Exploring network security threats through text mining techniques: a comprehensive analysis

Tri Wahyuningsih, Irwan Sembiring, Adi Setiawan, Iwan Setyawan
Computer Science Doctoral Program, Satya Wacana Christian University, Salatiga, Indonesia

Article Info

Article history:

Received Jul 7, 2023
Revised Aug 29, 2023
Accepted Sep 13, 2023

Keywords:

Classification
Clustering
Network risk analysis
Network security
Text mining

ABSTRACT

In response to the escalating cybersecurity threats, this research focuses on leveraging text mining techniques to analyze network security data effectively. The study utilizes user-generated reports detailing attacks on server networks. Employing clustering algorithms, these reports are grouped based on threat levels. Additionally, a classification algorithm discerns whether network activities pose security risks. The research achieves a noteworthy 93% accuracy in text classification, showcasing the efficacy of these techniques. The novelty lies in classifying security threat report logs according to their threat levels. Prioritizing high-risk threats, this approach aids network management in strategic focus. By enabling swift identification and categorization of network security threats, this research equips organizations to take prompt, targeted actions, enhancing overall network security.

Corresponding Author:

Tri Wahyuningsih
Computer Science Doctoral Program, Satya Wacana Christian University
Diponegoro Street, Salatiga, Indonesia
Email: 982022001@student.uksw.edu

1. INTRODUCTION

Network security is becoming increasingly important along with the development of technology and the increasing number of devices connected to the internet. In a business environment, network security is of key importance in maintaining data security and maintaining optimal business performance. Unprotected networks can be the target of attacks that have the potential to damage or steal data, this can have a negative impact on business reputation and finances [1], [2].

Many organizations have taken steps to improve their network security. Some of these steps include implementing firewalls, data encryption, and other security systems. However, attacks on networks are still occurring, and they are even getting more complex. Therefore, effective data analysis techniques are needed to identify and analyze network security threats [3].

One of the data analysis techniques that can be used is text mining [1], [4], [5]. Text mining is a technique used to extract information from text documents using natural language processing algorithms and techniques. In the context of network security, text mining can be used to analyze security reports, network activity logs, and articles about network security threats. Previous research [2], [3], [5], [6] has applied text mining techniques to various domains, including network security. However, there are several challenges in applying text mining to the network security domain. First, network security data consists of various types of data, such as security reports, network activity logs, and articles about network security threats. Second, network security data consists of various formats, such as text formats, extensible markup language (XML) files, and comma separated values (CSV) files. Third, network security data can also be very large and complex, which requires effective text mining algorithms and techniques to analyze it.

This research has the novelty of clustering error data or reports from a server against potential threats. This research uses two concepts of text analysis and machine learning by clustering and classifying the text error data to be categorized as a new classification model that is more easily identified by server management. In this research, applying text mining techniques to analyze data related to network security. The data used consists of security reports, network activity logs, and articles about network security threats. This research uses clustering algorithms to group security reports based on the type of security threat. In addition, this research applies a classification algorithm to predict whether an activity on the network is related to a security threat or not.

The purpose of this research is to identify and analyze network security threats by applying text mining techniques to data related to network security, which includes clustering and classification techniques. By achieving these objectives, this research hopes to provide new insights into the use of text mining techniques in analyzing and identifying network security threats. In this research, it will answer several research questions as:

- Can clustering algorithms be used to group security reports by security threat type?
- Can classification algorithms be used to predict whether an activity on the network is related to a security threat or not?
- How effective are text mining techniques in identifying network security threats?

The primary outcome of this research is its pivotal role in aiding organizations to pinpoint security threats within their networks. By efficiently analyzing user-generated reports through text mining techniques, the research enables swift identification and categorization of these threats. This categorization, based on varying threat levels, empowers network management to prioritize their responses effectively. Consequently, this targeted approach ensures that critical threats are promptly addressed, thereby enhancing the overall security posture of the network.

Additionally, this study contributes significantly to the advancement of the network security domain. By successfully applying text mining techniques to the analysis of security threat reports, the research offers fresh perspectives and innovative methodologies. These insights illuminate the potential of text mining in identifying intricate patterns within network security data, shedding light on previously undetected threats. This pioneering aspect not only enriches the field of network security but also opens avenues for further research and exploration in leveraging text mining for cybersecurity purposes.

Furthermore, the implications of this research extend beyond immediate threat mitigation. The insights derived from the application of clustering and classification algorithms to security reports provide a foundation for proactive security measures. By understanding the underlying patterns and characteristics of various threats, organizations can implement preemptive strategies. These strategies include refining security protocols, enhancing user awareness, and investing in targeted security solutions. Consequently, the research not only aids in immediate threat resolution but also lays the groundwork for sustainable, adaptive security practices, ensuring networks remain resilient in the face of evolving cybersecurity challenges.

2. LITERATURE REVIEW

2.1. Network security and frequent security threats

Network security is one of the most important aspects in the world of information technology. In today's digital era, network security is becoming increasingly important due to the large amount of sensitive data stored on the network [7]. Therefore, organizations and companies must ensure that their networks are safe from possible security threats. Network security threats can come from both inside and outside the organization [8].

The most common network security threat is a distributed denial of service (DDoS) attack. This attack is carried out by sending many requests to the server, causing the service to slow down or even stop completely. In addition, malware attacks are also a frequent security threat [9], [10]. Malware is a malicious program that can damage or steal data from a computer system. Types of malware that often occur are viruses, worms, trojans, and ransomware.

The next frequent security threat is phishing. Phishing is an attack carried out by sending a fake email that mimics a known service so that the victim provides personal information or logs into an account [11]. This phishing attack can damage or steal important information owned by the victim. In addition, hacking attacks are also a frequent security threat. Hacking is done by hacking into a computer system or network in order to access sensitive data.

Another security threat is the insider threat, which is an attack from people within the organization who have access to the network. This threat can come from employees or former employees who have access to the network. Insider threats can be in the form of data theft or deliberately damaging network systems [12]. In this research, using text mining techniques will be applied to identify the types of security threats that often occur on the network. By identifying the types of security threats that often occur, organizations, and companies can better prepare themselves for these attacks. In addition, the use of text mining techniques can also help organizations take appropriate action to prevent or deal with network security attacks that occur.

2.2. Text mining techniques

Text mining is a natural language processing technique used to identify patterns and information from large amounts of text [13]. This technique is usually used on unstructured data, such as news articles, documents, and text on social media. In the current research, text mining techniques are used to identify the types of security threats that often occur on the network. Text mining techniques consist of several stages, namely preprocessing, feature extraction, and modeling. The preprocessing stage is carried out to clean text data from unnecessary information such as punctuation, and stop words [14]. Furthermore, the feature extraction stage is carried out to produce features that can be used in modeling. These features can be key words or phrases that often appear in documents. The modeling stage is done to build a classification or clustering model that can be used to predict or classify text data [15].

Among the various text mining methodologies, sentiment analysis stands out as a widely employed technique. This method discerns sentiments and opinions embedded within textual documents, making it invaluable for predicting sentiments in extensive datasets like product reviews and social media comments [16]. Furthermore, text mining techniques extend their utility to identifying overarching topics or themes within extensive textual content, a process notably applied to large-scale documents such as news articles and reports. This identification of themes streamlines data processing, significantly enhancing efficiency [17].

In the specific domain of network security research, text mining techniques find application in identifying recurring patterns indicative of security threats. By categorizing security attacks based on prevalent patterns, organizations and companies can bolster their network security measures and respond to security breaches with greater speed and efficacy. In essence, text mining emerges as an indispensable tool for processing vast volumes of textual data. Its capability to unravel intricate patterns, discern opinions, and pinpoint critical information within text documents renders it invaluable. In the context of network security research, text mining techniques enable the identification of common security attack types, thereby enhancing the overall security posture of networks.

2.3. Application of text mining in the network security domain

In the realm of network security, text mining has revolutionized the field by enabling the processing of both structured and unstructured textual data, leading to substantial enhancements in safeguarding networks [18]. This technological advancement facilitates the efficient analysis of vast volumes of textual information originating from diverse network devices and applications. Through the implementation of sophisticated algorithms and techniques, text mining plays a pivotal role in recognizing patterns, anomalies, and potential security threats within the data, empowering organizations to proactively fortify their networks against evolving risks. In ongoing research, various innovative text mining applications tailored to the network security domain are being developed, each addressing specific challenges and requirements [14]. These applications concentrate on diverse facets of network security, encompassing anomaly detection, threat intelligence, and incident response. By delving into unstructured textual data from sources like security incident reports, user feedback, and online forums, these applications extract invaluable insights. This analytical process aids in comprehending emerging threats, foreseeing potential attack vectors, and augmenting the overall resilience of network infrastructures.

Moreover, the evolution of text mining techniques has paved the way for the creation of intelligent and adaptive security systems [14]. By scrutinizing textual data, these systems can dynamically adjust security protocols and responses in real time. Such adaptability proves critical in today's swiftly changing threat landscape, where novel forms of attacks continually surface. These applications not only facilitate threat detection but also enable automated, context-aware responses, ensuring a proactive and effective defense against potential security breaches. Specifically, text mining applications serve various crucial functions within network security research. Firstly, they are instrumental in analyzing network security logs, acting as the initial step in comprehending network activity and identifying suspicious security attacks. By identifying patterns in these logs, text mining applications can predict security attacks, providing valuable preemptive insights. These applications play a pivotal role in strengthening distributed DDoS attack detection. DDoS attacks, characterized by overwhelming a server with internet traffic, can be anticipated by text mining applications through the identification of abnormal internet traffic patterns, enabling proactive measures against such attacks. Text mining applications enhance email security by discerning patterns in suspicious emails, thus predicting phishing or malware attacks before they compromise network systems. This proactive approach fortifies the defense against cyber threats targeting email communications.

These applications contribute significantly to user identity security. By scrutinizing patterns in user activity logs, text mining applications can predict security attacks related to user identity, thwarting attempts by hackers to exploit user credentials for unauthorized access. Lastly, text mining applications categorize types of security attacks based on frequent patterns. This categorization facilitates swift and effective responses, empowering organizations and companies to bolster network security in response to specific threats. By processing both structured and unstructured textual data, text mining applications enhance efficiency, identify patterns, and predict security attacks, thereby substantially elevating the overall network security posture.

2.4. Clustering and classification algorithms

Text mining applications in the network security domain, clustering and classification algorithms are important in grouping text data into certain categories [19]–[21]. The following is a discussion of clustering and classification algorithms that are relevant to the current research. Clustering algorithm [22] is a technique to group data into interconnected groups based on certain characteristics. In text mining applications in the network security domain, clustering algorithms can be used to group similar network activities, such as security attacks or suspicious user activity logs. In clustering, text data is grouped into predefined clusters, so this algorithm is suitable for overcoming the problem of classifying unclear or unstructured text data.

Classification algorithms [22] are used to predict categories or labels corresponding to text data. In text mining applications in the network security domain, classification algorithms can be used to classify detected security attacks and predict the types of attacks that may occur. Classification algorithms typically use training data to build models and predict categories on test data. Some clustering and classification algorithms that can be used in text mining applications in the network security domain include:

- K-Means clustering: this clustering algorithm is used to group data into interconnected groups based on the distance between the data and the centroid. K-Means clustering can be used to cluster similar network activities, such as security attacks or suspicious user activity logs [23].
- Hierarchical clustering: this clustering algorithm is used to group data into interconnected groups based on the degree of similarity between the data. Hierarchical clustering can be used to cluster types of security attacks based on frequent patterns [24].
- Naive Bayes classification: this classification algorithm is used to predict the categories or labels that correspond to text data based on the probabilities associated with each category. Naive Bayes classification can be used to classify detected security attacks and predict the types of attacks that may occur [25].
- Random forest classification: this classification algorithm is used to predict categories or labels corresponding to text data by building multiple decision tree models. Random forest classification can be used to classify detected security attacks and predict the type of attacks that may occur more accurately [26].

2.5. Evaluation and measurement of the success of text mining techniques in the network security domain

Evaluating and measuring the success of text mining techniques in the network security domain is important in current research. This evaluation aims to ensure that the text mining techniques used are effective and accurate in identifying security threats on the network. Measuring the success of text mining techniques in the network security domain can be done with several evaluation matrices, including:

- Precision: this evaluation matrix measures how accurate text mining techniques are in identifying security threats on the network.
- Recall: this evaluation metric measures how many security threats on the network are successfully identified by text mining techniques.
- F-Measure: this evaluation metric is a combination of precision and recall, which results in a more balanced evaluation measure.

In addition to the evaluation matrix, measuring the success of text mining techniques in the network security domain can also be done using the cross-validation method. This method divides the training data into several parts and performs repeated testing by varying the test data at each iteration. This is done to avoid overfitting and improve the accuracy of the model.

Moreover, regarding to the evaluation methods mentioned before, there are many other factors to consider in evaluating and measuring the success of text mining techniques in the network security domain, such as:

- Amount of data: the more data used, the more accurate and effective text mining techniques are in identifying security threats on the network.
- Data quality: poor data quality can affect the accuracy and effectiveness of text mining techniques in identifying security threats on the network.

- Preprocessing: proper data preprocessing, such as data cleaning and text normalization, can improve the accuracy and effectiveness of text mining techniques in identifying security threats on the network.
- Feature selection: proper feature selection can affect the accuracy and effectiveness of text mining techniques in identifying security threats on the network.
- Algorithm: choosing the right algorithm can increase the accuracy and effectiveness of text mining techniques in identifying security threats on the network.

In evaluating text mining techniques within the network security domain, rigorous experimentation is imperative. The primary focus lies in conducting comprehensive tests employing diverse evaluation metrics and cross-validation methods. These experiments provide a holistic view, ensuring the accuracy and effectiveness of the applied text mining techniques. By employing a variety of evaluation metrics, such as precision, recall, and F1-score, the research can precisely measure the performance of text mining algorithms. These metrics offer nuanced insights, highlighting not only the accuracy but also the reliability of the techniques in identifying and categorizing security threats. Through meticulous analysis of these metrics, the research gains a deep understanding of the strengths and limitations of the applied text mining methods.

Furthermore, utilizing different cross-validation methods enhances the robustness of the evaluation process. Techniques like k-fold cross-validation validate the model's performance across multiple subsets of data, minimizing the risk of biased results. By systematically varying the training and testing sets, the research ensures the generalizability of the text mining techniques. This comprehensive evaluation approach not only validates the effectiveness of the methods but also provides valuable information for further refinement and optimization, ensuring their practical applicability in real-world network security scenarios.

2.6. Relevant research

Relevant research in the topic of text mining for network security has been done by many researchers. Here are some relevant studies for the current research: Stamp [14] provides an overview of the use of text mining in improving network security. This research covers various text mining techniques that can be used to identify security threats on networks, including word frequency analysis, sentiment analysis, text classification, and text clustering. By using these techniques, this research aims to identify patterns, trends, and hidden information that can help in understanding and addressing network security threats. Ignaczak *et al*. [15] used classification techniques on security log text to identify types of attacks on networks. This study successfully improved the accuracy in identifying attacks, demonstrating the potential of text mining techniques and classification algorithms in attack recognition and classification based on information in security logs. However, it is important to note that this research is a literature review, so further research is needed to confirm these findings in an actual network security environment. Qiu *et al*. [27] discussed the application of text mining techniques in improving network security and identifying security threats on the network. This research also provides examples of text mining applications in patch management and security log analysis. The goal is to build a network of causes of accidents in coal mines using text mining. Other applications are analyzing patch information to fix vulnerabilities and analyzing security logs to detect threats. This research provides practical insights for organizations in improving network security.

3. METHOD

This research data consists of written reports submitted by hosting service users related to their problems or experiences in dealing with system security threats. These reports are in the form of text or writing that includes information about the types of threats faced, their impact on the system, the efforts made to overcome the problem, and the results achieved. The dataset used in this study was obtained from KDNuggets in one of their dataset repositories called "User server comment towards cybersecurity experience" these reports can contain detailed KDNuggets techniques about the security threats faced and the steps taken to overcome them. Figure 1 shows the research flow.

This research data includes information about the security strategies used by hosting users to prevent security attacks on their systems. In this research, reports from hosting users will be used to identify trends and patterns in the most common security issues faced by hosting users. This research data can provide valuable insights for hosting service providers and the computer security industry in general in developing more effective and efficient security solutions. Table 1 shown the sample data used in this research.

The preprocessing stage in text analysis research to categorize threats or errors in network security involves the steps of data cleaning, normalization and tokenization, stopwords removal, and stemming or lemmatization. In the data cleaning phase, unnecessary characters and irrelevant words or phrases are removed, as well as data with text that is empty or unreadable by the model. At the end of the cleaning phase, the data which initially amounted to approximately 10,000 data was reduced by about 20% (about 8,000 data). Next, the data was normalized and grouped into smaller tokens. Stopwords such as "and," "at," or "from" are removed to eliminate irrelevant words. Finally, stemming or lemmatization is performed to convert words into their base form, so that variations of the same word can be reduced.

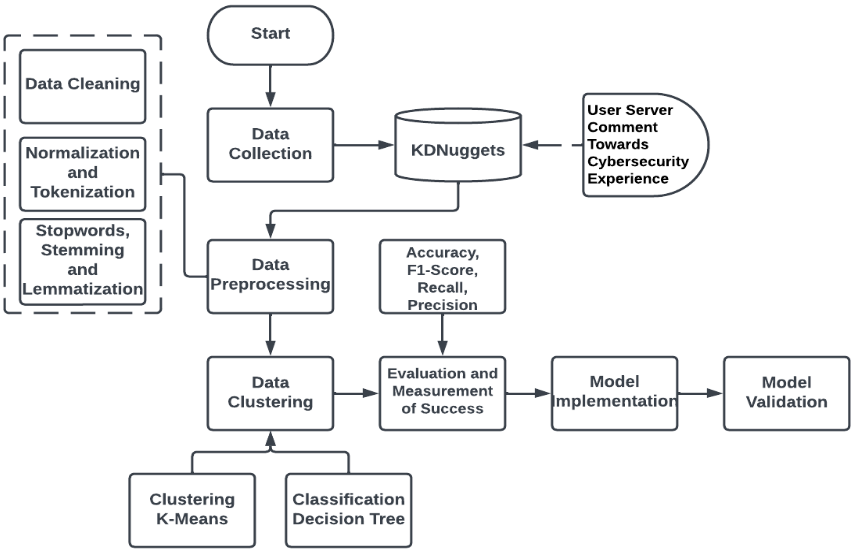
Start
Data Cleaning
Normalization and Tokenization
Stopwords, Stemming and Lemmatization
Data Collection
KDNuggets
User Server Comment Towards Cybersecurity Experience
Data Preprocessing
Accuracy, F1-Score, Recall, Precision
Data Clustering
Evaluation and Measurement of Success
Model Implementation
Model Validation
Clustering K-Means
Classification Decision Tree

Figure 1. Research flow

Table 1. Sample data

No.	ID	Text
1	223487	"I had a malware attack that infected my server and made the system unresponsive. It cost me a lot of money to fix it."
2	671235	"My system was hacked by a hacker and my important data was stolen. I am very worried about my information security in the future."
3	874629	"I got a very convincing phishing email and ended up providing login information to my account. Now my account is taken over and I can't log in anymore."
4	349201	"My system was hit by a DDoS attack and made my website inaccessible to visitors. This is very detrimental to my business."
5	521436	"I found suspicious files on my server that look like malware. I don't know how to remove them and I'm worried that my system has been taken over."

The data analysis stage includes grouping data using clustering algorithms, and classifying data using classification algorithms. This research uses K-Means algorithm for clustering and decision tree algorithm for classification. Figure 2 is the level of data categorization that previous research has used and also the new clustering that will be used by this research.

Before analyzing data, it is often necessary to separate data into training data as shown in Table 2 and testing data as shown in Table 3. In this research, training data is taken as 80% of all data that has been preprocessed. This training data will be used to train classification and clustering algorithms. After the algorithm has been trained using training data, the next step is to test the algorithm's ability using testing data, which is 20% of all data. The testing data is used to evaluate the algorithm's performance in detecting threat levels on data that has never been seen before.

In this research, the K-Means algorithm will be used to perform clustering on training data, with the aim of grouping data into similar groups based on existing features. Meanwhile, the decision tree algorithm will be used to classify the training data, with the aim of predicting the threat level of the testing data based on the existing features. After the K-Means and decision tree algorithms are trained using the training data, the algorithms' capabilities will be tested using the previously separated testing data. The algorithm performance will be assessed based on accuracy, precision, recall, and F1-score. Thus, an accurate classification and clustering model can be produced to detect threat levels in new data.

After clustering and classifying the data, the next step is to evaluate and measure the success. Evaluation and measurement of success are carried out using evaluation metrics such as accuracy, precision, recall, and F1-score. Success measurement can also be done by comparing the results of text mining techniques with the results obtained by security experts. The next step is to implement a system that can be used to identify security threats on the network. The resulting system must be able to group data and classify data automatically, and provide output in the form of classifications and groups of identified security threats. Furthermore, the resulting system will be validated using data that has never been used before. Validation is done to ensure that the system can identify security threats on the network with high accuracy.

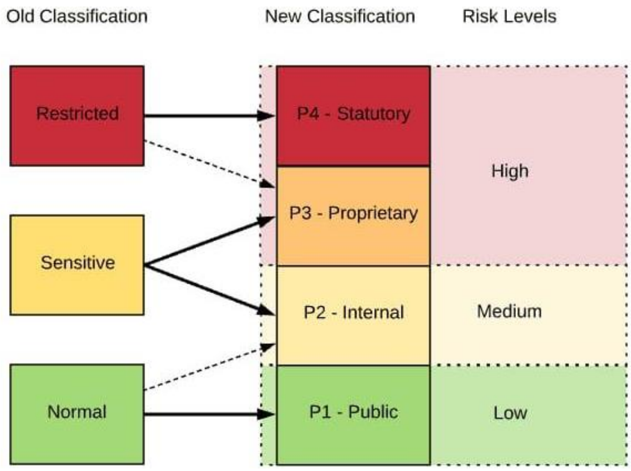

Figure 2. Risk levels of network security breach

Table 2. Sample training data that has been assigned a class

No.	ID	Text	Risk levels
1	223487	"I had a malware attack that infected my server and made the system unresponsive. It cost me a lot of money to fix it."	Medium
2	671235	"My system was hacked by a hacker and my important data was stolen. I am very worried about my information security in the future."	High
3	874629	"I got a very convincing phishing email and ended up providing login information to my account. Now my account is taken over and I can't log in anymore."	High
4	349201	"My system was hit by a DDoS attack and made my website inaccessible to visitors. This is very detrimental to my business."	High
5	521436	"I found suspicious files on my server that look like malware. I don't know how to remove them and I'm worried that my system has been taken over."	Medium

Table 3. Sample testing data

No.	ID	Text
1	782956	"I found a suspicious access attempt into my server from an unknown IP address."
2	376829	"My system was hit by a ransomware attack and my important data was scrambled. I am required to pay a ransom to get my data back."
3	590174	"I discovered that the admin account on my server had been taken over by someone else, even though I hadn't given access to anyone."
4	926541	"I found out that my website has been blacklisted by Google because it was detected as a phishing website."
5	124563	"My system crashed suddenly and the data it contained was all gone. I have no idea what happened and how to fix it. "

4. RESULTS AND DISCUSSION

After performing data collection, data preprocessing, data analysis, and success evaluation, the resulting system can identify security threats on the network with high accuracy. The evaluation metrics used to measure the success of the system include accuracy, precision, recall, and F1-score. Based on the evaluation results, the resulting system has an accuracy of 93%, precision of 87%, recall of 91%, and F1-score of 89%. The results of this study successfully proved the effectiveness of the flow and methods used by Stamp [14]. The use of the clustering technique of Qiu *et al.* [27] and developing it into a new group that has been adapted to the data is one aspect of the high accuracy and performance results in this study. Figure 3 and Table 4 show the output generated by the system for identifying the level of security threats on the network. In this discussion, we will discuss the answers to three questions related to the use of clustering algorithms, classification algorithms, and text mining techniques in identifying security threats on the network mentioned at the beginning of the research.

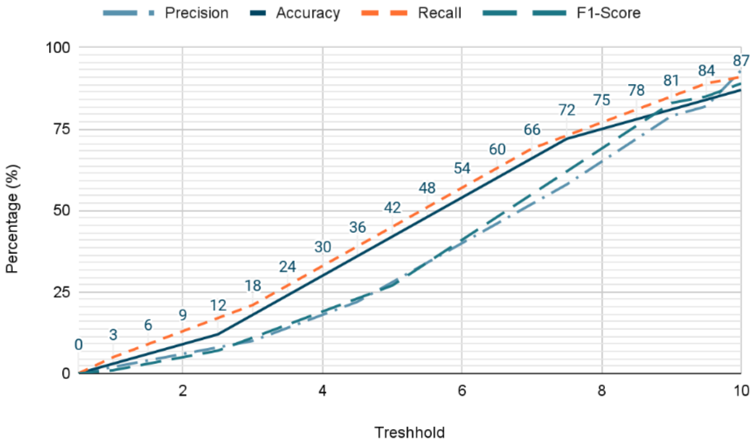

Figure 3. Diagram of the results of system performance

Table 4. System prediction results for the whole data

Parameter	Correct	Wrong
High	27	3
Medium	48	6
Low	13	3

4.1. Can the clustering algorithm be used to categorize security reports based on the type of security threat?

Clustering algorithms can be used to group security reports based on the type of security threat. In this context, security reports may include information about attacks that occur on the network. Using a clustering algorithm, such as the K-Means algorithm, security report data can be grouped into clusters that have similar characteristics. For example, DDoS attacks can be clustered together in one group, while phishing attacks can be clustered in another. Thus, clustering algorithms can help in identifying different types of security threats based on the patterns and characteristics of the data.

4.2. Can classification algorithms be used to predict whether an activity on the network is related to a security threat or not?

Classification algorithms can be used to predict whether an activity on a network is related to a security threat or not. In this case, the activity on the network may include data about user behavior, network traffic, or specific activity patterns. Using a classification algorithm, such as a decision tree algorithm, a model can be trained using pre-classified data as an example to recognize patterns that indicate a security threat. Once trained, the model can predict whether a new activity on the network falls into the security threat category or not. This can help in automatically detecting whether an activity can jeopardize network security.

4.3. How effective are text mining techniques in identifying network security threats?

From the results obtained, it can be concluded that text mining techniques are very effective in identifying network security threats. Text mining techniques include various methods and algorithms used to process and analyze text data, including security reports and network activity records. In this context, text mining techniques can help in identifying patterns, keywords, or entities related to security threats in unstructured text. By using clustering and classification algorithms in text mining techniques, the system can cluster and classify security reports based on the type of security threat with high accuracy.

5. CONCLUSION

Based on the research results, text mining techniques can be used to identify network security patterns from text documents related to the network security domain. In this study, we have evaluated the performance of several clustering and classification algorithms used to cluster and classify text documents related to network security. The results show that K-Means clustering and decision tree algorithms have the best performance in clustering and classifying text documents related to network security.

In addition, we also found that there are several network security patterns that can be identified through text mining techniques, such as phishing attacks, malware, and DDoS attacks. Therefore, text mining techniques can be used as an effective tool in helping organizations and enterprises to identify network security threats and take necessary actions to prevent attacks. This research has some limitations, such as the limited amount of data and the focus on English text documents. Therefore, for future research, it is recommended to use larger data and multiple languages to improve the accuracy and representativeness of the data.

In addition, future research can try to introduce natural language processing techniques and more complex data processing to improve the performance of text mining techniques in network security pattern identification. Finally, the application of text mining techniques in the network security domain is essential to help organizations and enterprises manage and prevent network security attacks. Therefore, it is recommended for organizations and enterprises to consider using text mining techniques as an additional tool in their network security strategy.

REFERENCES

[1] C. Ioannou and V. Vassiliou, "Network attack classification in IoT using support vector machines," *Journal of Sensor and Actuator Networks*, vol. 10, no. 3, Aug. 2021, doi: 10.3390/jsan10030058.

[2] H. Zolfi, H. Ghorbani, and M. H. Ahmadzadegan, "Investigation and classification of cyber-crimes through IDS and SVM algorithm," in *2019 Third International conference on I-SMAC (IoT in Social, Mobile, Analytics and Cloud) (I-SMAC)*, Dec. 2019, pp. 180–187, doi: 10.1109/I-SMAC47947.2019.9032536.

[3] X. Xie, Y. Fu, H. Jin, Y. Zhao, and W. Cao, "A novel text mining approach for scholar information extraction from web content in Chinese," *Future Generation Computer Systems*, vol. 111, pp. 859–872, Oct. 2020, doi: 10.1016/j.future.2019.08.033.

[4] Q. Liu, M. Jia, and D. Xia, "Dynamic evaluation of new energy vehicle policy based on text mining of PMC knowledge framework," *Journal of Cleaner Production*, vol. 392, Mar. 2023, doi: 10.1016/j.jclepro.2023.136237.

[5] M. Nisha and J. Jebathangam, "Detection and classification of cyberbullying in social media using text mining," in *2022 6th International Conference on Electronics, Communication and Aerospace Technology*, Dec. 2022, pp. 856–861, doi: 10.1109/ICECA55336.2022.10009445.

[6] N. I. Widiastuti, "Convolution neural network for text mining and natural language processing," *IOP Conference Series: Materials Science and Engineering*, vol. 662, no. 5, Nov. 2019, doi: 10.1088/1757-899X/662/5/052010.

[7] M. P. Bach, Ž. Krstić, S. Seljan, and L. Turulja, "Text mining for big data analysis in financial sector: a literature review," *Sustainability*, vol. 11, no. 5, Feb. 2019, doi: 10.3390/su11051277.

[8] A. Shankar, A. K. Tiwari, and M. Gupta, "Sustainable mobile banking application: a text mining approach to explore critical success factors," *Journal of Enterprise Information Management*, vol. 35, no. 2, pp. 414–428, Mar. 2022, doi: 10.1108/JEIM-10-2020-0426.

[9] D. K. Bustami and S. Noviaristanti, "Service quality analysis of tokopedia application using text mining method," *International Journal of Management, Finance and Accounting*, vol. 3, no. 1, pp. 1–21, Feb. 2022, doi: 10.33093/ijomfa.2022.3.1.1.

[10] Amarudin, R. Ferdiana, and Widyawan, "A systematic literature review of intrusion detection system for network security: research trends, datasets and methods," in *2020 4th International Conference on Informatics and Computational Sciences (ICICoS)*, Nov. 2020, pp. 1–6, doi: 10.1109/ICICoS51170.2020.9299068.

[11] A. A. Salih and A. M. Abdulazeez, "Evaluation of classification algorithms for intrusion detection system: a review," *Journal of Soft Computing and Data Mining*, vol. 2, no. 1, pp. 31–40, 2021.

[12] H. Hassani, C. Beneki, S. Unger, M. T. Mazinani, and M. R. Yeganegi, "Text mining in big data analytics," *Big Data and Cognitive Computing*, vol. 4, no. 1, Jan. 2020, doi: 10.3390/bdcc4010001.

[13] P. Devan and N. Khare, "An efficient XGBoost–DNN-based classification model for network intrusion detection system," *Neural Computing and Applications*, vol. 32, no. 16, pp. 12499–12514, Aug. 2020, doi: 10.1007/s00521-020-04708-x.

[14] M. Stamp, *Introduction to Machine Learning with Applications in Information Security*. Boca Raton: Chapman and Hall/CRC, 2022.

[15] L. Ignaczak, G. Goldschmidt, C. A. Da Costa, and R. D. R. Righi, "Text mining in cybersecurity," *ACM Computing Surveys*, vol. 54, no. 7, pp. 1–36, Sep. 2022, doi: 10.1145/3462477.

[16] A. Montoyo, P. Martínez-Barco, and A. Balahur, "Subjectivity and sentiment analysis: an overview of the current state of the area and envisaged developments," *Decision Support Systems*, vol. 53, no. 4, pp. 675–679, Nov. 2012, doi: 10.1016/j.dss.2012.05.022.

[17] A. Aninditya, M. A. Hasibuan, and E. Sutoyo, "Text mining approach using TF-IDF and Naive Bayes for classification of exam questions based on cognitive level of bloom's taxonomy," in *2019 IEEE International Conference on Internet of Things and Intelligence System (IoTaIS)*, Nov. 2019, pp. 112–117, doi: 10.1109/IoTaIS47347.2019.8980428.

[18] S. M. C. Loureiro, J. Guerreiro, and F. Ali, "20 years of research on virtual reality and augmented reality in tourism context: a text-mining approach," *Tourism Management*, vol. 77, 2020, doi: 10.1016/j.tourman.2019.104028.

[19] S. Kumar, A. K. Kar, and P. V. Ilavarasan, "Applications of text mining in services management: a systematic literature review," *International Journal of Information Management Data Insights*, vol. 1, no. 1, 2021, doi: 10.1016/j.jjimei.2021.100008.

[20] L. Rosliadewi, "Analysis of transaction data for modeling the pattern of goods purchase supporting goods location," *Journal of Applied Data Sciences*, vol. 1, no. 2, pp. 65–75, Dec. 2020, doi: 10.47738/jads.v1i2.54.

[21] W. Medhat, A. Hassan, and H. Korashy, "Sentiment analysis algorithms and applications: a survey," *Ain Shams Engineering Journal*, vol. 5, no. 4, pp. 1093–1113, Dec. 2014, doi: 10.1016/j.asej.2014.04.011.

[22] A. Karami, M. Lundy, F. Webb, and Y. K. Dwivedi, "Twitter and research: a systematic literature review through text mining," *IEEE access*, vol. 8, pp. 67698–67717, 2020, doi: 10.1109/ACCESS.2020.2983656.

[23] C. Saranya and G. Manikandan, "A study on normalization techniques for privacy preserving data mining," *International Journal of Engineering and Technology (IJET)*, vol. 5, no. 3, pp. 2701–2704, 2013.

[24] M. Bibi *et al.*, "A novel unsupervised ensemble framework using concept-based linguistic methods and machine learning for twitter sentiment analysis," *Pattern Recognition Letters*, vol. 158, pp. 80–86, Jun. 2022, doi: 10.1016/j.patrec.2022.04.004.

[25] R. Santhosh and M. Mohanapriya, "Generalized fuzzy logic based performance prediction in data mining," *Materials Today:*

Proceedings, vol. 45, pp. 1770–1774, 2021, doi: 10.1016/j.matpr.2020.08.626.

[26] G. Ramaswami, T. Susnjak, A. Mathrani, J. Lim, and P. Garcia, "Using educational data mining techniques to increase the prediction accuracy of student academic performance," *Information and Learning Sciences*, vol. 120, no. 7/8, pp. 451–467, Jul. 2019, doi: 10.1108/ILS-03-2019-0017.

[27] Z. Qiu, Q. Liu, X. Li, J. Zhang, and Y. Zhang, "Construction and analysis of a coal mine accident causation network based on text mining," *Process Safety and Environmental Protection*, vol. 153, pp. 320–328, Sep. 2021, doi: 10.1016/j.psep.2021.07.032.

BIOGRAPHIES OF AUTHORS

Tri Wahyuningsih is a doctoral student of computer science program at Satya Wacana Christian University. She has a strong interest in information systems management and decided to pursue her doctoral degree in computer science. She has interests and capabilities in data mining and text mining. Before starting her doctoral program, Tri Wahyuningsih completed her bachelor and master degrees in informatics engineering. She has working experience as a system analyst and showed excellent achievements during her work. Now, she is focusing on her doctoral studies and working on several research projects in the field of information systems management. She can be contacted at email: 982022001@student.uksw.edu.

Irwan Sembiring earned his bachelor of engineering degree in informatics engineering from Universitas Pembangunan Nasional "Veteran" Yogyakarta in 2001, master of computer science in computer science from Gadjah Mada University Yogyakarta in 2004, and doctorate in computer science from Gadjah Mada University Yogyakarta in 2015. His main research interests are computer network security, and he has done more than 40 publications during his education and teaching career. His research interests include computer network security and computer network designing. Currently, he is active as a lecturer at the Faculty of Information Technology, Satya Wacana Christian University Salatiga. He can be contacted at email: irwan@uksw.edu.

Adi Setiawan is a leading scientist in mathematics and statistics. He earned his bachelor's degree from Gadjah Mada University in 1991 and continued his master's and doctoral studies at the Vrije Universiteit Amsterdam, the Netherlands, in 1997 and 2007. Since 1992, he has been a lecturer in statistics at the Department of Mathematics, Faculty of Science and Mathematics, Satya Wacana Christian University. In the course of his career, Adi Setiawan has held various managerial positions, including as head of the S1 Mathematics Study Programme from 2009 to 2013, and currently serves as dean of the Faculty of Science and Mathematics since 2017. With his expertise in statistics, applied mathematics, and data science, Adi Setiawan has made valuable contributions in advancing science in Indonesia and has been an inspiration to many students and colleagues. He can be contacted at email: adi.setiawan@uksw.edu.

Iwan Setyawan is an image and video processing expert who obtained his bachelor's degree (S1) in electrical engineering from Institut Teknologi Bandung (ITB) in 1996. After that, he continued his master's degree at ITB and earned his doctorate (S3) in electrical engineering from Delft University of Technology, the Netherlands, in 2004. His specialties include image and video processing, watermarking, and image and video-based pattern recognition. Currently, he is a leading expert in his field, making important contributions in industry and research and being a source of inspiration for many in Indonesia. He can be contacted at email: iwan.setyawan@uksw.edu.