learning problem called (Multiclass classification) where the classification is usually applied to data with only two classes (binary classification) [35]. For multiclass classification or multinomial classification, there are two main strategies followed to solve this problem: One-*vs.*-rest strategy which trains one classifier per one class where the samples of that class are considered as positive samples on the other side all the rest or other samples are considered as negative. The second strategy is one-*vs.*-one which trains one classifier for each pair of classes [36]. Eq. (4) gives the number of class pairs to be trained:

$$\mathrm{P} = (\mathrm{N}^{*}(\mathrm{N} - 1))/2 \tag{4}$$

where P is the total number of class pairs and N is equal to the total number of classes [37].

In this work, one-*vs.*-rest strategy is followed to classify the data using six different classification algorithms: SVM, Naïve Bayes (NB), RF, LR, DT, and KNN. One-*vs.*-rest strategy is used widely in literature and it shows good results with classification algorithms such LR [38], SVM [37], and RF [39]. Table 3 illustrates the parameter setting for each algorithm.

Table 3: Machine learning algorithm parameters settings

Algorithm	Parameter	Value
SVM	kernel	linear
	C parameter	1
RF	n_estimators	1000
	random_state	0
LR	Default parameters	Default values
DT	max_depth	2
NB	Default parameters	Default values
KNN	n_neighbors	7

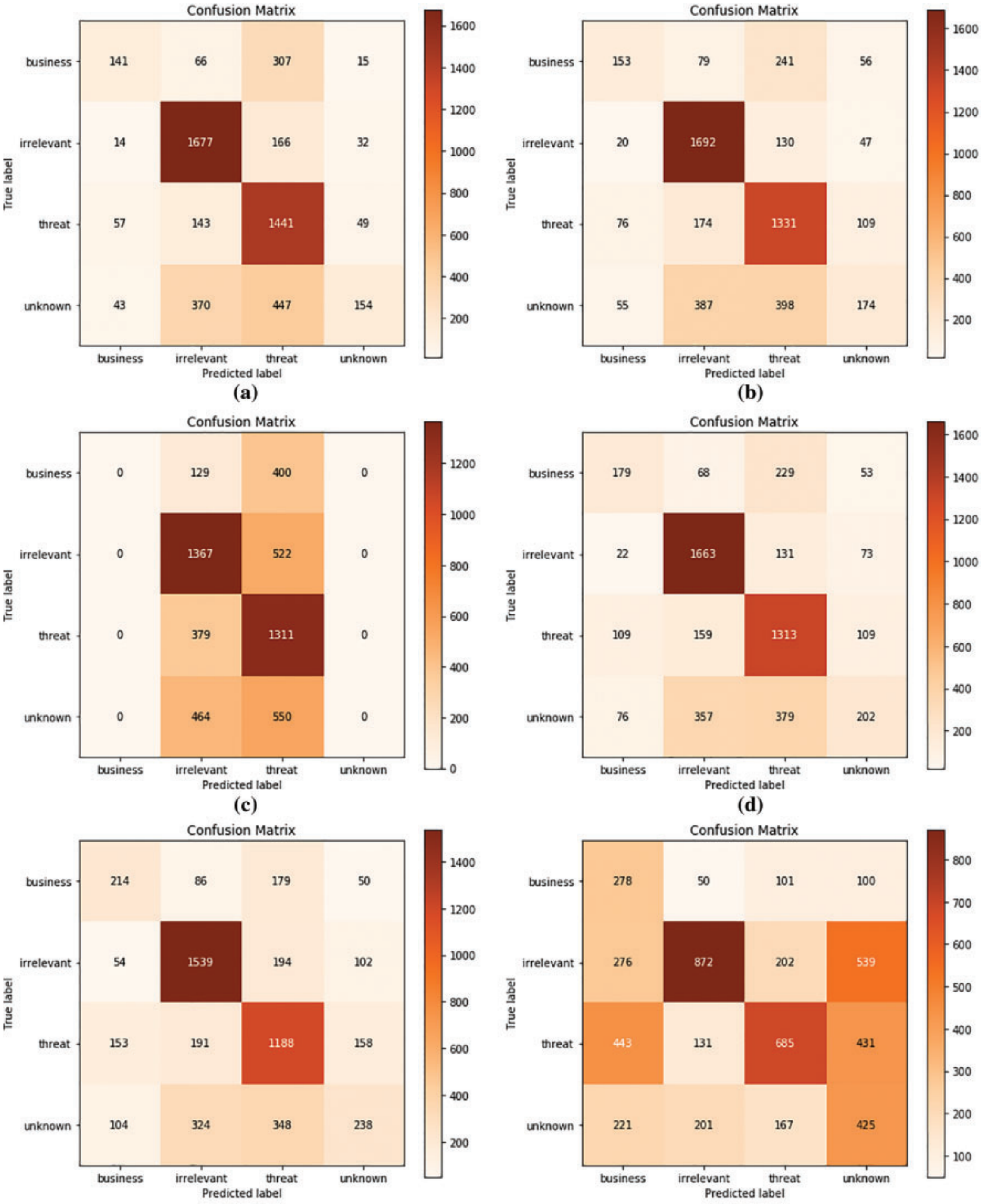
4 Experiment Results and Discussion

In this study, all the experiments are conducted using Python programming language and run using the Anaconda environment and its integrated editor of Jupiter Notebook. Windows operating system PC was used to run the all experiments with an Intel processor, Core i7-4770, speed of 3.40 GHz, and 8 GB RAM.

The experiments flow to start with preparing the data and preprocessing with a total number of 21368 tweets. Preprocessing stage outcomes out with 15519 tweets after filtering and removing the empty and duplicated tweets. Feature extraction is followed using the TF-IDF technique to extract the importance frequency of each feature in the dataset. A total number of 17805 features is obtained from the feature extraction process, which became in ready format to be fed into the machine learning algorithms. Before proceeding to the classification process, the whole dataset is divided into two parts: the training dataset and the testing dataset with 66% and 34% division, respectively. Each algorithm is trained using the training data then the prediction of the algorithm is tested using the testing data.

The result is evaluated by calculating the most popular evaluation metrics: precision, recall, F1-score, and accuracy. The confusion matrix is produced for each algorithm and visualized using clear readable diagrams. Fig. 3 shows the confusion matrix and Table 4 shows the classification report for each algorithm.

From the above evaluation results, the classification algorithms have achieved different scores of accuracy. Notably, the RF algorithm achieved the best results with the highest accuracy (0.67) among the other algorithms, while the DT achieved the lowest results and NaÏve Bayes achieved the lowest accuracy (0.44). In general, the results are quite low and there are some reasons behind this low accuracy. The first reason is related to the representation of unstructured textual data. The tweet text is short, unstructured, and has many issues that cause the low accuracy of the machine learning tasks [40]. One of the issues in analyzing unstructured text is the high dimensionality of the resulting feature vectors [41]. The second reason is related to the data representation technique TF-IDF, which resulted in a very sparse feature vector matrix. Most of the values of these sparse vectors are zeros and this affects negatively the accuracy of the analysis results [41,42].



(a) (b) (c) (d) (e) (f)

Figure 3: The confusion matrix of the classification results: (a) RF result; (b) LR result; (c) DT result; (d) SVM result; (e) KNN result; (f) NB result

Table 4: The classification results report

Algorithm	Class	Precision	Recall	F1-score	Accuracy
SVM	Business	0.46	0.34	0.39	0.66
	Irrelevant	0.74	0.88	0.80	
	Threat	0.64	0.78	0.70	
	Unknown	0.46	0.20	0.28	
RF	Business	0.55	0.27	0.36	0.67
	Irrelevant	0.74	0.89	0.81	
	Threat	0.61	0.85	0.71	
	Unknown	0.62	0.15	0.24	
LR	Business	0.50	0.29	0.37	0.65
	Irrelevant	0.73	0.90	0.80	
	Threat	0.63	0.79	0.70	
	Unknown	0.45	0.17	0.25	
DT	Business	0.00	0.00	0.00	0.52
	Irrelevant	0.58	0.72	0.65	
	Threat	0.47	0.78	0.59	
	Unknown	0.00	0.00	0.00	
NB	Business	0.23	0.53	0.32	0.44
	Irrelevant	0.70	0.46	0.55	
	Threat	0.59	0.41	0.48	
	Unknown	0.28	0.42	0.34	
KNN	Business	0.41	0.40	0.41	0.62
	Irrelevant	0.72	0.81	0.76	
	Threat	0.62	0.70	0.66	
	Unknown	0.43	0.23	0.30	

The classification result of the proposed study can be compared with the other methods of threat detection. Table 5 shows a brief comparison between some studies from the literature with the proposed study in the form of precision, recall, and F1-score. The chosen comparsion studies were selected based on the used data and the multiclass classification approach. From the comparison table, it is clear that the proposed study has achieved the best recall and F1 using the RF classification method among the other studies. The high F1-score indicates that the proposed method achieved good balanced predictions of true positive and true negative. As observed from the Table 5, the study [13] achieved high precision but low recall and F1-score, that is meaning the model can predict true positive accurately but it gives low accuracy with true negative prediction. This unbalanced results of precision and recall cause the low results of F1-score.

Table 5: Results comparison

Study	Method	Precision	Recall	F1
[43]	k-means (k = 10)	–	–	68.62
[44]	Ensemble-based approach	70.84	75.65	73.99
[13]	Threat detection by moving boundaries around normal samples (THEODORA)	**99.65**	42.78	59.86
[11]	RF classification	73.6	87.27	79.85
Proposed study	RF classification	74	**89**	**81**

5 Conclusion and Future Works

The detection of cyber threats is an important process and it is not an easy task. There are many works in the literature that used different approaches to overcome this task. The integrity between the several sciences opens a new analysis era to use other approaches in one field to solve issues in other fields. Machine learning is one of the promising approaches that achieve good results in several aspects of other fields. Resolving cybersecurity issues with machine learning still needs more investigations. In this study, a machine learning supervised multiclass classification approach is used to classify data from Twitter into several classes to detect cyber threads. By using several machine learning algorithms, this study achieved a detection accuracy of (67%). For future work, more investigation must be conducted to discover the other sides of cyber threat detection by using machine learning. In addition, the study suggests using different data representations for the feature extraction process such as Word2Vec to refine and obtain a high score of accuracy. Furthermore, the study suggests adding a new process to the framework pipeline which is the feature selection process to select the optimal set of features and ignore the unimportant features that affect the accuracy of the detection.

Acknowledgement: The authors extend their appreciation to the Deputyship for Research & Innovation, Ministry of Education in Saudi Arabia for funding this research work through Project Number MoE-IF-UJ-22-04100409-5.

Funding Statement: This research was funded by Deputyship for Research & Innovation, Ministry of Education in Saudi Arabia, Project Number MoE-IF-UJ-22-04100409-5.

Author Contributions: The authors confirm contribution to the paper as follows: study conception and design: A. Hussein, A. A. Almazroi; data collection: A. Hussein; analysis and interpretation of results: A. Hussein; draft manuscript preparation: A. Hussein, A. A. Almazroi. All authors reviewed the results and approved the final version of the manuscript.

Availability of Data and Materials: The data that support the findings of this study are openly available in “github” at http://doi.org/10.1109/BigData.2018.8622506, reference number [22].

Conflicts of Interest: The authors declare that they have no conflicts of interest to report regarding the present study.

References

[1] E. Sala, G. Cerati and A. Gaia, “Are social media users more satisfied with their life than non-users? A study on older Italians,” *Ageing & Society*, vol. 43, pp. 76–88, 2023.

[2] Z. Saeed, M. Masood and M. U. Khan, “A review: Cybersecurity challenges and their solutions in connected and autonomous vehicles (CAVs),” *JAREE (Journal on Advanced Research in Electrical Engineering)*, vol. 7, pp. 44–51, 2023.

[3] R. R. Karuniawan, S. Santoso, M. Al Fikri and M. Argadilah, “Learning cyber security and machine engineering at the university,” *Blockchain Frontier Technology*, vol. 3, pp. 89–94, 2023.

[4] V. O. Kayhan, M. Agrawal and S. Shivendu, “Cyber threat detection: Unsupervised hunting of anomalous commands (UHAC),” *Decision Support Systems*, vol. 168, pp. 113928, 2023.

[5] A. A. Almazroi, L. Abualigah, M. A. Alqarni, E. H. Houssein, A. Q. M. AlHamad *et al.,* “Class diagram generation from text requirements: An application of natural language processing,” in *Deep Learning Approaches for Spoken and Natural Language Processing*, 1st ed., Switzerland: Springer Cham, pp. 55–79, 2021.

[6] O. Alshaikh, S. Parkinson and S. Khan, “On the variability in the application and measurement of supervised machine learning in cyber security,” in *Int. Conf. on Ubiquitous Security*, pp. 545–555, 2023.

[7] S. Silvestri, S. Islam, S. Papastergiou, C. Tzagkarakis and M. Ciampi, “A machine learning approach for the NLP-based analysis of cyber threats and vulnerabilities of the healthcare ecosystem,” *Sensors*, vol. 23, no. 2, pp. 651, 2023.

[8] F. Alrowais, S. Althahabi, S. S. Alotaibi, A. Mohamed, M. A. Hamza *et al.,* “Automated machine learning enabled cybersecurity threat detection in Internet of Things environment,” *Computer Systems Science and Engineering*, vol. 45, no. 1, pp. 687–700, 2023.

[9] K. Shaukat, S. Luo, S. Chen and D. Liu, “Cyber threat detection using machine learning techniques: A performance evaluation perspective,” in *2020 Int. Conf. on Cyber Warfare and Security (ICCWS)*, pp. 1–6, Islamabad, Pakistan, 2020.

[10] Q. Ke, “Research on threat detection in cyber security based on machine learning,” *Journal of Physics: Conference Series*, vol. 2113, no. 1, pp. 012074, 2021.

[11] G. Apruzzese, M. Colajanni, L. Ferretti, A. Guido and M. Marchetti, “On the effectiveness of machine and deep learning for cyber security,” in *2018 10th Int. Conf. on Cyber Conflict (CyCon)*, Tallinn, Estonia, pp. 371–390, 2018.

[12] D. Nandakumar, R. Schiller, C. Redino, K. Choi, A. Rahman *et al.,* “Zero day threat detection using metric learning autoencoders,” arXiv preprint arXiv:2211.00441, 2022.

[13] G. Andresini, A. Appice, F. Paolo Caforio and D. Malerba, “Improving cyber-threat detection by moving the boundary around the normal samples,” in *Machine Intelligence and Big Data Analytics for Cybersecurity Applications*, 1st ed., Switzerland: Springer Cham, Chapter 5, pp. 105–127, 2021.

[14] E. Kocyigit, M. Korkmaz, O. K. Sahingoz and B. Diri, “Real-time content-based cyber threat detection with machine learning,” in *Int. Conf. on Intelligent Systems Design and Applications*, Warsaw, Poland, pp. 1394–1403, 2021.

[15] I. Ullah, B. Raza, S. Ali, I. A. Abbasi, S. Baseer *et al.,* “Software defined network enabled fog-to-things hybrid deep learning driven cyber threat detection system,” *Security and Communication Networks*, vol. 2021, pp. 1–15, 2021.

[16] V. Christopher, T. Aathman, K. Mahendrakumaran, R. Nawaratne, D. de Silva *et al.,* “Minority resampling boosted unsupervised learning with hyperdimensional computing for threat detection at the edge of Internet of things,” *IEEE Access*, vol. 9, pp. 126646–126657, 2021.

[17] N. Peppes, E. Daskalakis, T. Alexakis, E. Adamopoulou and K. Demestichas, “Performance of machine learning-based multi-model voting ensemble methods for network threat detection in Agriculture 4.0,” *Sensors*, vol. 21, no. 22, pp. 7475, 2021.

[18] K. Simran, P. Balakrishna, R. Vinayakumar and K. Soman, “Deep learning approach for enhanced cyber threat indicators in Twitter stream,” in *Int. Symp. on Security in Computing and Communication*, pp. 135–145, 2020.

[19] P. Sornsuwit and S. Jaiyen, “A new hybrid machine learning for cybersecurity threat detection based on adaptive boosting,” *Applied Artificial Intelligence*, vol. 33, pp. 462–482, 2019.

[20] M. S. Raval, R. Gandhi and S. Chaudhary, “Insider threat detection: Machine learning way,” in *Versatile Cybersecurity*, 1st ed., Switzerland: Springer Cham, pp. 19–53, 2018.

[21] S. Kumar, A. Viinikainen and T. Hamalainen, “Evaluation of ensemble machine learning methods in mobile threat detection,” in *2017 12th Int. Conf. for Internet Technology and Secured Transactions (ICITST)*, New York, NY, USA, pp. 261–268, 2017.

[22] V. Behzadan, C. Aguirre, A. Bose and W. Hsu, “Corpus and deep learning classifier for collection of cyber threat indicators in twitter stream,” in *2018 IEEE Int. Conf. on Big Data (Big Data)*, Seattle, WA, USA, pp. 5002–5007, 2018.

[23] O. Abiola, A. Abayomi-Alli, O. A. Tale, S. Misra and O. Abayomi-Alli, “Sentiment analysis of COVID-19 tweets from selected hashtags in Nigeria using VADER and Text Blob analyser,” *Journal of Electrical Systems and Information Technology*, vol. 10, pp. 1–20, 2023.

[24] T. Singh and M. Kumari, “Role of text pre-processing in twitter sentiment analysis,” *Procedia Computer Science*, vol. 89, pp. 549–554, 2016.

[25] S. Pradha, M. N. Halgamuge and N. T. Q. Vinh, “Effective text data preprocessing technique for sentiment analysis in social media data,” in *2019 11th Int. Conf. on Knowledge and Systems Engineering (KSE)*, Da Nang City, Vietnam, pp. 1–8, 2018.

[26] V. Wisdom and R. Gupta, *An Introduction to Twitter Data Analysis in Python*. Artigence Inc., India, 2016.

[27] S. Bird, “NLTK: The natural language toolkit,” in *Proc. of the COLING/ACL, 2006 Interactive Presentation Sessions*, Sydney, Australia, pp. 69–72, 2006.

[28] NLTK, “NLTK stopwords list,” 2023. [Online]. Available: https://www.nltk.org/search.html?q=stopwords

[29] M. F. Porter, “Snowball: A language for stemming algorithms,” 2001. Available: http://snowball.tartarus.org/

[30] M. Botlagunta, M. D. Botlagunta, M. B. Myneni, D. Lakshmi, A. Nayyar *et al.,* “Classification and diagnostic prediction of breast cancer metastasis on clinical data using machine learning algorithms,” *Scientific Reports*, vol. 13, pp. 485, 2023.

[31] Y. Zhang, R. Jin and Z. H. Zhou, “Understanding bag-of-words model: A statistical framework,” *International Journal of Machine Learning and Cybernetics*, vol. 1, pp. 43–52, 2010.

[32] S. Robertson, “Understanding inverse document frequency: On theoretical arguments for IDF,” *Journal of Documentation*, vol. 60, no. 5, pp. 503–520, 2004.

[33] H. P. Luhn, “A statistical approach to mechanized encoding and searching of literary information,” *IBM Journal of Research and Development*, vol. 1, pp. 309–317, 1957.

[34] K. Sparck Jones, “A statistical interpretation of term specificity and its application in retrieval,” *Journal of Documentation*, vol. 28, pp. 11–21, 1972.

[35] A. Robles-Guerrero, T. Saucedo-Anaya, E. González-Ramírez and J. I. de la Rosa-Vargas, “Analysis of a multiclass classification problem by lasso logistic regression and singular value decomposition to identify sound patterns in queenless bee colonies,” *Computers and Electronics in Agriculture*, vol. 159, pp. 69–74, 2019.

[36] J. A. Rohwer and C. T. Abdullah, “One-vs-one multiclass least squares support vector machines for direction of arrival estimation,” *The Applied Computational Electromagnetics Society Journal (ACES)*, vol. 18, no. 2, pp. 34–45, 2003.

[37] J. H. Hong and S. B. Cho, “A probabilistic multi-class strategy of one-vs.-rest support vector machines for cancer classification,” *Neurocomputing*, vol. 71, pp. 3275–3281, 2008.

[38] Y. Alhessi and R. Wicentowski, “SWATAC: A sentiment analyzer using one-vs-rest logistic regression,” in *Proc. of the 9th Int. Workshop on Semantic Evaluation (SemEval 2015)*, Denver, Colorado, pp. 636–639, 2015.

[39] J. Ramírez, J. Górriz, A. Ortiz, F. Martínez-Murcia, F. Segovia *et al.,* “Ensemble of random forests One vs. Rest classifiers for MCI and AD prediction using ANOVA cortical and subcortical feature selection and partial least squares,” *Journal of Neuroscience Methods*, vol. 302, pp. 47–57, 2018.

[40] R. Qasim, W. H. Bangyal, M. A. Alqarni and A. Ali Almazroi, "A fine-tuned BERT-based transfer learning approach for text classification," *Journal of Healthcare Engineering*, vol. 2022, pp. 3498123, 2022.

[41] A. Hussein, F. K. Ahmad and S. S. Kamaruddin, "Cluster analysis on COVID-19 outbreak sentiments from twitter data using K-means algorithm," *Journal of System and Management Sciences*, vol. 11, pp. 167–189, 2021.

[42] M. Chen, K. Q. Weinberger and F. Sha, "An alternative text representation to TF-IDF and Bag-of-Words," arXiv preprint arXiv:1301.6770, 2013.

[43] K. Alperin, E. Joback, L. Shing and G. Elkin, "A framework for unsupervised classificiation and data mining of tweets about cyber vulnerabilities," arXiv preprint arXiv:2104.11695, 2021.

[44] A. Mehmood, M. S. Farooq, A. Naseem, F. Rustam, M. G. Villar *et al.,* "Threatening URDU language detection from tweets using machine learning," *Applied Sciences*, vol. 12, no. 20, pp. 10342, 2022.