OPEN ACCESS
ISSN 2460-9056
http://socj.telkomuniversity.ac.id/ijoict/

Intl. Journal on ICT
Vol. 9, No. 1, Jun 2023. pp. 1-10
doi:10.21108/ijoict.2018.42.203

Machine Learning Sentiment Analysis in Cyber Threat Intelligence Recommendation System

Marastika wicaksono aji bawono [1], Stevani Dwi Utomo [2], Sachlany Kasman [3*]

[1] *Telkom Unviersity* , [2] *Universitas Kristen Duta Wacana*
[1] *Jl. Telekomunikasi No.1 Terusan Buah Batu, Bandung, Jawa Barat, Indonesia,* [2] *Jl. Dr. Wahidin Sudirohusodo No.5-25, Kotabaru, Kec. Gondokusuman, Kota Yogyakarta, Daerah Istimewa Yogyakarta*

[3]*Swiss German University*
[3]*The Prominence Tower Alam Sutera, Jl. Jalur Sutera Bar. No.Kav 15, RT.003/RW.006, Panunggangan Tim., Kec. Pinang, Kota Tangerang, Banten*

[1]marastika@telkomuniversity.ac.id, [2] Stevani.utomo@ti.ukdw.ac.id, [3]sachlany.kasman@student.sgu.ac.id

Abstract
This research focuses on the problem identification of common cyber threats, particularly viruses, using sentiment analysis based on machine learning. The challenge of analyzing unstructured data, such as articles and technical reports, to identify and categorize cybercrime attacks. The novelty of this research is creating a trend dashboard about news trends using a Cyber Threat Intelligence (CTI) engine approach to identify cyber threats. This research method uses cosine similarity to search for news articles by matching them to cybercrime incidents that frequently occur and leverage AI techniques such as TF-IDF and Bag of Words to extract relevant information from CTI documents. Our study highlights the importance of this approach in improving cybersecurity. The findings of this research are that there is an increasing trend in news of cybercrime incidents in Indonesia with the type of Trojan virus with a cosine similarity of 73.41 according to data processing using the BSSN (National Cyber and Crypto Agency) table in Indonesia from 2019 to 2021, frequent incidents were found. appears in the type of virus trojan-downloader: win32small and heur:trojan win32.generik.

Keywords: Natural Language Processing, Text Mining, Cybersecurity,Threat Detection,Data Crawling, Sentiment Classification

I. PENDAHULUAN

The data from January 2021 shows that internet usage increased to 4.66 billion, which is 69.5% of the population in Indonesia. Social media usage also increased by 4.20% among the 53.6% of Indonesia's population, and it continues to grow in the current digital era [2]. According to the BSSN Report, the threat of Trojan attacks continues to rise in Indonesia, with the following trends: in 2019, there were 26,460,689 incidents; in 2020, there were 50,320,126 incidents; and most recently, in 2021, there were 35,286,819 incidents [3]. The more people use the internet, the greater the risk of cyber attacks, making it important to develop tools to protect oneself from cyber attacks.This background information is the basis for conducting research to develop a model for monitoring cybercrime news updates and seeking risk mitigation solutions. This motivates

Received on 14 Aug 2020. Revised on 1 July 2020. Accepted and Published on 2 July 2020.

us to conduct research to provide some information to the public or those who do not understand the threats of cybercrime attacks. The purpose of this research is to provide practical impacts and can be implemented to benefit the community, organizations, or individuals for preventive actions against cybercrime regarding viruses that often attack in a country and how to overcome them so as not to become victims of cybercrime by categorizing tactics, techniques, and procedures (TTP) collected from cybersecurity news, which represents significant threats that describe behaviors and attack patterns against enemies in cybercrime [4]. Cyber Threat Intelligence (CTI) refers to the collection, analysis, and dissemination of information regarding cyber threats and vulnerabilities. The primary purpose of CTI is to help organizations understand and mitigate risks associated with cyberattacks. This involves analyzing data about threats and vulnerabilities to provide actionable and relevant intelligence that can help organizations defend against current and future cyber threats. CTI can include information about malware, threat actors, attack vectors, and vulnerabilities, and it is used to enhance an organization's cybersecurity posture through informed decision-making and strategic planning.

We dissect data from articles, online news, and blogs concerning cybercrime activities to study and discover problem-solving methods. There are several cybersecurity threats, including malware viruses, worms, ransomware, adware, spyware, and trojans. Malware will be classified as a threat to prevent attacks [5]. Our contribution from this research is to develop a machine learning-based framework that uses Cyber Threat Intelligence (CTI) documents as input. Cyber Threat Intelligence (CTI) serves as a solution to mitigate risks by understanding adversaries [21]. CTI strategies involve risk impact analysis and decision support in the form of high-level information [23]. The research problem is to identify frequently occurring cybercrime threats. The latest news updates on cybersecurity in society and the community. Due to the need to gather news data on cybersecurity threats and then analyze them using the tactics, techniques, and procedures (TTP) identification process, it can be concluded that the research problem formulation is: Lack of language and analytic standards, automated information about cyber threats primarily focuses on identifying emerging Tactics, Techniques, and Procedures (TTPs) [24].

II. LITETURE REVIEW

A bibliometric literature analysis was conducted on Scopus data to search for research related to cyber threat intelligence from 2021 to 2023, to be visualized in the VOS Viewer application. This has been done to explore relevant journal literature [21].

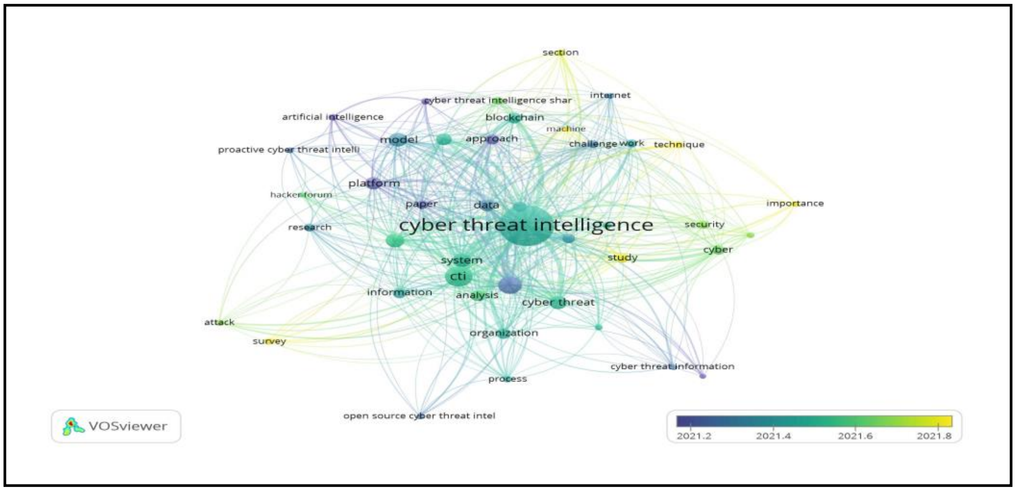

Fig. 1 Analysis of vos viewer literature review

In Figure 1 corresponds to the research findings on data crawling in the Perish of Perish application by extracting Scopus journals from 2020 to 2022 related to cyber threat intelligence. The results of data extraction from all Scopus journals yielded research that can be further developed with the latest techniques aimed at providing recommendations for preventive maintenance steps against security incidents. Correlation methods that are commonly used and have been previously studied include text mining, big data analysis, malware analysis, and network security [21]. This study analyzes applications using machine learning techniques to protect personal data in the virtual world from cybercriminals.

The authors also outline various challenges encountered during the implementation of machine learning. They conclude that machine learning techniques expand various ways to protect the virtual world from cybercriminals. However, there is still much progress needed to safeguard classifiers from cybercrime attacks [9]. Based on Table I, this is previous research to look for methods and problem gaps from previous research

TABLE I
"RELATED PREVIOUS RESEARCH LITERATURE REVIEW"

Journal	Insight	Method	Contribution
[29]	The paper discusses the use of topic modeling algorithms for cyber threat intelligence applications, specifically in the Open Web Application Security Project (OWASP) Maryam framework.	BERTopic and Top2Vec approach	- Proposed idea of utilizing topic modeling approaches for cyber threat intelligence (CTI) applications - Implemented BERTopic and Top2Vec approaches for topic modeling in the Open Web Application Security Project (OWASP) Maryam: Open-Source Intelligence (OSINT) framework
[31]	The paper discusses the taxonomy of cyber threat intelligence frameworks and their components.	- Pyramid of Pain, MITRE ATT&CK framework, Cyber Kill Chain, and The Diamond Model of Intrusion Analysis are used as examples of cyber security frameworks. - Researchers have developed a cyber threat intelligence framework.	- Review of cyber threat intelligence framework - Identification of three main components of the framework.
[30]	The paper discusses the aggregation, cleaning, processing, management, validation, and analysis of cyber threat intelligence for the GTO project.	- Aggregation, cleaning, processing, management, validation, and analysis - Gathering, processing, managing, and analyzing cyber threat intelligence	- Gathering, processing, managing, and analyzing cyber threat intelligence - Providing predictions and empowering cyber threat defenders
[32]	The paper evaluates different cyber threat intelligence maturity models based on design principles and recommends the CTIM model developed by TUDelft University and Cyber Threat Intelligence Lab.	Survey on the significant impact of CTI on their ability to detect, respond to, and prevent cyber threats	- Comparison of publicly accessible CTI maturity models - Identification of the most comprehensive CTI maturity model
[33]	TTP Hunter is an automated tool that extracts Tactics, Techniques, and Procedures	- TTPHunter uses BERT embeddings and linear classifiers.	- Proposed TTPHunter for automated extraction of TTPs

	(TTPs) from APT reports with an F1-score of 88%.	- TTPHunter outperforms rcATT and AttacKG baseline models. techniques, and procedures (TTPs) from e-commerce cyber intelligence datasets.	- Achieved F1-score of 88% and 75% on two datasets
[34]	The paper discusses the definition and modeling of tactics, techniques, and procedures in cyberspace operations, but does not specifically address E-commerce cyber threat intelligence.	- TTPs in relation to cybersecurity - Development of ML algorithms for predictive analytic	.- Proposed model is suitable for describing real-world operations
[37]	Research on automated approaches in cyber risk management and works on predictive attack algorithms and threat hunting	Automated incident response using simple heuristics and ombination of static base defense with adaptive incidence response	Development of methods for efficient automatic incident respons and combination of static base defense with adaptive incidence response for generating a bio-inspired artificial immune system for computerized network
[38]	The paper proposes a data analytic approach using adversary playbooks of tactics, techniques, and procedures (TTPs) for cyber threat intelligence.	- Data analytic approach using association rule mining - Weighted Jaccard similarity for attack attribution	- Data analytic approach for threat attribution - Extending known threat playbooks with probable
[39]	The paper proposes a methodology to extract features from unstructured cyber threat intelligence reports, including tactics, techniques, and tools used by cyber threat actors.	- Natural language processing (NLP) techniques - Machine learning algorithms (decision tree, random forest, support vector machine)	- Development of a mechanism to attribute cyber threat actors (CTA) - Use of natural language processing (NLP) techniques and machine learning algorithms for feature extraction and classification of CTA.
[40]	The paper combines cyber threat intelligence with methods for incident response and defense techniques to enhance cyber security.	- Automated incident response using simple heuristics - Combination of static base defense with adaptive incidence response	- Development of methods for efficient automatic incident response - Combination of static base defense with adaptive incidence response for generating a bio-inspired artificial immune system for computerized networks
[41]	The paper discusses the progress made in applying artificial intelligence techniques to refine the flow of threat intelligence.	- Artificial intelligence techniques for refining threat intelligence - Integration of learning models with firewalls, rules, and heuristics	- Applying artificial intelligence techniques to refine threat intelligence

- Developing learning models integrated with firewalls and heuristics

III. Research Method

In this section, we address the question: "How can we use cyber threat intelligence reports to automatically discover ATT&CK methods and techniques and detect emerging new TTPs?" by determining how to perform pre-processing, feature extraction, and selecting the best features, and how we can discover emerging new TTPs using Diamond Model analysis as part of the cyber kill chain. The steps involved in the design process are included in the study conducted for this work, as illustrated in Figure 2.

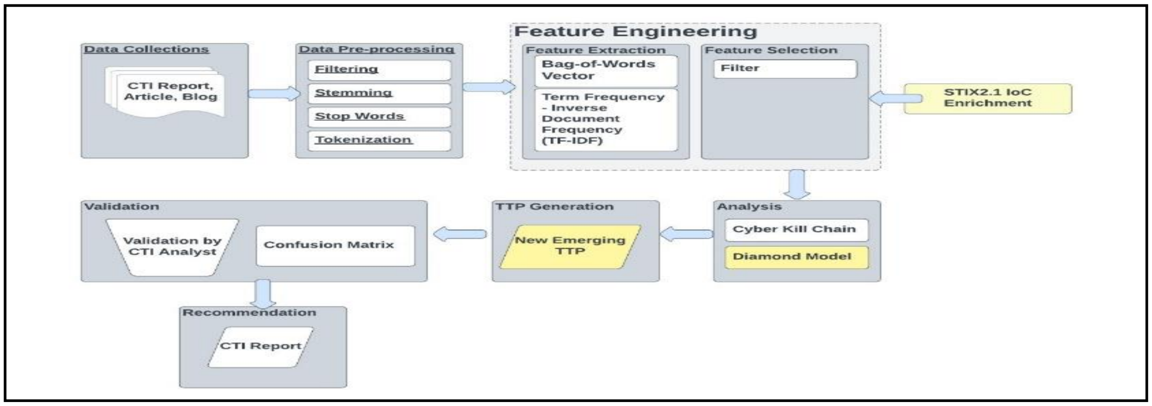

Fig 2. Flowchart of the research process

In Figure 2 exlpain about data source used in this research is gathered from various news websites. A scraping process is conducted to collect all the data. The collected data is then processed in the next stages. The preprocessing stage is the initial step in data processing. There are several preprocessing stages commonly performed in text processing. In this stage, various tasks are carried out, including filtering, stemming, stopwords removal, and tokenization, until the data is prepared for further processing. In the next stage, feature engineering is performed, which is divided into two parts: feature extraction and feature selection. In feature extraction, a bag of words and tf-idf weighting are applied. The Bag of Words is used to represent text documents as a bag containing information that is transformed into vectors that are more comprehensible to computers. It is followed by the TF-IDF algorithm, which assigns weights to each keyword in each category to determine the similarity of keywords to the available categories. This study proposes a framework by analyzing text input from intelligence reports on threats provided, processing it to identify emerging new TTPs, and finally incorporating them into reports as recommendations. We utilize the ATT&CK framework (Adversarial Tactics, Techniques, and Common Knowledge) [19]. Therefore, this framework aims to describe each possible technique and tactic used to prioritize threats and understand the TTPs (Techniques, Tools, and Procedures) of the attack [35].These tactics are explained within the MITRE ATT&CK Framework [7]. The temporal relationships of tactics are their distinguishing feature. Linking reports to tactics and specific techniques within the ATT&CK framework can also be used for faster methods of prevention, detection, and mitigation of threats mentioned in the text [8]. We hope this framework functions effectively, especially in reports that provide further analysis of newly emerging TTPs.

Combining methods from updated natural language processing (NLP), data mining, and machine learning to filter information from large text data sets like news and social media [19].The aim of this research is focused on identifying current updates on cybercrime news using NLP and data mining techniques to extract features that support text classification and approaches to text classification using machine learning techniques and the data used to train it [37].

IV. RESULTS AND DISCUSSION

The dataset was collected based on news about cyber security in Indonesia using data crawling. Subsequently, data analysis using the cosine similarity method was performed to observe the results of actions taken against ground truth data. Once the data was obtained, a comparison was made between the data and the BSSN (National Cyber and Code Agency) reports from 2019 to 2021 regarding the countries most affected by cybercrime viruses and the types of viruses involved.

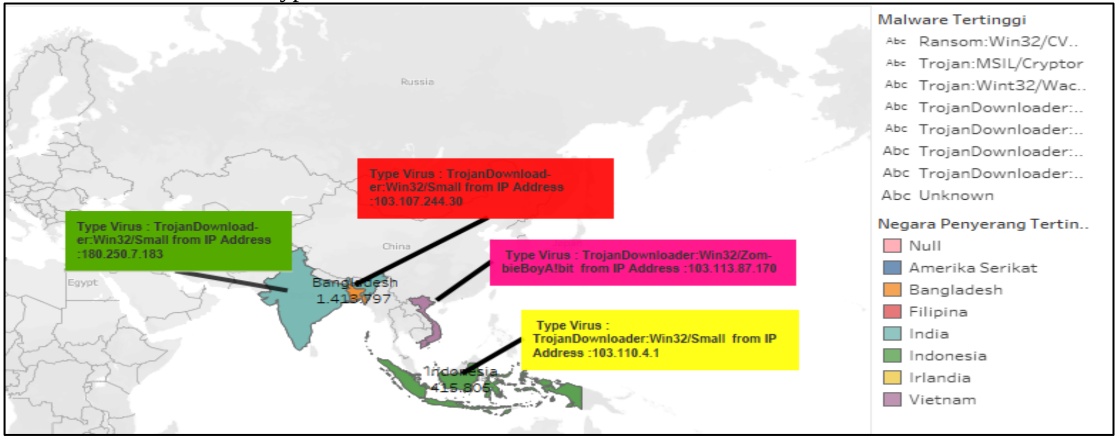
Russia
China
Egypt
Type Virus : TrojanDownloader:Win32/Small from IP Address :103.107.244.30
Type Virus : TrojanDownloader:Win32/Small from IP Address :180.250.7.183
Type Virus : TrojanDownloader:Win32/ZombieBoyA!bit from IP Address :103.113.87.170
Type Virus : TrojanDownloader:Win32/Small from IP Address :103.110.4.1
Bangladesh 1.413.797
Indonesia 415.805
Malware Tertinggi
Abc Ransom:Win32/CV..
Abc Trojan:MSIL/Cryptor
Abc Trojan:Wint32/Wac..
Abc TrojanDownloader:..
Abc TrojanDownloader:..
Abc TrojanDownloader:..
Abc TrojanDownloader:..
Abc Unknown
Negara Penyerang Tertin..
Null
Amerika Serikat
Bangladesh
Filipina
India
Indonesia
Irlandia
Vietnam

Fig 3. Countries hit by the highest cybersecurity threat attacks.

From the chart in Figure 3 in the data analysis below, the countries with the highest number of attacks are India, Indonesia, Bangladesh, and Vietnam. The most frequently occurring virus types are Trojan-Downloader: Win32/Small and HEUR: Trojan .Win32.Generik. The first stage involves analysis and evaluation, which is then visualized using Tableau software, using cyber kill chain indicators and the diamond model to assist in detection, protection, recovery, and system restoration using TTP (Techniques, Tools, and Procedures) methods to mitigate the risk of attacks.

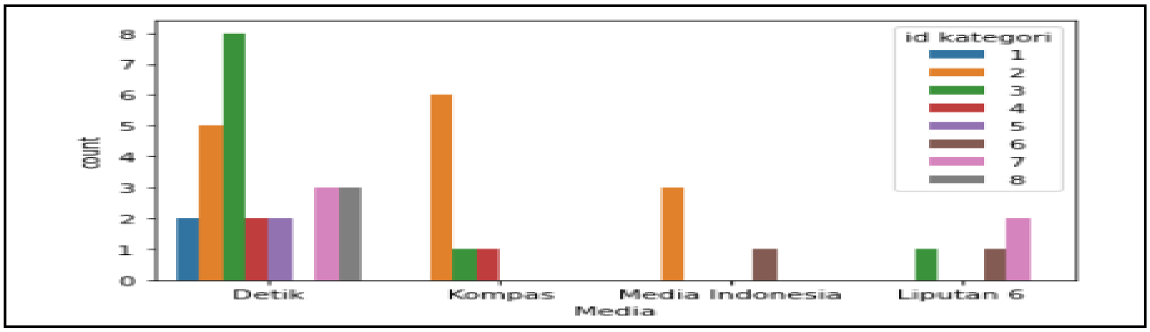

Fig 4. Visualization of media news sources regarding cybercrime threats

In Figure 4 above, it is concluded from various kinds of electronic media that often discuss cybercrime, Detik.com is highlighted due to the increase in Category ID on the trojan virus, which is green in the graph. Therefore, prevention and mitigation can be observed to avoid the trojan virus. The distribution of news sources collected from the Detik, Kompas, Media Indonesia, and Liputan 6 websites, which regularly report on cyber security crimes in cyberspace. We then performed an analysis using cosine similarity on the available sample data. Based on the data above, the following table provides an explanation regarding the category ID below.

TABLE II. COSINES SIMILARITY RESULT

Category	Type	Standard	Cosine Similatiry
Id category 1	Cyber security	1,00	62.92
Id category 2	Hacking	1,00	67.98
Id category 3	Trojan	1,00	73.41
Id category 4	Malware	1,00	56.65
Id category 6	Phishing	1,00	77.18
Id category 7	Ransomware	1,00	71.87
Id category 8	Botnet	1,00	55.94

Based on Table II, it is found that the most frequently occurring incidents are Trojan threats with a cosine similarity of 73.41, which is found in Detik.com news.

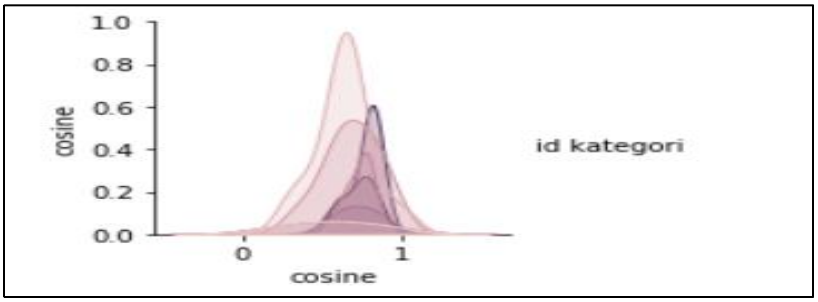

Fig 5. graph of the cosine similarity results for each category

From the Figure 5 above, Cosine Similarity is a measure of similarity between two numerical sequences in data analysis. To define it, consider a sequence as a vector in a dot product space, and cosine similarity is the dot product of the vectors divided by the product of their magnitudes. After conducting the analysis, it is evaluated to determine the comparison score with the ground truth using cosine similarity. Cosine Similarity is used to calculate the cosine angle value between the document vector and the query vector [37]. The smaller the resulting angle, the higher the similarity level of the essays. This evaluation uses the cosine similarity method to assess the level of similarity between the processed results and the ground truth. The samples used in this evaluation consist of several categories, as seen in Figure 6.

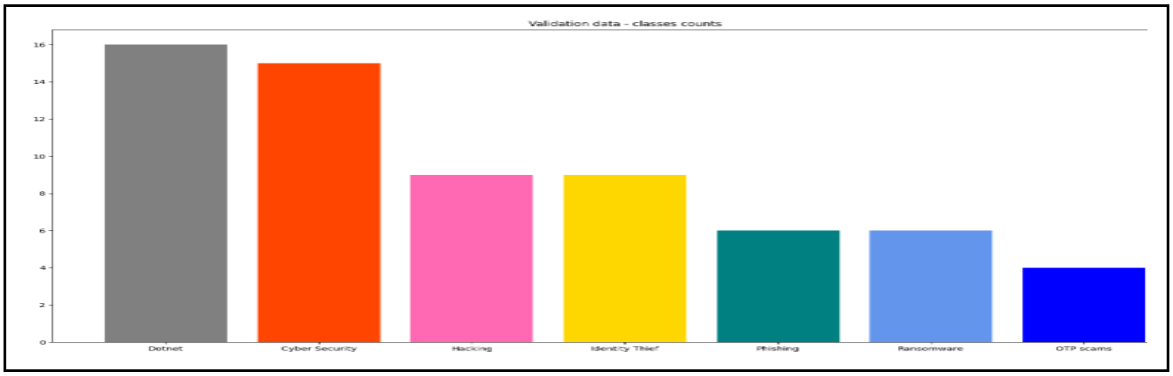

Fig 6. cybersecurity category results

Based on the results in Figure 6, the most prevalent types of cybercrime attacks are dotnet and trojan. Exploitation opportunities use dotnets and trojans are commonly used by cybercriminals because they provide various opportunities for exploitation. Dotnet vulnerabilities can be exploited to compromise systems, while trojans often trick users into downloading malicious files or disguising themselves as legitimate software. Implementing dotnet and trojan attacks can be relatively low-cost compared to other cybercrime methods, making them attractive to threat actors.

V. CONCLUSION

The research findings indicate a rising trend in cybercrime incident reports within Indonesia, particularly pertaining to Trojan viruses, with a cosine similarity of 73.41, according to data processed using the BSSN (National Cyber and Crypto Agency) table from 2019 to 2021. Notably, frequent incidents were identified involving the trojan-downloader: win32/Small and heur :trojan win32/generik virus types. The limitation of this research solely focuses on prevention measures against viruses or cybercrimes in Indonesia to enhance cybersecurity in the country. Future research could discuss other viruses as well as ways to address them using machine learning or cybersecurity.

Here are the steps to prevent being infected by the trojan virus :

- Use Antivirus Software: Ensure you have reliable antivirus software installed and regularly updated. Enable real-time protection to scan and monitor your system continuously.
- Software Updates:Regularly update your operating system and all software to patch vulnerabilities. Ensure that your antivirus definitions are updated to recognize new threats.
- Safe Browsing Habits:Be cautious when downloading files or clicking on links from unknown sources.Avoid visiting suspicious websites.
- Email Safety: Do not open email attachments or click on links from unknown or untrusted sources. Be wary of phishing emails that may appear legitimate.
- Use a Firewall: Employ a firewall to monitor and control incoming and outgoing network traffic.
- Backup Data:Regularly back up your data to recover information in case of a ransomware attack.
- User Education: Educate users about the risks of downloading and executing unknown files from the internet.

How to overcome/eemove trojan virus:

- Safe Mode: Boot your computer in Safe Mode to restrict the Trojan’s operation.
- Scan and Remove: Use your antivirus software to perform a full system scan and remove detected threats. Consider using a specialized malware removal tool for a thorough scan.
- Manual Removal: Identify and delete malicious files manually (this requires technical knowledge and is risky).
- System Restore: Perform a system restore to revert your computer to a state before the infection.
- Professional Help:If the infection is persistent or damaging, seek professional help to remove the malware without risking data.
- Change Passwords: After removal, change all passwords as they might have been compromised.
- Update Software: Ensure all software and systems are updated to prevent re-infection.
- Review Security Practices: Review and enhance security practices to prevent future infections