UDC 004.85:004.056
DOI: 10.15587/1729-4061.2024.306522

DETECTION AND CLASSIFICATION OF THREATS AND VULNERABILITIES ON HACKER FORUMS BASED ON MACHINE LEARNING

Saken Mambetov
PhD Student
Department of Information Systems
Al-Farabi Kazakh National University
al-Farabi ave., 71, Almaty, Republic of Kazakhstan, 050040

Yenlik Begimbayeva
PhD, Head of Department
Department of Cybersecurity
AUPET named after Gumarbek Daukeyev
Baytursynuli str., 126/1, Almaty, Republic of Kazakhstan, 050013
Satbayev University
Satbayev str., 22, Almaty, Republic of Kazakhstan, 050013

Oleksandr Gurko
Corresponding author
Doctor of Technical Sciences, Head of Department
Department of Automation and Computer-Aided Technologies*
E-mail: gurko@khadi.kharkov.ua

Hanna Doroshenko
Doctor of Economic Sciences, Professor, Head of Department**

Serik Joldasbayev
Master of Science
Department of Computer Engineering
International IT University
Manasa str., 34/1, Almaty, Republic of Kazakhstan, 050040

Olena Fridman
Associate Professor**

Bakytzhan Kulambayev
Candidate in Technical Sciences
Department of Radio Engineering, Electronics and Telecommunications
Turan University
Satpaeva str., 16a, Almaty, Republic of Kazakhstan, 050013

Vitalina Babenko
Doctor of Economic Sciences, PhD, Professor, Head of Department
Department of Computer Systems*
Department of Law, Management & Economics
Daugavpils University
Vienības str., 13, Daugavpils, Latvia, 5401

Ihor Ilhe
Associate Professor
Department of Automation and Computer-Aided Technologies*

Serhii Neronov
Senior Lecturer
Department of Computer Systems*
*Kharkiv National Automobile and Highway University
Yaroslava Mudroho str., 25, Kharkiv, Ukraine, 61002
**Department of Economics and Management
V. N. Karazin Kharkiv National University
Svobody sq., 4, Kharkiv, Ukraine, 61022

The object of this study is the process of detecting threats and vulnerabilities in hacker forums, which are a well-known source of potential dangers for Internet users. However, the problem of analyzing and classifying data from these forums is its complexity due to such features of the participants' language as specific slang, jargon, etc., which requires the use of modern tools of their processing. This paper explores the application of machine learning to devise an effective method for analyzing sentiment and trends in hacker forums to identify potential threats and vulnerabilities in cyberspace. All necessary stages of the process of detecting threats and vulnerabilities have been developed, ranging from data collection and preprocessing to the training of a model that is capable of processing "raw" unstructured data from hacker forums. The implementation of six popular machine learning algorithms, namely k Nearest Neighbors (kNN), Random Forest, Naive Bayes, Logistic Regression, Support Vector Machines (SVM), and Decision Tree algorithms have been studied with a view to determining their efficiency of threat and vulnerability detection and classification. The experiments have been conducted on real data (150,000 messengers). It has been determined that the Random Forest algorithm coped with the task the best (accuracy=0.89, recall=0.84, precision=0.91, F1-score=0.87 and ROC-AUC=0.89). The proposed tool based on machine learning not only collects data that poses a potential threat but also processes and classifies it according to the specified keywords. This allows detecting threats and vulnerabilities at a high speed. The results of the study make it possible to identify potential trends in threats and vulnerabilities. This will contribute to the improvement of cybersecurity systems and ensure more reliable protection of information resources

Keywords: cybersecurity, hacker forum, threats identification, data classification, machine learning

Received date 05.04.2024
Accepted date 12.06.2024
Published date 30.06.2024

How to Cite: *Mambetov, S., Begimbayeva, Ye., Gurko, O., Doroshenko, H., Joldasbayev, S., Fridman, O., Kulambayev, B., Babenko, V., Ilhe, I., Neronov, S. (2024). Detection and classification of threats and vulnerabilities on hacker forums based on machine learning. Eastern-European Journal of Enterprise Technologies, 3 (9 (129)), 16–27. https://doi.org/10.15587/1729-4061.2024.306522*

1. Introduction

Given the constant expansion of the digital domain and the increase in the amount of information circulating in the virtual space, issues of cyber security are becoming an inseparable part of modern reality. In a world where technological advances replace each other at an incredible speed, threats to cyber security and vulnerabilities of information systems are taking on new forms and dimensions. Maintaining the integrity, confidentiality, and availability of data becomes an integral task for organizations, individual users, and society as a whole [1]. Gathering, analyzing, and interpreting threat and vulnerability data are critical steps in ensuring effective cybersecurity. One of the important sources for this is hacker forums [2], which have a complex social structure and their own hierarchy [3–7]. On such platforms, hackers and cybercriminals exchange information about attack methods, new vulnerabilities, and other related topics. This data is an important resource for analyzing and preventing potential threats. Hacker forums are also good for analyzing the consequences of cyberattacks that have already been carried out, which also makes it possible to prevent further attacks by attackers in advance.

Threat and vulnerability analysis faces a number of problems and challenges. The first of them is the volume and complexity of data because a huge amount of information is generated in the Internet space, including messages of various styles, topics and levels of technical complexity [8]. Processing and analysis of this textual data requires natural language processing tools and algorithms that can distinguish important signals from the noise [9]. The second is the classification and clustering of data, which can contain information about a variety of threats, from infrastructure attacks to social engineering. Effective classification and clustering of data will help identify different types of threats and allow for more accurate protection strategies to be developed [10]. The third pressing problem is the lack of labeled data, which in the field of cyber security limits the ability to train appropriate machine learning algorithms. Building effective models requires reliable, labeled datasets, which can be difficult to achieve in a hacker forum environment. The fourth in this series is anonymity and information substitution when attackers can intentionally distort information, making it difficult to analyze it. Anonymity of participants can also complicate the process of identification and classification of threat sources. And the biggest problem is the rapidity of information, that is, threats and vulnerabilities evolve quickly. All these challenges create the need to devise new methods and tools for data analysis from hacker forums. Since computer methods are a powerful device for automatic data processing, it is appropriate to use them to analyze the content of hacker forums. The above justifies the relevance of this study, which proposes an effective machine learning-based tool for reducing the risks and vulnerabilities posed by the content of hacker forums.

2. Literature review and problem statement

In [11], an analysis of the social network was carried out in combination with the technique of textual analysis of the data of hacker forums in the Darknet. A special crawler has been designed for this purpose. The received content was analyzed by statistical methods, which made it possible to identify extremist sentiments. However, the main emphasis in the work is on the collection of data, and not on their analysis and prediction of potential threats.

In [12, 13], Linguistic Inquiry and Word Count(LIWC) software was used to determine the linguistic characteristics of each forum/subreddit. Hacking-related texts from websites were collected and a thematic analysis of a portion of the text from each source was conducted. The results of the LIWC analysis showed that there are differences between these forums and sub-resources on several psychologically significant factors, including the extent to which users use language that indicates their honesty, confidence, analytical level, and emotionality. However, those works are aimed at understanding the process of creation and dynamics of hacker communities and do not consider the issue of analysis and forecasting of threats and vulnerabilities.

In [14], the authors consider the application of the English Event Analysis of Systematic Teamwork (EAST) methodology to study the interaction of participants of an illegal trading platform on the dark Internet. Special attention is paid to the investigation of the procedure of the initial registration of a participant on an illegal trading platform, when a potential applicant must demonstrate to the dark web market provider that he can be trusted. The application of the methodology of sociotechnical systems in a little-studied but popular environment is proposed. However, the work examines the relationships between participants of illegal platforms in the Darknet and does not consider the degree of danger of the content.

Paper [15] proposed a Darknet research concept called DICE-E. Four steps for conducting darknet forum research are outlined: identifying data sources, data collection strategies, data evaluation, and ethical issues related to darknet research. It should be noted that DICE-E is a valuable tool for finding collection of Darknet content, for further detection of threats and vulnerabilities. However, the solution of the specified task is not considered directly in the work.

To facilitate the implementation of effective countermeasures against cybercrime, study [16] proposes an intelligent system called KADetector, which automates the analysis of hacker forums to identify their key users who play an important role in the criminal chain. A structured heterogeneous information network (HIN) is introduced to model rich semantic relationships, and then a meta-paths approach is used to incorporate higher-level semantics to build the relationship between users in a hacker forum. This work also proposes a way (an HIN embedding model called ActorHin2Vec) to reduce the high computational and spatial costs. Work [17] also considers the categorization of users of hacker forums. However, these papers are more focused on identifying key forum participants who pose a real threat to cyber security, rather than content classification.

In work [18], messages from a specific (selective) hacker forum were collected using a special web crawler. Posts were analyzed using a part-of-speech tag that helped identify a list of keywords used to query the data. The SentiStrength software was used to analyze sentiment on the forums. However, it should be noted that the dictionary used in the work does not take into account the peculiarities of the language of cybercriminals, their jargon and other words and expressions inherent in this subculture. This limits the effectiveness of detecting potential threats.

In [19], a specialized web crawler was developed, specifically designed to collect structured content posted on three hacker forums that represented different aspects of the hacker community. The results of the analysis allowed the authors to understand what threats they pose to critical systems in particular and cyber security in general. However, the results were analyzed manually.

In paper [20], systematic identification and automatic collection of data from hacker forums, carding shops, Internet-Relay-Chat and Dark Net Marketplaces was carried out. It collected information from 102 platforms, a total of 43,981,647 records, which is currently the largest collection of data from the hacking community in an academic environment. The resulting data is provided to Cyber Threat Intelligence (CTI) specialists through the AZSecure Hacker Assets Portal. The shortcoming of the study is the lack of a detailed analysis of the collected information and its practical applicability for predicting and preventing cyber threats. The solution to this shortcoming can be deeper research and analysis of the collected data using machine learning methods.

Machine learning has great potential in improving the efficiency and security of the Internet. For example, work [21] reports a study on the possibilities of machine learning during the organization of the interaction of network elements, where the effectiveness of several machine learning algorithms under different scenarios of requests received by the server is analyzed. And in work [22], the authors indicate ways to reduce damage from accidental or intentional incorrect actions of users and administrators when working with external sources of information. Those two papers apply machine learning to improve network performance but do not directly address how machine learning can reduce network threats.

According to [23, 24], the mood of the authors of the texts, their motivation, productivity, etc. can be estimated based on the evaluation of the text data using machine learning, which can also contribute to the early identification of the future event. However, those studies are also not focused on cyber security.

The analysis of Internet forum data is carried out by parsing web pages, which is complicated by the presence of heterogeneous data. Therefore, to highlight the necessary fragments of the text, subject to certain distortions in the form of abbreviated words, local slang, and peculiarities of the language of information carriers, becomes an extremely difficult task in the field of cyber security and cryptography. Another significant limitation in this area is the limited resources of clients who use heterogeneous security systems. If data processing takes too much time, it can negatively affect the performance and availability of data for end users, as has been shown in [25]. But the use of developed scripts and the implementation of machine learning significantly improved the solution to this problem.

Thus, studies [21–25] are examples of successful application of machine learning to solve various tasks in the network. However, they do not address cyber security issues.

Work [26] is important, in which, similarly to [16, 17], the characteristics of users of underground forums were analyzed and their comprehensive assessment was formed. Dirichlet's Latent Placement Model is used to predict users' topic preferences. When analyzing social networks, user influence is obtained using the improved Topic-specific PageRank algorithm, which is based on comprehensive evaluations and thematic preferences. By ranking a user's influence, you can identify key hackers in underground forums. Experiments compare HackerRank (HR) with methods using only content analysis or social network analysis. However, as already noted, despite the fact that the identification of key participants in hacking forums has a significant impact on cyber security, the analysis of the content of the forums is also necessary.

Understanding the functions and characteristics of assets on hacker forums using classification methods and thematic modeling was carried out in [27]. The study provides a deeper understanding of the hacker assets in known forums and organizes them in such a way that they can be reused for learning purposes. However, the work focuses on the analysis of code posted on hacker forums and does not consider the sentiments and intentions of their participants, which would help prevent potential threats from forum content.

Work [28] also confirms the hypothesis about the potential threats of the content of hacker forums. A combination of machine learning methodology and information retrieval methods is used to detect threats. However, the narrow focus of the research limits the possibilities for identifying other types of cyber threats. This is due to the purpose of the work, which is focused on the detection and prevention of financial fraud and fraud in the network.

In [29], a solution is proposed using a hybrid model of machine learning. The model automatically searches for messages on hacker forums, identifies the most relevant cybersecurity messages, and then groups them according to the ratings of the topics discussed by hackers. In the first stage, the Support Vector algorithm (SVM) is used for identification, and in the second stage, the Dirichlet Latent Placement method is used for grouping. The model is tested on data from a real hacker forum to automatically extract information about various threats such as credential leaks, malicious proxies, malware that evades antivirus detection, and more. However, the work does not evaluate the quality of models and compare different algorithms using metrics typical for machine learning tasks, which makes it difficult to determine the effectiveness of the proposed method.

Work [30] is aimed at studying methods of deep learning and natural language processing to detect unwanted and offensive content in popular social networks. The selection of features was carried out on the basis of a fuzzy fully convolutional neural network model. Extraction of selected features and classification were performed using the ensemble architecture of the Bidirectional Long Short-Term Memory (Bi-LSTM) model with a hybrid architecture of Naive Bayes with SVM-based machines. However, work [30] considers only the detection and classification of cyberbullying in social networks, and the issues of preventing related threats are not considered. Another work [31] contains a systematic review of modern strategies, machine learning methods, and technical means for detecting cyberbullying and aggressive personality management in the information space. All steps for detecting cyberbullying in social networks are considered, including data collection, pre-processing, preparation, selection, and feature extraction. In addition, the effectiveness of the application of machine learning methods for the analysis and classification of texts is evaluated. However, the authors only propose a possible architecture of a cyberbullying and online harassment detection model and do not consider its practical implementation.

Although large organizations tend to attract more attention from attackers due to their extensive infrastructure, large amounts of data and potential financial benefits, they can be the target of large and sophisticated cyber-attacks. These can be attacks on infrastructure, theft of intellectual property and large-scale security breaches. However, larger organizations also tend to have more sophisticated defenses and can invest significant resources in cybersecurity. On the other hand, small and medium-sized organizations may be less secure and more vulnerable to certain types of cyber-attacks. Although hacking attempts may be less sophisticated and aimed at avoiding detection, the consequences for such organizations can be severe as they may lose valuable data or face financial problems. For example, in December 2020, attackers carried out a large-scale hack of users of Orion, a network monitoring product from SolarWinds. Among the victims of the attack were leading US federal agencies such as the Department of Justice, the Treasury, the National Security Service, Fortune 500 companies and their clients, the cybersecurity firm FireEye [32]. Study [32] proposed a text mining-based cyber risk assessment and mitigation system that classifies them by experience and calculates the financial consequences for each hacker/attack type combination. The proposed system is a significant help to company managers in making decisions regarding measures to improve cyber security. However, the results reported in the paper refer to the study of only one hacking forum.

Forums often use anti-overflow tools such as authentication, restriction, and obfuscation. Such limitations prevent many researchers from collecting data in real time. Research [33] considers a web crawler with numerous anti-crawling measures for the continuous collection of hacker exploits and their automatic classification using a recurrent neural network with a long short-term memory. Classification of exploits is carried out on the fly according to predefined categories. Interactive visualizations have also been created to allow CTI professionals to explore the collected exploits. The results of this study indicate, among other things, that exploits of systems and networks are distributed much more often than exploits of other types. Although the crawler presented in the work collects only attachments from hacker forums. Analyzing the rest of the content could help identify more threats and vulnerabilities.

Thus, the shortcomings of existing solutions for the application of machine learning for the analysis of hacker forums are their narrow focus on identifying a specific type of threats, as well as the lack of information about the effectiveness of the proposed solutions. This makes it possible to argue for the need for further research on the use of machine learning to analyze data from hacker forums in order to identify and classify potential threats and vulnerabilities.

3. The aim and objectives of the study

The purpose of our study is to improve the security of Internet users' data by developing an effective tool for detecting and classifying threats and vulnerabilities in online hacker forums using machine learning algorithms. This will enable CTI specialists to devise effective risk mitigation strategies in a timely manner.

To achieve the specified goal, the following tasks must be completed:

– to collect data from hacker forums for use in testing algorithms;

– to perform preliminary preparation of the collected data for preparation for marking and effective training;

– to mark up data based on hacker terminology to determine forum user intentions;

– to train models to detect threats and vulnerabilities containing data obtained from hacker Internet forums, using popular machine learning algorithms based on labeled data, and to evaluate the effectiveness of the application of trained models.

4. The study materials and methods

The object of our study is the process of identifying threats and vulnerabilities on hacker forums. The main hypothesis of the study assumes that the application of machine learning methods will increase the effectiveness of detecting and categorizing the content of hacker forums, which may pose a potential threat.

This study is limited to analyzing data only from active English-language hacker forums, the content of which is periodically updated.

Data from hacker forums were collected and pre-processed using a script written in Python using additional libraries. The data was partitioned based on pre-collected and derived word terminology, hacker dictionary slang, and manually assigned weights. Training and testing of the model were carried out with the help of scripts, also written in the Python language.

The threat and vulnerability detection technology proposed in this paper consists of the following (Fig. 1) phases:

1. Data collection.
2. Data preparation.
3. Data marking.
4. Training and testing.

The content of these phases is shown in Fig. 1. In Phase A, data was collected using a parser. In Phase B, incorrect and uninformative characters were removed from the data and feature scaling was performed. During phase C, determination of tonality and classification of data by topics, types of threats and other features were carried out.

In phase D, the training of search models in text data for content that may pose potential threats to Internet users and the evaluation of the effectiveness of the trained models were carried out.

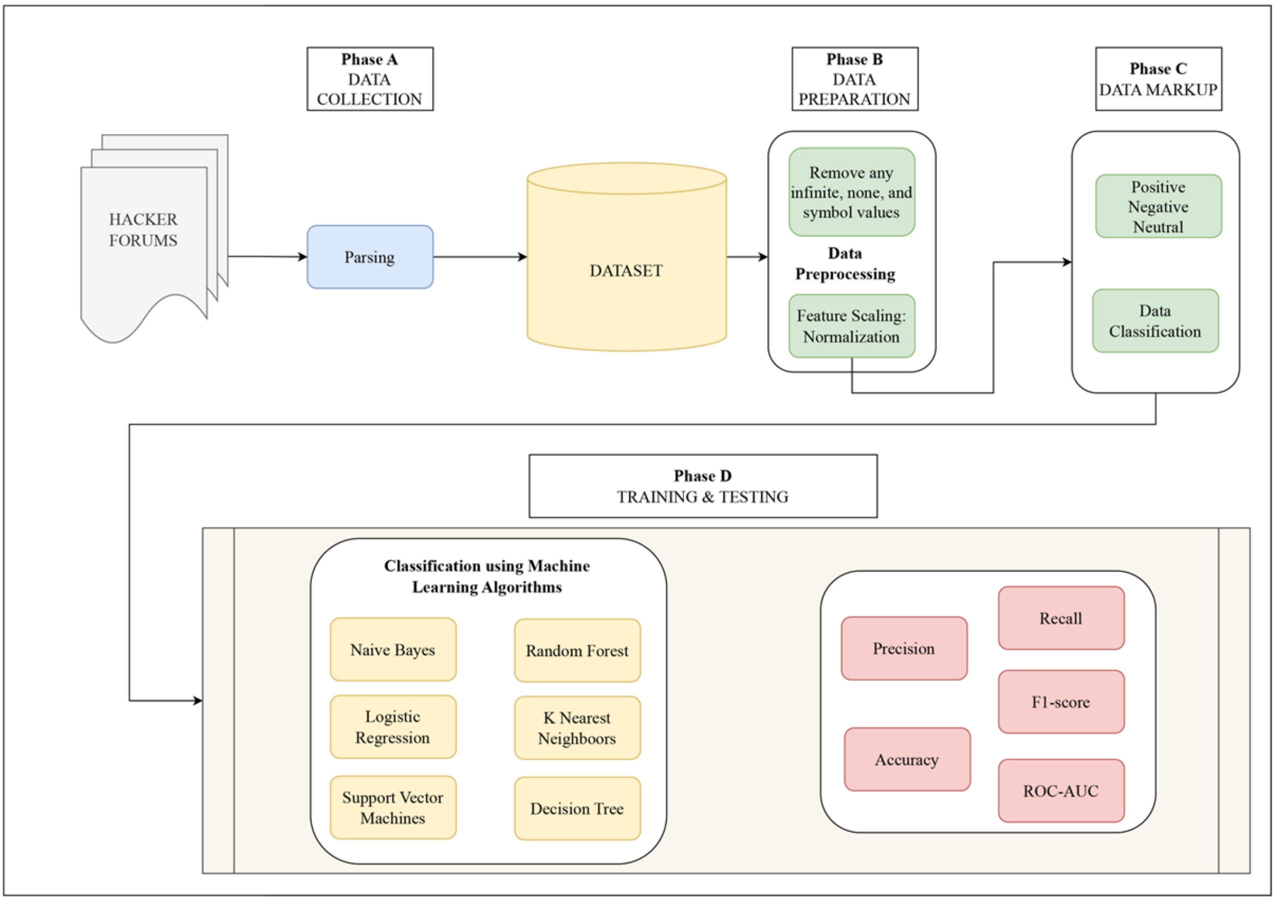


Fig. 1. Phases in identifying potential threats and vulnerabilities

5. Results of investigating the effectiveness of detection and classification of threats and vulnerabilities using machine learning

5. 1. Data collection from hacker forums

In the first step of phase A (Fig. 1) of identifying potential threats and vulnerabilities, links to the most popular English-language hacker forums on the Internet are collected, where new topics are most often opened, and content is updated. At the next stage, data collection (messages, discussions, and other content) from the specified forums was carried out. There are several approaches to collecting data from Internet resources; in this work, it was decided to apply parsing [34, 35].

Thus, the data was collected using a parser developed using the BeautifulSoup4 Python library [36]. A total of 150,000 posts were collected from the forums. These data were taken on the basis of the semantic features indicated in Table 1. Collected data was saved in csv format. A visual representation of the parsing operation is shown in Fig. 2.

Inputs for semantic text analysis can be described as follows:

$$X=\{x_1,x_2,...,x_i,...,x_n\}, \qquad (1)$$

where x_i is a key feature, which can be expressed by a single word, a combination of words, a formulated expression, etc., and which is used in data collection; $i\in[0, N]$, N is the volume of key data.

Table 1

Semantic features

No.	Feature	Description
1	message_text	Forum post/comment
2	message_date	Time to write a post/comment on the forum
3	thread_name	Name of a local forum thread
4	username	Username
5	username_link	Link to the user's personal account on the forum
6	user_title	Nickname of the user
7	user_msg_number	Number of user posts on the forum
8	user_reputation	User's reputation compared to other users on the forum
9	user_join_date	Time of user registration on the forum
10	user_reactions	Number of reactions to user messages left by other users

The next step was the identification of resources, where key units are defined. At this stage, after data collection, a predictive analysis [37] was conducted taking into account the latest resources from the Internet. Here, too, the number of researched resources was limited. Intermediate data using (1) can be represented as:

$$X^*=A_j(x_i), \qquad (2)$$

where A_j is a resource with key features; $j\in[0, M]$, M is the number of resources under consideration.

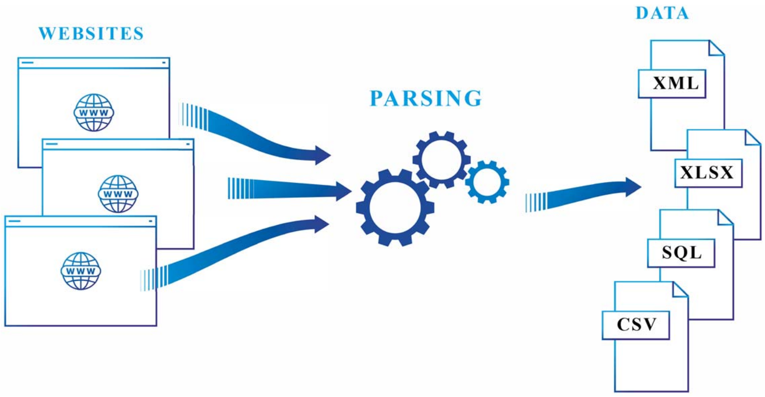


Fig. 2. Parsing scheme

Input data for analysis is in the form of a vector of Internet resources. After the resources were determined, the output was discarded, leaving a list of the resulting resources that can be written as a vector:

$$A^* = \left(A_1, A_2, \ldots, A_j, \ldots, A_m\right)^T. \tag{3}$$

The definition of output data is described as follows:

$$Y = \left\{y_1, y_2, \ldots, y_i, \ldots, y_n\right\}, \tag{4}$$

where y_i is the result of the analysis, new knowledge obtained as a result of data processing, $i \in [0, N]$, N is the number of additional or limited results that the user can specify.

By analogy with (3), (4) can be transformed into the following form:

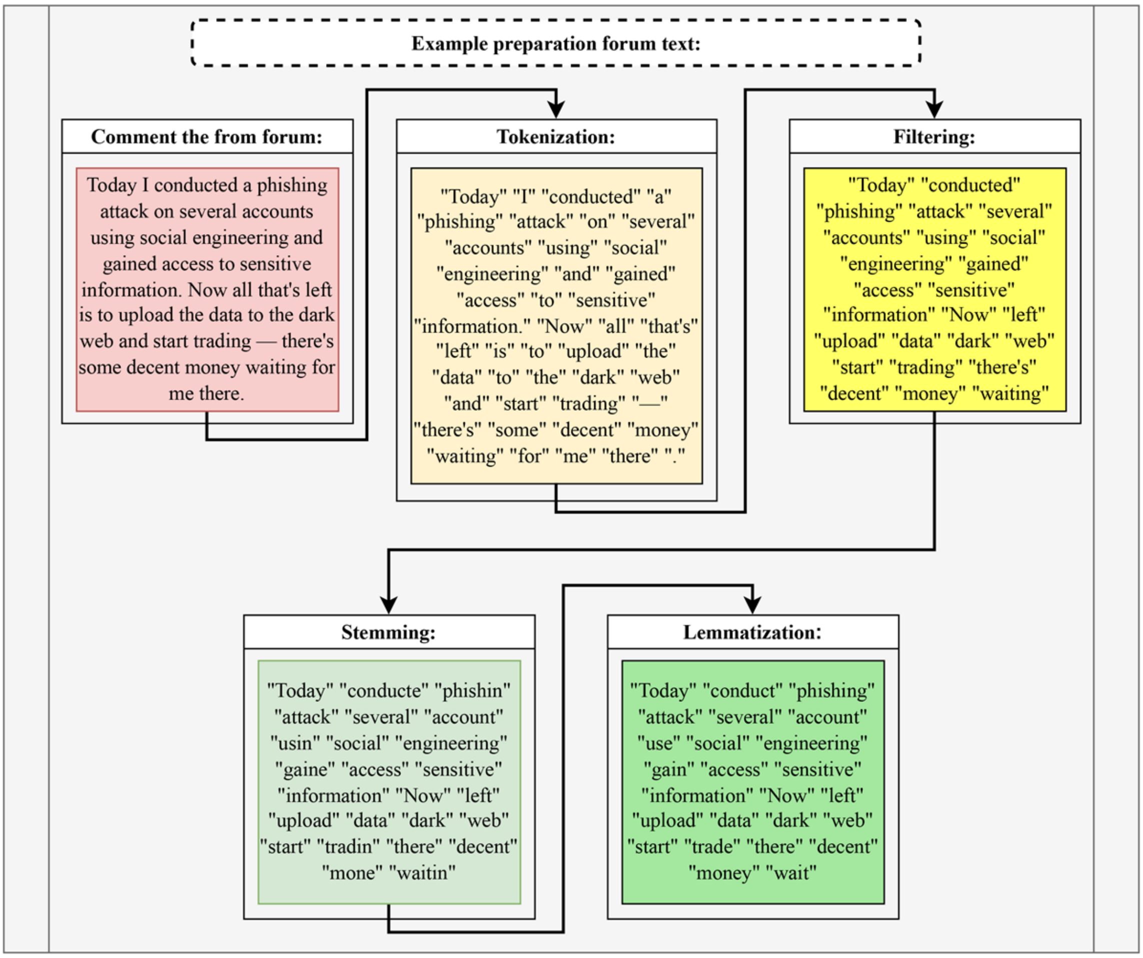
$$Y^* = \left(A_1^*, A_2^*, \ldots, A_j^*, \ldots, A_m^*\right)^T, \tag{5}$$

where A_j^* is the result of the analysis of intermediate data in the form of a vector A^*, $j \in [0, M]$.

After the successful execution of the script, a large array of data appeared, which reflects all discussions, opinions and trends on the selected topic on the forum. This array allowed us to provide additional analysis where it is possible to determine the mood, emotions of users, dominant topics of discussion in the forum and study of user behavior patterns based on interaction and content creation. This facilitates the integration of the script into the monitoring system, which is periodically practiced in real time, and will allow for continuous data collection and tracking of trends, increasing the depth and accuracy of conclusions from the obtained data.

5. 2. Data preparation

The second phase after the collection was data preparation (Fig. 3), in which normalization and pre-processing of the text data from the forums was carried out. Of all the data collected in the previous phase, 120,000 were prepared for training the model, and the remaining 30,000 were left for testing the effectiveness of the trained models.



Example preparation forum text:

Comment the from forum:

Today I conducted a phishing attack on several accounts using social engineering and gained access to sensitive information. Now all that's left is to upload the data to the dark web and start trading — there's some decent money waiting for me there.

Tokenization:

"Today" "I" "conducted" "a" "phishing" "attack" "on" "several" "accounts" "using" "social" "engineering" "and" "gained" "access" "to" "sensitive" "information." "Now" "all" "that's" "left" "is" "to" "upload" "the" "data" "to" "the" "dark" "web" "and" "start" "trading" "—" "there's" "some" "decent" "money" "waiting" "for" "me" "there" "."

Filtering:

"Today" "conducted" "phishing" "attack" "several" "accounts" "using" "social" "engineering" "gained" "access" "sensitive" "information" "Now" "left" "upload" "data" "dark" "web" "start" "trading" "there's" "decent" "money" "waiting"

Stemming:

"Today" "conducte" "phishin" "attack" "several" "account" "usin" "social" "engineering" "gaine" "access" "sensitive" "information" "Now" "left" "upload" "data" "dark" "web" "start" "tradin" "there" "decent" "mone" "waitin"

Lemmatization:

"Today" "conduct" "phishing" "attack" "several" "account" "use" "social" "engineering" "gain" "access" "sensitive" "information" "Now" "left" "upload" "data" "dark" "web" "start" "trade" "there" "decent" "money" "wait"

Fig. 3. Example of data preparation

Normalization was carried out by removing unnecessary symbols, tags, punctuation marks that can distort the structure and content of the text. After that, the text is divided into separate words and/or phrases, which is called tokenization [38]. Tokenization made it possible to conduct a deeper analysis of certain elements of the text. After tokenization, lemmatization and stemming were performed to reduce the various forms and word-forms to their basic forms [39]. Lemmatization is necessary for simplification, eliminating the need to analyze each form separately, preserving its semantic value. Stemming, in turn, reduces words to their root forms for unification and makes it easier to compare words with the same root. Next comes the filtering of words that do not carry a semantic load, but are often found in the text: prepositions, conjunctions, etc. Filtering reduces the dimensionality of data and thereby increases informativeness [40]. After that, empty tokens and words that may have lost their meaning after previous operations are removed. At this stage, the data is ready for sentiment analysis. This step emphasizes the importance of data cleanliness and structuring to produce more accurate, interpretable results.

5. 3. Marking up received data

After preparation, the collected text data were marked to determine their tonality based on a previously compiled dictionary (Fig. 4): positive, neutral, or negative. The pre-compiled dictionary contains a set of keywords, phrases, idioms, slang, etc. related to hacker terminology, which refer to positive, negative, and neutral emotional shades. It should be noted that Fig. 4 shows an example of key text data, and the dictionary development process continued with model training.

The markup process began by analyzing each token that was prepared in the previous phase for matching keywords from the dictionary. If the token matched a positive or negative keyword, the text was labeled accordingly. If a token did not match any of the keywords, it was assigned a neutral label. This made it possible to assess the emotional color of each text in the analyzed sample.

After the marking process was completed, each text was enriched with information about its tonality, which allowed for a deeper analysis of emotional trends and relationships in forum discussions. This step emphasizes the importance of tonality analysis for understanding the emotional component of the data and its additional use in research and analysis.

5. 4. Training of models to detect threats and vulnerd abilities and evaluation of effectiveness

Training the models to identify potential threats and vulnerabilities in textual data was performed on a single training data set of 120,000 posts collected from hacker forums. In total, six models were trained using some of the most common machine learning algorithms: kNN, Random Forest, Naive Bayes, Logistic Regression, SVM, and Decision Tree [41]. Python 3.6, Scikit-learn 0.23.2, Numpy 1.19.2, Pandas 1.1.3 and SciPy 1.5.2 were used to develop the scripts. All tests were performed on an Asus Vivobook Pro 15 M6500QC-MA145 laptop with an AMD Ryzen 7 processor with a clock frequency of 3.2 GHz and 16 GB of RAM with Windows 10 installed. The training time of each model is shown in Fig. 5.

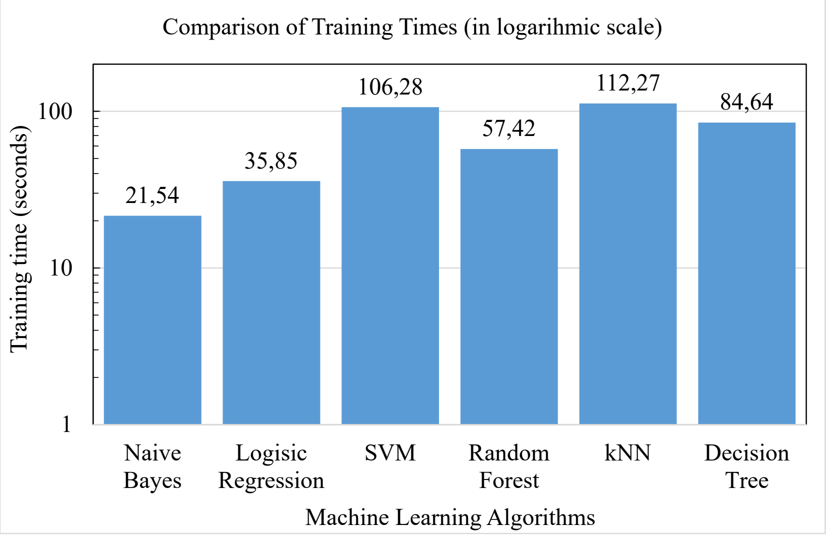


Fig. 5. Time to train models to identify threats and vulnerabilities

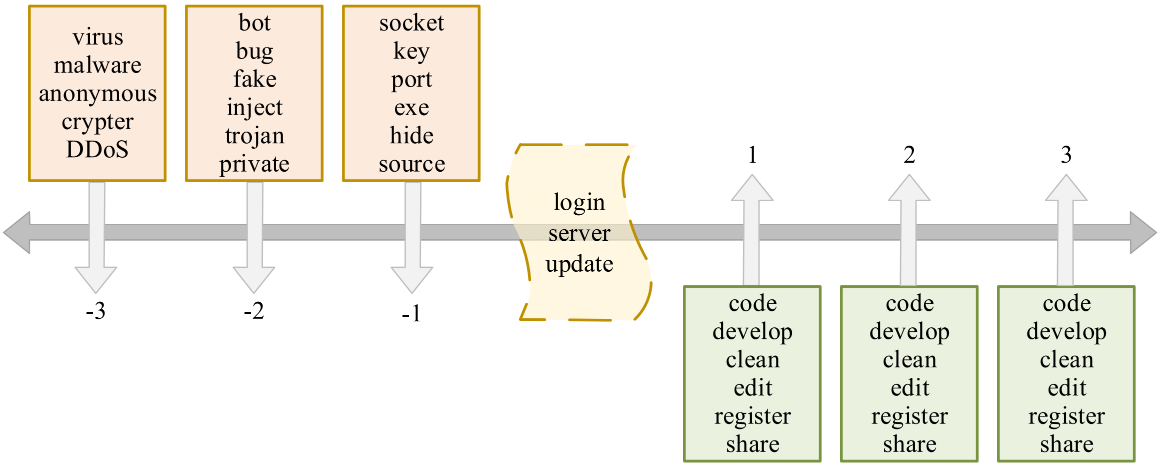


Fig. 4. Suggested sentiment dictionary for positive and negative keywords

After training the models on labeled data, they were tested on their performance in detecting potentially dangerous content contained in unlabeled data from forums that were not used during training. A total of 30,000 text messages were used in the test out of the 150,000 comments previously collected.

Table 2 gives the number of correctly and incorrectly classified objects, where *TR* is True Positive, *TN* is True Negative, *FP* is False Positive, *FN* is False Negative.

Table 2

Confusion matrix for the machine learning algorithms used

Algorithms	*TP*	*FN*	*FP*	*TN*
Naive Bayes	10,550	2,637	1,862	14,951
Logistic Regression	10,960	2,406	1,494	15,140
SVM	10,571	2,480	1,720	15,229
Random Forest	11,407	2,173	1,127	15,293
kNN	9,842	2,940	2,460	14,758
Decision Tree	12,652	1,725	2,772	12,851

In addition to the confusion matrix, other metrics were also used to conduct a comparative analysis of the effectiveness of detecting threats and vulnerabilities, such as precision, recall, accuracy, *F*1-score, and ROC-AUC [42]. Accuracy measures the proportion of correct predictions relative to all predictions, recall measures the proportion of all real positive cases found by the model, *F*1-score combines recall and precision for a more objective assessment. ROC-AUC allows one to evaluate the performance of a model with a single number that is equal to the area under the ROC curve. The values of these indicators were calculated according to the following formulas:

$$accuracy = \frac{TP + TN}{TP + TN + FP + FN}, \qquad (6)$$

$$precision = \frac{TP}{TP + FP}, \qquad (7)$$

$$recall = \frac{TP}{TP + FN}, \qquad (8)$$

$$F1\text{-score} = \frac{2 \cdot precision \cdot recall}{precision + recall}, \qquad (9)$$

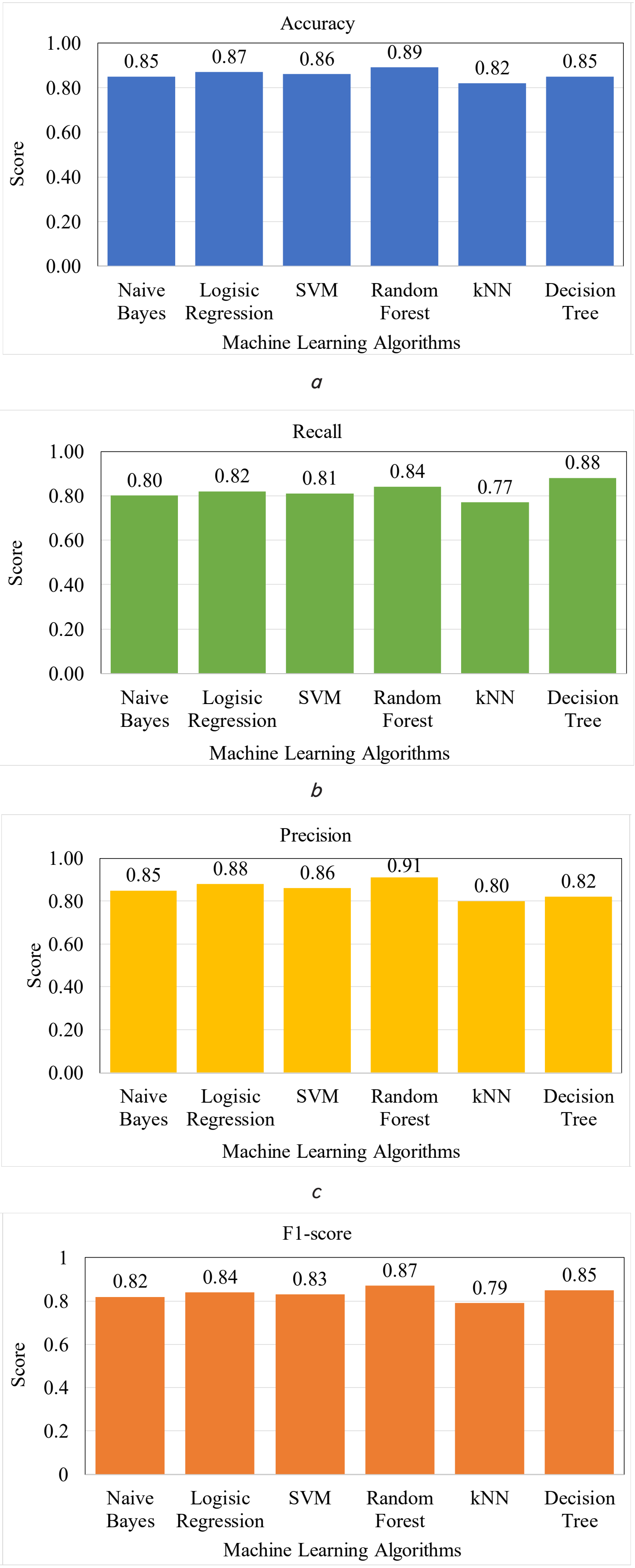
$$FPR = \frac{TP}{FP + TN}, \qquad (10)$$

where *FPR* is the False Positive Rate, which together with (8) is used to construct the ROC curve and calculate the AUC under this curve.

The values of efficiency indicators calculated according to the above formulas (6) to (9), as well as the ROC-AUC result, are given in Table 3.

For the convenience of analysis, a comparison of indicators in graphical form is shown in Fig. 6, 7.

Given in Tables 2, 3, and Fig. 6, 7, the data show that most of the tested machine learning algorithms used to detect threats and vulnerabilities contained in hacker forums show reasonably high results. Almost all metric values except recall and *F*1-score for the kNN algorithm have a value of 0.8 or more with an ideal value of 1.0.



a

b

c

d

Fig. 6. Comparison of the effectiveness of six machine learning algorithms: *a* – accuracy; *b* – recall; *c* – precision; *d* – *F*1-score

Table 3

Evaluation of the effectiveness of various machine learning algorithms

Algorithm	Accuracy	Recall	Precision	*F*1-score	ROC-AUC
Naive Bayes	0.85	0.80	0.85	0.82	0.84
Logistic Regression	0.87	0.82	0.88	0.84	0.87
SVM	0.86	0.81	0.86	0.83	0.85
Random Forest	0.89	0.84	0.91	0.87	0.89
kNN	0.82	0.77	0.80	0.79	0.81
Decision Tree	0.85	0.88	0.82	0.85	0.85

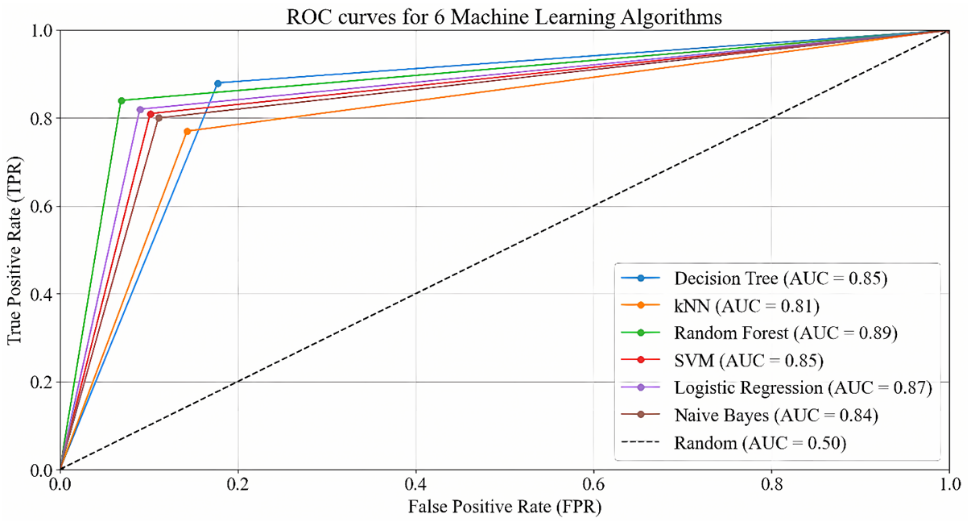


Fig. 7. ROC curves for different learning algorithms along with their AUCs

6. Discussion of results of the application of various algorithms for detection and classification of threats and vulnerabilities

Data collection and preparation performed in phases A to C (Fig. 1) are preparatory actions necessary for training models to detect threats and vulnerabilities contained in the content of hacker forums. Careful preparation and marking of data provided a reduction in model training time and affected the quality of the trained models. However, Fig. 5 shows that training models on the same data with different algorithms required different times. The minimum time of 21.54 s was required to train the model with the Naive Bayes algorithm, while 106.28 s and 112.27 s were spent on training the SVM and kNN algorithms, respectively. Such different training time of models on the same data set is explained by the internal characteristics of the algorithms, the peculiarities of setting hyperparameters, the number of iterations, the efficiency of calculations, and the possibility of parallel data processing.

As can be seen from Table 3 and Fig. 6, 7, when detecting and classifying threats and vulnerabilities on hacker Internet forums, the Random Forest algorithm showed the best result, showing the highest values for such metrics as accuracy – 0.89, precision – 0.91, *F*1-score – 0.87, and ROC-AUC – 0.89. Conversely, the kNN algorithm showed the lowest result, namely accuracy equal to 0.82, recall equal to 0.77, precision equal to 0.80, and *F*1-score equal to 0.79.

Despite the fact that the Decision Tree algorithm is slightly better in terms of recall (0.88 versus 0.84 in the Random Forest algorithm), higher values for the rest of the metrics determine the advantage of Random Forest. This advantage is due to its ensemble approach, which combines the results of multiple decision trees, reducing the probability of overtraining and improving the generalization ability of the model. A high value of precision indicates accuracy in predicting positive cases, and a high recall indicates the ability to detect most threats and vulnerabilities. Other algorithms, such as Logistic Regression and SVM, also performed well, but were inferior to Random Forest due to lower robustness to overtraining and lower generalization ability. The smallest values of the metrics shown by the model trained by the kNN algorithm can be explained by its sensitivity to the amount of data and the number of features. Thus, our results indicate that the Random Forest algorithm is the most attractive for training models for detecting and classifying threats and vulnerabilities in hacker forums.

It is worth noting that improving the quality of the classification of threats and vulnerabilities is related to the thoroughness of the collection and preparation of input data. This preparation made it possible to form a high-quality database for further processing. Data preparation consisted of pre-processing and text normalization (Fig. 3). Data marking was performed on the basis of a previously prepared dictionary of hacker terminology (Fig. 4).

The high performance indicators of the model obtained during the experiment show that the proposed tool is effective in detecting threats and vulnerabilities discussed on hacker Internet forums, which allows for prompt application of appropriate measures.

A limitation of our study is that some algorithms are sensitive to changes in data or model parameters. This can lead to unstable results and make them difficult to interpret. As a limitation, it should also be noted that the given results are obtained on the basis of specific data taken from hacker Internet forums. The behavior of forum participants cannot be modeled 100 % based on the amount of data obtained during the research.

It should also be noted the shortcomings of the proposed tool for detecting risks and vulnerabilities that the content of hacker forums carries. The first disadvantage is its sensitivity to noise in forums, which are specially created by their participants for the purpose of their own security. Another important factor that attracts attention is the correct data markup. If models are trained on incomplete or inaccurate data, this will lead to poor performance.

Another disadvantage is that some threats and vulnerabilities may be hidden or unknown at the time the model is trained. This can lead to underestimation of certain types of threats or vulnerabilities. In addition, some data characteristics that may be important for detecting threats and vulnerabilities may not be known at the time the models are trained. This can make it difficult to effectively use the models to detect new threats.

A possible further development of the research may consist in the construction of a hybrid model for detection of threats and vulnerabilities using artificial neural networks.

7. Conclusions

1. To identify and classify threats and vulnerabilities, a Python script was developed that collects text data from hacker forums. 150,000 posts and comments have been collected from the most active forums. The obtained data are transformed into a type suitable for further analysis and saved as a csv file using the proposed semantic features of the text and its author. This data was later used to train models and test algorithms when identifying potential threats to Internet users that contain hacker forum content.

2. A procedure for pre-processing data obtained from forums by removing non-informative symbols and normalizing essential features has been devised. This made it possible to prepare the data after preprocessing for qualitative training of threat detection models. Of all the collected data, 120,000 were prepared for training the model, and the remaining 30,000 were prepared for testing the performance of the trained models.

3. Marking of the processed data was performed based on a pre-compiled dictionary of hacker terminology and slang with manually adjusted weights. Marking made it possible to classify posts and comments on forums: negative, positive, or neutral to quickly determine the degree of danger contained in a particular message.

4. 120,000 prepared data were used to train models to detect potential threats contained in them. Six popular tutored machine learning algorithms were used to determine the most appropriate algorithm for training the model: kNN, Random Forest, Naive Bayes, Logistic Regression, SVM, and Decision Tree. The trained models were used to detect threats on new unlabeled data contained in 30,000 text messages. The analysis of the test results revealed that the best indicators according to the total accuracy/recall/precision criterion were shown by the Random Forest algorithm, which gave an efficiency of 0.89 according to the accuracy criterion, 0.84 according to the recall criterion, 0.91 according to the precision criterion, and $F1$-score – 0 .87 points, as well as ROC-AUC – 0.89. This indicates that the Random Forest algorithm is the most resistant to the quality of the collected data and the approach to its marking.

Conflicts of interest

The authors declare that they have no conflicts of interest in relation to the current study, including financial, personal, authorship, or any other, that could affect the study and the results reported in this paper.

Funding

The study was conducted without financial support.

Data availability

The data will be provided upon reasonable request.

Use of artificial intelligence

The authors confirm that they did not use artificial intelligence technologies when creating the current work.

References

1. Mambetov, S., Begimbayeva, Y., Joldasbayev, S., Kazbekova, G. (2023). Internet threats and ways to protect against them: A brief review. 2023 13th International Conference on Cloud Computing, Data Science & Engineering (Confluence). https://doi.org/10.1109/confluence56041.2023.10048858
2. Dhake, B., Shetye, C., Borhade, P., Gawas, D., Nerurkar, A. (2023). Stratification of Hacker Forums and Predicting Cyber Assaults for Proactive Cyber Threat Intelligence. 2023 2nd International Conference on Paradigm Shifts in Communications Embedded Systems, Machine Learning and Signal Processing (PCEMS). https://doi.org/10.1109/pcems58491.2023.10136033
3. Leukfeldt, E. R., Kleemans, E. R., Stol, W. P. (2016). Cybercriminal Networks, Social Ties and Online Forums: Social Ties Versus Digital Ties within Phishing and Malware Networks. British Journal of Criminology, azw009. https://doi.org/10.1093/bjc/azw009
4. Shakarian, J., Gunn, A. T., Shakarian, P. (2016). Exploring Malicious Hacker Forums. Cyber Deception, 259–282. https://doi.org/10.1007/978-3-319-32699-3_11
5. Mikhaylov, A., Frank, R. (2016). Cards, Money and Two Hacking Forums: An Analysis of Online Money Laundering Schemes. 2016 European Intelligence and Security Informatics Conference (EISIC). https://doi.org/10.1109/eisic.2016.021
6. Abbasi, A., Li, W., Benjamin, V., Hu, S., Chen, H. (2014). Descriptive Analytics: Examining Expert Hackers in Web Forums. 2014 IEEE Joint Intelligence and Security Informatics Conference. https://doi.org/10.1109/jisic.2014.18
7. Zhang, X., Li, C. (2013). Survival analysis on hacker forums. SIGBPS workshop on business processes and service, 106–110.
8. Tariq, E., Akour, I., Al-Shanableh, N., Alquqa, E. K., Alzboun, N., Al-Hawary, S. I. S., Alshurideh, M. T. (2024). How cybersecurity influences fraud prevention: An empirical study on Jordanian commercial banks. International Journal of Data and Network Science, 8 (1), 69–76. https://doi.org/10.5267/j.ijdns.2023.10.016
9. Karuna, P., Purohit, H., Jajodia, S., Ganesan, R., Uzuner, O. (2021). Fake Document Generation for Cyber Deception by Manipulating Text Comprehensibility. IEEE Systems Journal, 15 (1), 835–845. https://doi.org/10.1109/jsyst.2020.2980177

10. Rebafka, T. (2023). Model-based clustering of multiple networks with a hierarchical algorithm. Statistics and Computing, 34 (1). https://doi.org/10.1007/s11222-023-10329-w
11. Fu, T., Abbasi, A., Chen, H. (2010). A focused crawler for Dark Web forums. Journal of the American Society for Information Science and Technology, 61 (6), 1213–1231. https://doi.org/10.1002/asi.21323
12. McAlaney, J., Kimpton, E., Thackeray, H. (2019). Fifty shades of grey hat: A socio-psychological analysis of conversations on hacking forums. CyPsy24: Annual CyberPsychology, CyberTherapy & Social Networking Conference. Available at: https://eprints.bournemouth.ac.uk/32495
13. McAlaney, J., Hambidge, S., Kimpton, E., Thackray, H. (2020). Knowledge is power: An analysis of discussions on hacking forums. 2020 IEEE European Symposium on Security and Privacy Workshops (EuroS&PW). https://doi.org/10.1109/eurospw51379.2020.00070
14. Lacey, D., Salmon, P. M. (2015). It's Dark in There: Using Systems Analysis to Investigate Trust and Engagement in Dark Web Forums. Lecture Notes in Computer Science, 117–128. https://doi.org/10.1007/978-3-319-20373-7_12
15. Benjamin, V., Valacich, J. S., Chen, H. (2019). DICE-E: A Framework for Conducting Darknet Identification, Collection, Evaluation with Ethics. MIS Quarterly, 43 (1), 1–22. https://doi.org/10.25300/misq/2019/13808
16. Zhang, Y., Fan, Y., Ye, Y., Zhao, L., Wang, J., Xiong, Q., Shao, F. (2018). KADetector: Automatic Identification of Key Actors in Online Hack Forums Based on Structured Heterogeneous Information Network. 2018 IEEE International Conference on Big Knowledge (ICBK). https://doi.org/10.1109/icbk.2018.00028
17. Park, A. J., Frank, R., Mikhaylov, A., Thomson, M. (2018). Hackers Hedging Bets: A Cross-Community Analysis of Three Online Hacking Forums. 2018 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM). https://doi.org/10.1109/asonam.2018.8508613
18. Macdonald, M., Frank, R., Mei, J., Monk, B. (2015). Identifying Digital Threats in a Hacker Web Forum. Proceedings of the 2015 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining 2015. https://doi.org/10.1145/2808797.2808878
19. Frank, R., Macdonald, M., Monk, B. (2016). Location, Location, Location: Mapping Potential Canadian Targets in Online Hacker Discussion Forums. 2016 European Intelligence and Security Informatics Conference (EISIC). https://doi.org/10.1109/eisic.2016.012
20. Du, P.-Y., Zhang, N., Ebrahimi, M., Samtani, S., Lazarine, B., Arnold, N. et al. (2018). Identifying, Collecting, and Presenting Hacker Community Data: Forums, IRC, Carding Shops, and DNMs. 2018 IEEE International Conference on Intelligence and Security Informatics (ISI). https://doi.org/10.1109/isi.2018.8587327
21. Joldasbayev, S., Sapakova, S., Zhaksylyk, A., Kulambayev, B., Armankyzy, R., Bolysbek, A. (2023). Development of an Intelligent Service Delivery System to Increase Efficiency of Software Defined Networks. International Journal of Advanced Computer Science and Applications, 14 (12). https://doi.org/10.14569/ijacsa.2023.0141267
22. Balakayeva, G., Ezhilchelvan, P., Makashev, Y., Phillips, C., Darkenbayev, D., Nurlybayeva, K. (2023). Digitalization of enterprise with ensuring stability and reliability. Informatyka, Automatyka, Pomiary w Gospodarce i Ochronie rodowiska, 13 (1), 54–57. https://doi.org/10.35784/iapgos.3295
23. Balakayeva, G., Zhanuzakov, M., Kalmenova, G. (2023). Development of a digital employee rating evaluation system (DERES) based on machine learning algorithms and 360-degree method. Journal of Intelligent Systems, 32 (1). https://doi.org/10.1515/jisys-2023-0008
24. Balakayeva, G., Darkenbayev, D., Zhanuzakov, M. (2023). Development of a software system for predicting employee ratings. Informatyka, Automatyka, Pomiary w Gospodarce i Ochronie rodowiska, 13 (3), 121–124. https://doi.org/10.35784/iapgos.3723
25. Joldasbayev, S., Balakayeva, G., Joldasbayev, O. (2020). Application of load balancing algorithms to improve the quality of service delivery using modifications of the least connections algorithm. Journal of Theoretical and Applied Information Technology, 98 (12), 2063–2077. Available at: http://www.jatit.org/volumes/Vol98No12/7Vol98No12.pdf
26. Huang, C., Guo, Y., Guo, W., Li, Y. (2021). HackerRank: Identifying key hackers in underground forums. International Journal of Distributed Sensor Networks, 17 (5), 155014772110151. https://doi.org/10.1177/15501477211015145
27. Samtani, S., Chinn, R., Chen, H. (2015). Exploring hacker assets in underground forums. 2015 IEEE International Conference on Intelligence and Security Informatics (ISI). https://doi.org/10.1109/isi.2015.7165935
28. Benjamin, V., Li, W., Holt, T., Chen, H. (2015). Exploring threats and vulnerabilities in hacker web: Forums, IRC and carding shops. 2015 IEEE International Conference on Intelligence and Security Informatics (ISI). https://doi.org/10.1109/isi.2015.7165944
29. Deliu, I., Leichter, C., Franke, K. (2018). Collecting Cyber Threat Intelligence from Hacker Forums via a Two-Stage, Hybrid Process using Support Vector Machines and Latent Dirichlet Allocation. 2018 IEEE International Conference on Big Data (Big Data). https://doi.org/10.1109/bigdata.2018.8622469
30. Anand, M., Sahay, K. B., Ahmed, M. A., Sultan, D., Chandan, R. R., Singh, B. (2023). Deep learning and natural language processing in computation for offensive language detection in online social networks by feature selection and ensemble classification techniques. Theoretical Computer Science, 943, 203–218. https://doi.org/10.1016/j.tcs.2022.06.020
31. Sultan, D., Omarov, B., Kozhamkulova, Z., Kazbekova, G., Alimzhanova, L., Dautbayeva, A. et al. (2023). A Review of Machine Learning Techniques in Cyberbullying Detection. Computers, Materials & Continua, 74 (3), 5625–5640. https://doi.org/10.32604/cmc.2023.033682

32. Biswas, B., Mukhopadhyay, A., Bhattacharjee, S., Kumar, A., Delen, D. (2022). A text-mining based cyber-risk assessment and mitigation framework for critical analysis of online hacker forums. Decision Support Systems, 152, 113651. https://doi.org/10.1016/j.dss.2021.113651
33. Williams, R., Samtani, S., Patton, M., Chen, H. (2018). Incremental Hacker Forum Exploit Collection and Classification for Proactive Cyber Threat Intelligence: An Exploratory Study. 2018 IEEE International Conference on Intelligence and Security Informatics (ISI). https://doi.org/10.1109/isi.2018.8587336
34. Benjamin, G. (2021). What we do with data: a performative critique of data "collection." Internet Policy Review, 10 (4). https://doi.org/10.14763/2021.4.1588
35. Jain, S., de Buitleir, A., Fallon, E. (2020). A Review of Unstructured Data Analysis and Parsing Methods. 2020 International Conference on Emerging Smart Computing and Informatics (ESCI). https://doi.org/10.1109/esci48226.2020.9167588
36. Thivaharan., S., Srivatsun., G., Sarathambekai., S. (2020). A Survey on Python Libraries Used for Social Media Content Scraping. 2020 International Conference on Smart Electronics and Communication (ICOSEC). https://doi.org/10.1109/icosec49089.2020.9215357
37. Sarkar, S., Almukaynizi, M., Shakarian, J., Shakarian, P. (2019). Predicting enterprise cyber incidents using social network analysis on dark web hacker forums. The Cyber Defense Review, 87–102. Available at: https://www.jstor.org/stable/26846122
38. Ampel, B., Samtani, S., Zhu, H., Ullman, S., Chen, H. (2020). Labeling Hacker Exploits for Proactive Cyber Threat Intelligence: A Deep Transfer Learning Approach. 2020 IEEE International Conference on Intelligence and Security Informatics (ISI). https://doi.org/10.1109/isi49825.2020.9280548
39. Ampel, B., Chen, H. (2021). Distilling Contextual Embeddings Into A Static Word Embedding For Improving Hacker Forum Analytics. 2021 IEEE International Conference on Intelligence and Security Informatics (ISI). https://doi.org/10.1109/isi53945.2021.9624848
40. Samtani, S., Zhu, H., Chen, H. (2020). Proactively Identifying Emerging Hacker Threats from the Dark Web. ACM Transactions on Privacy and Security, 23 (4), 1–33. https://doi.org/10.1145/3409289
41. Sen, P. C., Hajra, M., Ghosh, M. (2019). Supervised Classification Algorithms in Machine Learning: A Survey and Review. Emerging Technology in Modelling and Graphics, 99–111. https://doi.org/10.1007/978-981-13-7403-6_11
42. Sokolova, M., Lapalme, G. (2009). A systematic analysis of performance measures for classification tasks. Information Processing & Management, 45 (4), 427–437. https://doi.org/10.1016/j.ipm.2009.03.002