КІБЕРБЕЗПЕКА: освіта, наука, техніка
CYBERSECURITY: EDUCATION, SCIENCE, TECHNIQUE
№ 2 (26), 2024
ISSN 2663 - 4023

DOI 10.28925/2663-4023.2024.26.694
UDC 004.056

Oleksii Chalyi
Master's Degree in Cyber Security, Master's Degree in Computing,
Laborant at the Institute of Social Sciences and Applied Informatics
Kaunas Faculty Vilnius University, Kaunas, Lithuania
ORCID ID: 0009-0006-3536-9715
oleksii.chalyi@knf.vu.lt

Iryna Stopochkina
PhD, Associate Professor,
Associate Professor at the Institute of Physics and Technology
National Technical University of Ukraine "Igor Sikorsky
Kyiv Polytechnic Institute", Kyiv, Ukraine
ORCID ID: 0000-0002-0346-0390
i.stopochkina@kpi.ua

INFORMATION RETRIEVAL AND DEANONYMIZATION IN THE TASKS OF EARLY DETECTION OF POTENTIAL ATTACKS ON CRITICAL INFRASTRUCTURE

Abstract. Information about cyberattacks that attackers plan to carry out against critical infrastructure facilities is partly distributed on malicious information channels, chats or sites. Investigation of information materials and their analysis can provide an understanding of the stages of attack planning and their prevention. Part of this problem is to provide information search and analysis tools to detect linguistic patterns, similarities in text data, which are capable of deanonymizing cybercriminals and establishing relationships between published data. This work proposes a new model and a corresponding prototype of the system, based on the vector space model and the TF-IDF algorithm. The system is designed to analyze publicly available text data (both internet and darknet), and differs with a probabilistic approach to analyzing the identifiers of the information publisher. The proposed system also focuses on identifying latent connections between anonymous accounts by analyzing unique stylistic and linguistic traits. It leverages these traits to trace patterns in communication, uncovering hidden associations among cybercriminal entities. Experiments conducted based on the analysis of real chats, including chats of cybercriminals, demonstrate the potential of the system for detecting identifiers and determining stylistic features. If a sufficiently complete set of data is available and a list of target words is available, it is possible to analyze the stages of preparing an attack, malicious individuals or groups involved in it. The results underline the significance of integrating advanced linguistic analysis techniques with probabilistic models to enhance investigative capabilities against evolving cyber threats.

Keywords: information retrieval; cybersecurity; deanonymization; vector space model; critical infrastructure; cybercriminal; TF-IDF algorithm.

INTRODUCTION

In recent years, the escalation of cybercrime against critical infrastructure has increased the request for sophisticated methods to track, identify, and combat new threats [1].

At the stage of attack implementation, there are tools that allow establish a pattern of malicious intrusion using information from MITRE [2]. However, a more successful solution is to identify the potential intrusion and its nature based on freshly collected data.

There are various methods of information gathering used by attackers [3] which can also be used to detect information about planned attacks. However, they are not focused on the task of early detection of information about attacks, and require a creative approach from those who use them.

A separate task remains to establish the relationship of texts published by attackers. They often use different nicknames to mask authorship, even when writing in the same information space (the same chat). Establishing joint authorship of posts will allow focusing attention on posts by well-known in cyberspace (although anonymous in physical space) individuals. Analysis of texts of a certain authorship will allow indirectly learning about the plans and intentions of the targeted criminal person or representatives of a criminal group, and identifying new potential threats.

The anonymity has enabled cybercriminals to operate under multiple pseudonymous or additional accounts, which complicates efforts by law enforcement to trace cybercriminals [4]. Identifying the individuals behind these accounts is challenging, especially when the cybercriminal strategically disguises their identities to evade detection [5]. This is why deanonymization, the process of uncovering hidden identities, has become crucial in the field of cybersecurity.

One potential approach to deanonymization lies in analyzing written text for linguistic and stylistic patterns unique to an individual [6]. Information Retrieval (IR) techniques are instrumental in this approach, as they can systematically analyze textual content [7] to establish probabilistic matches between multiple accounts. By detecting stylistic similarities, IR methods can reduce the pool of suspects, enhancing investigative efficiency in cybercrime cases. Despite the promise of IR methods, however, few accessible and reliable deanonymization systems exist within the public domain, leaving a gap in the tools available for practical law enforcement use.

Problem statement. Despite the growing need for effective deanonymization methods in combating cybercrime, particularly for identifying individuals behind pseudonymous or additional accounts, existing Information Retrieval (IR) techniques face limitations. These methods are not specifically tailored for early detection of attack plans or establishing authorship links in cybercriminal texts, and there is a lack of accessible and reliable deanonymization systems for practical use by law enforcement.

ANALYSIS OF RECENT RESEARCH AND PUBLICATIONS

This section provides an analysis of scientific methods for cybercriminal deanonymization. It offers a brief review of developed deanonymization methods, which will be used for comparison with the proposed method based on IR.

Rüdian et al. [8] in their article investigate the confidentiality issues resulting from the use of the “like” system on Facebook. They describe the applied algorithms and methods in their methodology section. They outline the required steps for cybercriminals to use likes for deanonymization purposes:

1. Information Gathering
2. Using applications to identify useful or critical information from gathered data
3. Combining the collected data with datasets
4. Deanonymization

Rüdian et al. indicated that it is important to have a connection between targets and processed datasets. They conducted several experiments, attempting to simulate the attack using two different methods. For the first method, they used the Seorld [9] tool, which can gather information from Facebook. The second method involves collecting browser history. By combining both of these methods, a cybercriminal can deanonymize the target. In their experiments, they collected seven million likes from more than 920,000 Facebook users. If a user reacted with at least four likes, the success rate of their system in deanonymization was approximately 99.91%. This system is a significant example of an existing method for deanonymization using internet technologies, including social media.

Simioni et al., in their article [10], study a deanonymization method on the internet. As a target, they chose the infrastructure and conducted two experiments to monitor it. In the first experiment, they estimated the access (the quantity of data that could be collected) from their infrastructure over several days. In the second experiment, they attempted to deanonymize a hidden node using the collected data. After analyzing the data from the information-gathering phase, they achieved visibility of approximately 15% of the total nodes in the network. In their conclusions, they noted the complexity of their method and its dependence on the infrastructure. They acknowledged that the results could be unsuccessful, which they aim to address in future work.

Boldyrikhin et al., in their article, describe their deanonymization method using correlation analysis [11]. First, they collected a database of their targets, gathered from different sources. Second, they estimated the correlation coefficient for all data in the selected database. Finally, they compared the target with other users. If the correlation coefficient was high, the user could be deanonymized. This article provides an example of using comparison methods to succeed in deanonymization tasks.

Beato et al. developed a deanonymization method called “Friend in the Middle”, using complex mathematical formulas and algorithms [12]. By “Friend”, they refer to common contacts in social media. They conducted an experiment in which a cybercriminal has several friends and uses one of them as the middle point. As a result, they can create a dataset with nodes representing users and links representing social media connections. This dataset can be used to reverse social media contacts and perform deanonymization.

Peng et al. developed a two-step deanonymization attack on social media users called “Seed-and-Grow”, based on graph methods [13]. In their experiment, they first collected two datasets from different social media platforms. Using these datasets, they created a graph with nodes and links. They assumed that the cybercriminal also had their own graph with nodes and links. The first step, called “Seed”, involved developing an algorithm to define the primary subgraph, which could either be added by the cybercriminal or discovered by users. The second step, “Grow”, expanded the primary subgraph based on the cybercriminal's knowledge of the network. In their conclusions, Peng et al. highlighted the potential functionality of their method and its success in deanonymization.

Hongu et al. describe a deanonymization method based on spectral graph division [14]. First, they collected three datasets from Facebook, Twitter, and Google Plus social media platforms. They developed a two-step algorithm: in the first step, they applied a spectral method for graph division to divide the graph into smaller subgraphs. In the second step, they used their algorithm to process these subgraphs, identifying structural and attribute similarities. The results demonstrate the success of the experiment and high precision in deanonymization.

Our approach differs by its targeting not only on cybercrime deanonymization, but also by selecting the texts with the relevant information concerning certain type of attacks on critical infrastructure.

PURPOSE OF THE RESEARCH

This research aims to bridge the gap in the tools available for practical law enforcement use by designing and evaluating a prototype system for deanonymization, based on IR methodologies. The study addresses key objectives: it evaluates existing deanonymization approaches, compares popular methods, and proposes a functional prototype for real-world testing. Additionally, the research outlines the algorithmic content, including text preprocessing, tokenization, and reverse indexing, which forms the foundation of the proposed system. Through experimental analysis, the study assesses the performance of the prototype and its practical applicability, offering insights into the feasibility of using IR technologies for deanonymization tasks.

The practical significance of this research could be significant for organizations and law enforcement agencies and state services, which monitor the new attack patterns for critical infrastructure. By narrowing down suspects based on linguistic evidence, the proposed system offers a novel tool that may streamline the process of identifying cybercriminals. This study provides an evidence-based assessment of IR's suitability for deanonymization and paves the way for further innovation in the field.

METHODS

Development of a methodology for deanonymizing a cybercriminal using Information Retrieval

Before developing the system, a methodology was created, defining six stages necessary for the successful and effective use of the system:

- Planning Stage: Sets the target and goal, including criteria and resources. The criteria may include attack features and critical infrastructure peculiarities. Resources may include the list of cybercrime chats and media to be analyzed.
- Preparing Stage: Defines the wordlists, environment, and other initial programming stages.
- System Developing Stage: Outlines the required steps for creating the deanonymization system using IR technologies.
- Setting-up Stage: Describes the debugging steps.
- Exploitation Stage: Applies the deanonymization system.
- Result Analysis Stage: Draws conclusions about the deanonymization process and its success.

The following factors can be used as a basis for the search:

- S — scope of attack that interests researchers (hardware, OS type, edge device, software, network devices).
- H — the need to involve the human factor.
- V — zero day vulnerabilities, known vulnerabilities.
- F — features of the exploit (LPE, ACE, RCE).
- P — which manufacturer of controllers, IoT devices or other hardware equipment is the attack targeted at.

It is also necessary to specify the search source — public sources, well-known cybersecurity articles, cybersecurity forums, as well as channels and chats with a criminal focus to identify posts that may lead to understanding the essence of a new attack on critical infrastructure.

When searching for relevant information on critical infrastructure objects of a certain focus — it is necessary to specify the above data in order to draw conclusions about the current state of the exploit market and trends in the field of organized cybercrime.

The set of texts of a cybercrime person (or cybercrime group) can be further analyzed for the factors of interest to the researcher (S, H, V, F, P). Text analysis can be done by tokenization and detection of common words.

On the other hand, each of the factors can be represented by a set of specific values {Si, Hi, Vi, Fi, Pi}, which narrows the search area.

A proposed methodology of searching certain cybercrime messages, texts, and possible deanonymization of cybercrime, is given in Fig. 1.

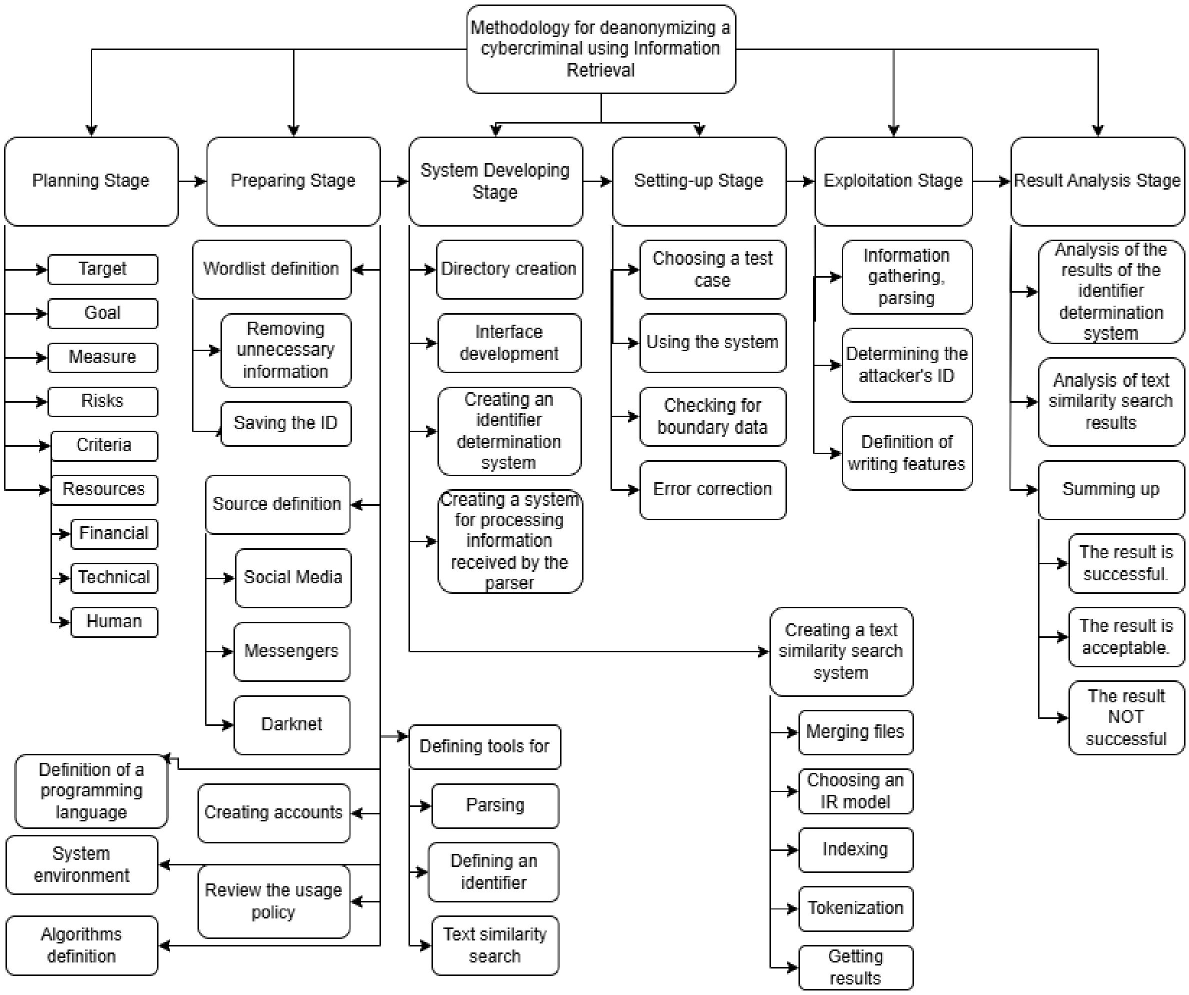

Methodology for deanonymizing a cybercriminal using Information Retrieval
Planning Stage
Preparing Stage
System Developing Stage
Setting-up Stage
Exploitation Stage
Result Analysis Stage
Target
Goal
Measure
Risks
Criteria
Resources
Financial
Technical
Human
Wordlist definition
Removing unnecessary information
Saving the ID
Source definition
Social Media
Messengers
Darknet
Definition of a programming language
Creating accounts
System environment
Review the usage policy
Algorithms definition
Directory creation
Interface development
Creating an identifier determination system
Creating a system for processing information received by the parser
Defining tools for
Parsing
Defining an identifier
Text similarity search
Choosing a test case
Using the system
Checking for boundary data
Error correction
Information gathering, parsing
Determining the attacker's ID
Definition of writing features
Creating a text similarity search system
Merging files
Choosing an IR model
Indexing
Tokenization
Getting results
Analysis of the results of the identifier determination system
Analysis of text similarity search results
Summing up
The result is successful.
The result is acceptable.
The result NOT successful

Fig. 1. Proposed methodology for deanonymizing a cybercriminal using IR

Development of an architecture for deanonymizing a cybercriminal using Information Retrieval

Based on the Methodology from the previous section, the architecture of the proposed system was created, which is shown in Fig. 2.

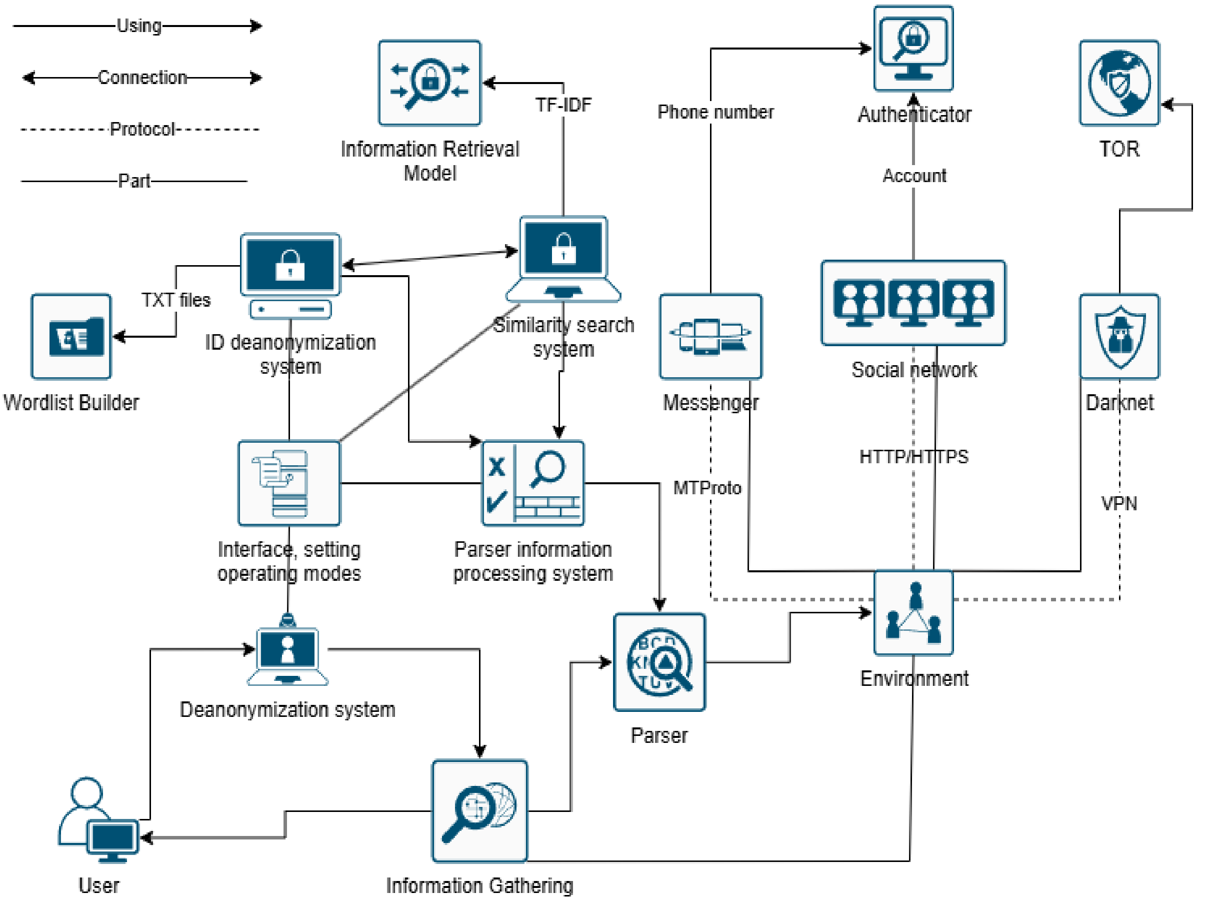


Fig. 2. Proposed architecture of the proposed cybercriminal deanonymization system

This architecture presents the main components of the system, as described in the methodology. It also illustrates the different protocols required for functionality. For example, social networks like Facebook use the HTTPS protocol, which is depicted in Fig. 2. Additionally, the figure shows the usage and connections between different parts of the system.

The ID deanonymization system and the Similarity Search system are represented as interconnected because they exchange data obtained after algorithm processing. The “using” lines indicate which modules are utilized by specific systems. For instance, the ID deanonymization system uses the wordlist builder, as it creates the required dataset for the system’s functionality. Data transfer operates through specialized text files.

Simple lines represent components of the system. For example, the architecture shows that the environment consists of the following parts:

1. Social Networks.
2. Messengers.
3. Darknet.

Using this data, the algorithm diagram of the Similarity Search system was created and is represented in Fig. 3.

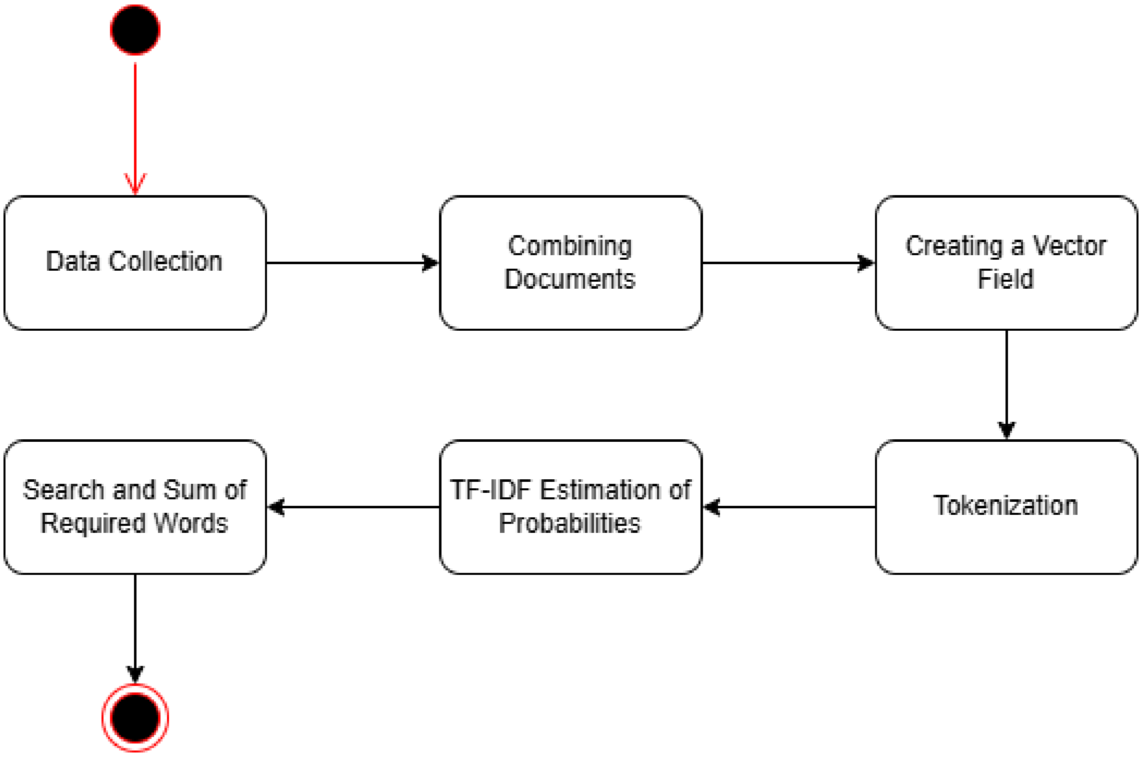


Fig. 3. Similarity search system operation algorithm

Fig. 3 illustrates the sequence of the similarity search system, which also can be used for the following deanonymization. The similarity search system plays a crucial role in determining the final results. Defining the functional algorithm for probability calculation is essential.

The proposed system uses the TF-IDF algorithm, which is well-regarded for generating relevant results and is commonly employed by Google [15]. The vector model was utilized as the IR model, as it offers a semantic approach, considering word meanings and context for a better understanding of user queries. It supports ranking through techniques like TF-IDF. Nevertheless, it faces challenges related to word order and the inclusion of irrelevant terms in the vector representation [16]. As a vector, it consists of one component called “document”, which is represented as user text files in this research. The proposed system also adheres to the fundamental principles of IR processes, including indexing, tokenization, and preprocessing [7].

As a result, the system can process any combination of user words or word sequences and calculate probabilities according to the defined objectives. All possible word combinations are considered, meaning their order is taken into account, and the system processes all word combinations.

To enhance the interface, all fields containing useless data, such as zero-value columns, were removed.

Fig. 4 presents the diagram for the proposed ID deanonymization system algorithm.

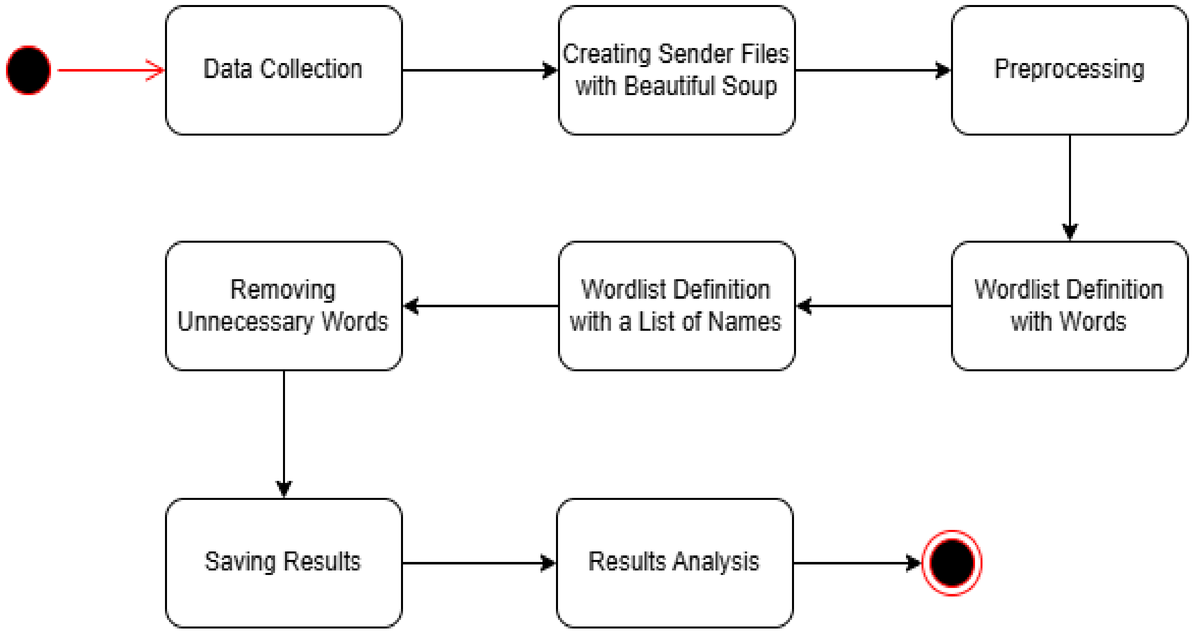

Data Collection
Creating Sender Files with Beautiful Soup
Preprocessing
Wordlist Definition with Words
Wordlist Definition with a List of Names
Removing Unnecessary Words
Saving Results
Results Analysis

Fig. 4. ID deanonymization system operation algorithm

This figure illustrates the sequence of steps required for ID deanonymization. The system operates after the parser has collected all data and before the similarity search system. It is proposed to use the Beautiful Soup [17] tool for processing the collected data due to its accessibility and speed.

The wordlist definition step is crucial, as it helps remove unnecessary words. In this research, experiments were conducted in Ukraine using Ukrainian chats. For languages with a limited number of available wordlists on the internet, it is proposed to include an additional step in this system. This step ensures the preservation of critical information from accidental removal.

The process of user text selection is described by Formula 1

$$R = \bigcup_{j=1}^{n} (T_j \cap W_d) \qquad (1)$$

where R is the result stored in the directory, T_j is user text, j is user ID, n is the number of users parsed, W_d is text containing necessary IDs, where ID may be specification of the user name, and factors (S_i, H_i, V_i, F_i, P_i). Union in (1) means processing for each user.

As a result, text files (*.txt) are generated for each user. These files contain the user nicknames (and also factors S_i, H_i, V_i, F_i, P_i if we search the relevant information by known factors among user messages).

Another case is related with ID selection in the user texts. It can be made using the regular expressions, e.g:

```
   @\w+     # Matches @username;
   #\w+     # Matches #tag;
   [a-zA-Z0-9._%+-]+@[a-zA-Z0-9.-]+\.[a-zA-Z]{2,}    # Matches email formats;
   \b[a-zA-Z0-9._-]{3,20}\b     # Matches words with lengths suitable for nicknames
(3-20 chars).
```

Then it is required to remove duplicates, filter out common words or false positives (e.g., “user”, “admin”) by comparing against a stoplist of excluded terms.

Next step is to optionally validate results against a predefined format or rules for platform (e.g., minimum/maximum length).

Technical features of the system

As the programming language for the system, Python [18] was used due to its functionality, and integration with Linux distributions. Table 1 shows the libraries used in the system and their respective functionalities.

Table 1

Python libraries involved in the developed system

Python Library	Functionality
BeautifulSoup [16]	Analysis of documents obtained by parsing
TfidfVectorizer [19]	TF-IDF algorithm using
Pandas	Table creation
Numpy	Working with arrays
os	Working with files
re	Working with regular expression

To check the system, a device with the following specifications in Table 2 was used.

Table 2

Test Device Specifications

Component	Specification
Device	Laptop
CPU	AMD Ryzen 7 4800H
RAM	6 Гб
Virtual Machine	VirtualBox 6.1.46
Operating System	Kali Linux 2023.3

Datasets

The experiments were conducted with the cyrillic-based languages, so special wordlists for removing words were defined [20], [21]. The wordlists should consist of as many words as possible. If the system is used for English, there are many more available wordlists that could be used. The wordlist with identifiers were also used [22].

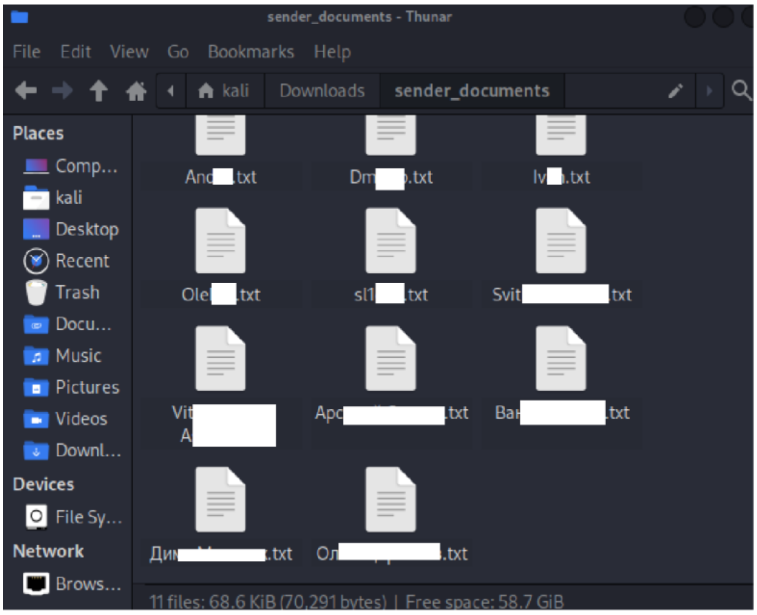
In future it is relevant to use existing dictionaries of hacker jargon [23], [24], and others for target person language.

RESULTS

Two experiments were conducted based on the methodology described in the previous section. Since the experiments were carried out using cyrillic-based cybercriminal chats, the figures in this section contain words in Cyrillic. This method is designed to work with any other language; the only difference lies in the selected wordlists.

Results with system operation in a test chat

The first stage involved information gathering. Telegram Messenger was used as the target for the test chat, specifically a university chat related to cybersecurity. The Telegram export chat function was used to collect all chat data, excluding multimedia. Beautiful Soup was employed to extract the text. A separate *.txt file was created for each user, containing their user ID. Additionally, all punctuation symbols, except for spaces, were removed. The result of data gathering using Beautiful Soup is shown in Fig. 5. To comply with the journal's confidentiality rules, critical information was removed.

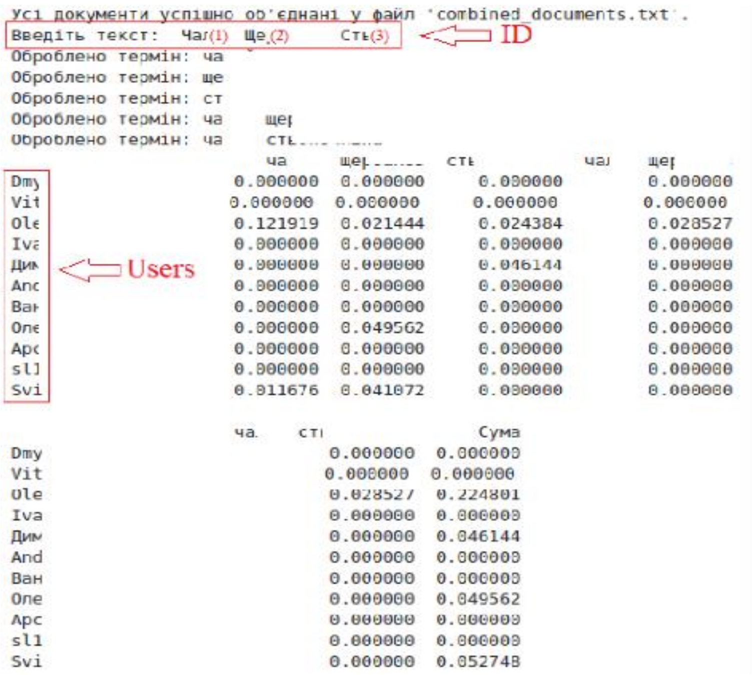


And[redacted].txt
Dm[redacted].txt
Iv[redacted].txt
Ole[redacted].txt
sl1[redacted].txt
Svit[redacted].txt
Vit A[redacted]
Apc[redacted].txt
Ba[redacted].txt
Дим[redacted].txt
Ол[redacted].txt

11 files: 68.6 KiB (70,291 bytes) | Free space: 58.7 GiB

Fig. 5. Information Gathering using Beautiful Soup

As shown in Fig. 5, these files contain useful information, including user IDs. To process these files, it is necessary to use the ID deanonymization system. As described in the methodology, the main purpose of this system is to remove all words except the IDs, which will be used further. After applying the ID deanonymization system a set of files (e.g., Output/Oleksii.txt) containing only the IDs was obtained. The next step is to use the created IR system to find text similarities. Fig. 6 illustrates the results of processing this system.



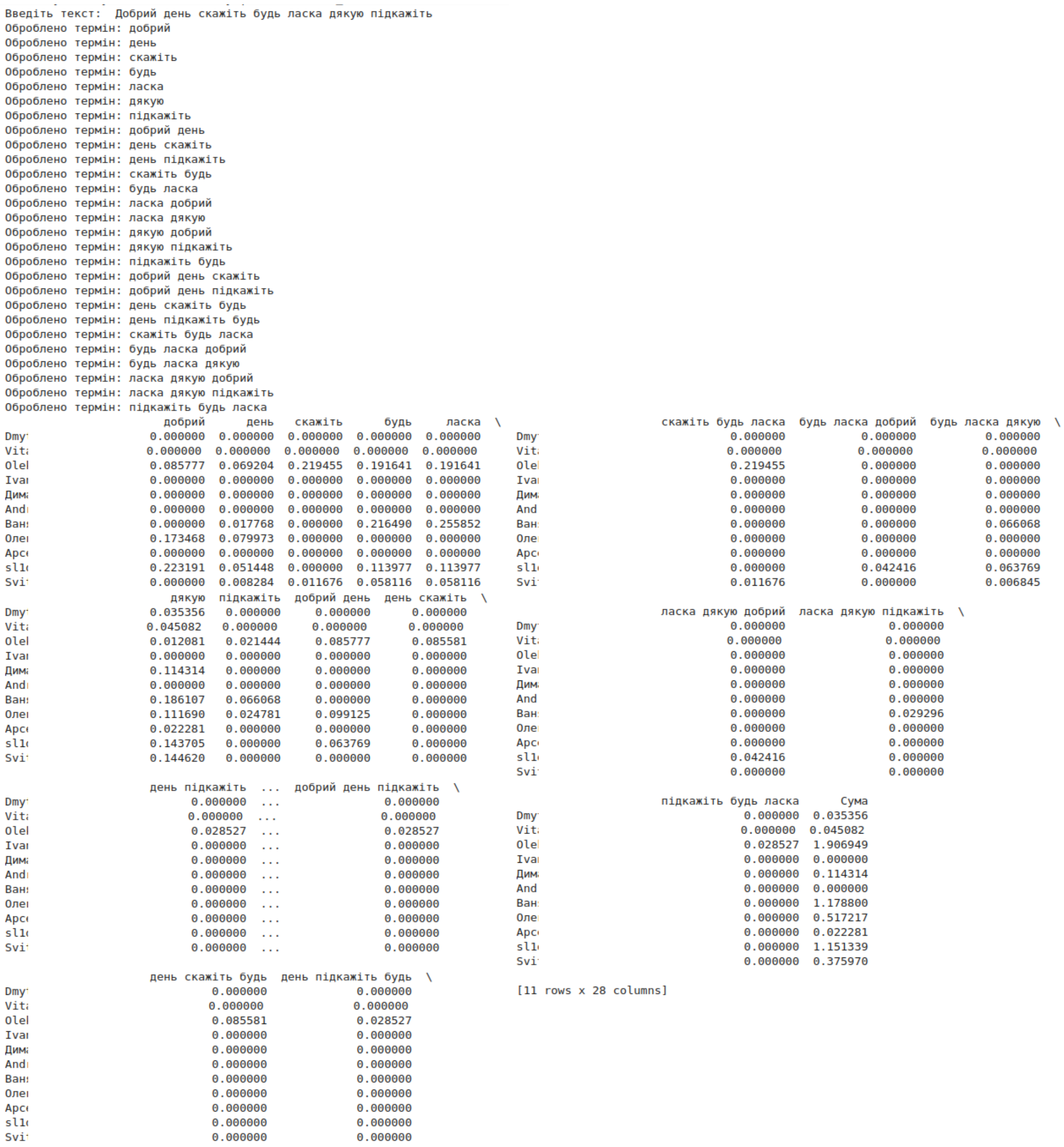
```
Усі документи успішно об'єднані у файл 'combined_documents.txt'.
Введіть текст:   Чаı(1)  Ще(2)      Ств(3)
Оброблено термін: ча
Оброблено термін: ще
Оброблено термін: ст
Оброблено термін: ча     щеı
Оброблено термін: ча     ств
                         ча        щеı        ств              чаı      щеı
Dmy                  0.000000  0.000000    0.000000            0.000000
Vit                 0.000000  0.000000    0.000000            0.000000
Ole                  0.121919  0.021444    0.024384            0.028527
Iva                  0.000000  0.000000    0.000000            0.000000
Дим                  0.000000  0.000000    0.046144            0.000000
Anc                  0.000000  0.000000    0.000000            0.000000
Ваı                  0.000000  0.000000    0.000000            0.000000
One                  0.000000  0.049562    0.000000            0.000000
Apc                  0.000000  0.000000    0.000000            0.000000
sl1                  0.000000  0.000000    0.000000            0.000000
Svi                  0.011676  0.041072    0.000000            0.000000

                     ча.    сті                Сума
Dmy                         0.000000  0.000000
Vit                        0.000000  0.000000
Ole                         0.028527  0.224801
Iva                         0.000000  0.000000
Дим                         0.000000  0.046144
And                         0.000000  0.000000
Ван                         0.000000  0.000000
One                         0.000000  0.049562
Apc                         0.000000  0.000000
sl1                         0.000000  0.000000
Svi                         0.000000  0.052748
```

Fig. 6. Result of using IR

As the figure shows, this system can facilitate user deanonymization and identify additional useful information. In Fig. 6, three IDs were used.

The task was to identify user identifiers based on their Telegram IDs located in sender_documents/Oleksii.txt. The TF-IDF algorithm determined the probabilities, with the file Oleksii.txt showing the highest probability (0.121) for the ID (1), indicating the strongest similarity. The ID (2) demonstrated the highest probability with the user Ole. Regarding ID (3), it displayed probabilities for both users, suggesting a connection between them. One advantage of this system is its ability to identify and document all possible combinations and their associated probabilities. This system can also be applied to analyze the features of user text. Fig. 7 demonstrates the results of using this system to identify text features in a test chat.



```
Введіть текст:  Добрий день скажіть будь ласка дякую підкажіть
Оброблено термін: добрий
Оброблено термін: день
Оброблено термін: скажіть
Оброблено термін: будь
Оброблено термін: ласка
Оброблено термін: дякую
Оброблено термін: підкажіть
Оброблено термін: добрий день
Оброблено термін: день скажіть
Оброблено термін: день підкажіть
Оброблено термін: скажіть будь
Оброблено термін: будь ласка
Оброблено термін: ласка добрий
Оброблено термін: ласка дякую
Оброблено термін: дякую добрий
Оброблено термін: дякую підкажіть
Оброблено термін: підкажіть будь
Оброблено термін: добрий день скажіть
Оброблено термін: добрий день підкажіть
Оброблено термін: день скажіть будь
Оброблено термін: день підкажіть будь
Оброблено термін: скажіть будь ласка
Оброблено термін: будь ласка добрий
Оброблено термін: будь ласка дякую
Оброблено термін: ласка дякую добрий
Оброблено термін: ласка дякую підкажіть
Оброблено термін: підкажіть будь ласка
                       добрий      день   скажіть      будь     ласка  \
Dmy                  0.000000  0.000000  0.000000  0.000000  0.000000
Vit                 0.000000  0.000000  0.000000  0.000000  0.000000
Olel                 0.085777  0.069204  0.219455  0.191641  0.191641
Iva                  0.000000  0.000000  0.000000  0.000000  0.000000
Дим                  0.000000  0.000000  0.000000  0.000000  0.000000
And                  0.000000  0.000000  0.000000  0.000000  0.000000
Ван                  0.000000  0.017768  0.000000  0.216490  0.255852
Оле                  0.173468  0.079973  0.000000  0.000000  0.000000
Арс                  0.000000  0.000000  0.000000  0.000000  0.000000
sl1                  0.223191  0.051448  0.000000  0.113977  0.113977
Svi                  0.000000  0.008284  0.011676  0.058116  0.058116
                        дякую  підкажіть  добрий день  день скажіть  \
Dmy                  0.035356   0.000000     0.000000      0.000000
Vit                 0.045082   0.000000     0.000000      0.000000
Olel                 0.012081   0.021444     0.085777      0.085581
Iva                  0.000000   0.000000     0.000000      0.000000
Дим                  0.114314   0.000000     0.000000      0.000000
And                  0.000000   0.000000     0.000000      0.000000
Ван                  0.186107   0.066068     0.000000      0.000000
Оле                  0.111690   0.024781     0.099125      0.000000
Арс                  0.022281   0.000000     0.000000      0.000000
sl1                  0.143705   0.000000     0.063769      0.000000
Svi                  0.144620   0.000000     0.000000      0.000000

                     день підкажіть  ...  добрий день підкажіть  \
Dmy                        0.000000  ...               0.000000
Vit                       0.000000  ...               0.000000
Olel                       0.028527  ...               0.028527
Iva                        0.000000  ...               0.000000
Дим                        0.000000  ...               0.000000
And                        0.000000  ...               0.000000
Ван                        0.000000  ...               0.000000
Оле                        0.000000  ...               0.000000
Арс                        0.000000  ...               0.000000
sl1                        0.000000  ...               0.000000
Svi                        0.000000  ...               0.000000

                     день скажіть будь  день підкажіть будь  \
Dmy                           0.000000             0.000000
Vit                          0.000000             0.000000
Olel                          0.085581             0.028527
Iva                           0.000000             0.000000
Дим                           0.000000             0.000000
And                           0.000000             0.000000
Ван                           0.000000             0.000000
Оле                           0.000000             0.000000
Арс                           0.000000             0.000000
sl1                           0.000000             0.000000
Svi                           0.000000             0.000000
```

```
                     скажіть будь ласка  будь ласка добрий  будь ласка дякую  \
Dmy                            0.000000           0.000000          0.000000
Vit                           0.000000           0.000000          0.000000
Ole                            0.219455           0.000000          0.000000
Iva                            0.000000           0.000000          0.000000
Дим                            0.000000           0.000000          0.000000
And                            0.000000           0.000000          0.000000
Ван                            0.000000           0.000000          0.066068
Оле                            0.000000           0.000000          0.000000
Арс                            0.000000           0.000000          0.000000
sl1                            0.000000           0.042416          0.063769
Svi                            0.011676           0.000000          0.006845

                     ласка дякую добрий  ласка дякую підкажіть  \
Dmy                            0.000000               0.000000
Vit                           0.000000               0.000000
Ole                            0.000000               0.000000
Iva                            0.000000               0.000000
Дим                            0.000000               0.000000
And                            0.000000               0.000000
Ван                            0.000000               0.029296
Оле                            0.000000               0.000000
Арс                            0.000000               0.000000
sl1                            0.042416               0.000000
Svi                            0.000000               0.000000

                     підкажіть будь ласка      Сума
Dmy                              0.000000  0.035356
Vit                             0.000000  0.045082
Ole                              0.028527  1.906949
Iva                              0.000000  0.000000
Дим                              0.000000  0.114314
And                              0.000000  0.000000
Ван                              0.000000  1.178800
Оле                              0.000000  0.517217
Арс                              0.000000  0.022281
sl1                              0.000000  1.151339
Svi                              0.000000  0.375970

[11 rows x 28 columns]
```

Fig. 7. Result of determining the user's text feature

This figure shows that the user Ole has the highest probability in the analyzed text, which assists in identifying their text features.

Results with system operation in a cybercriminal chat

To comply with the journal's confidentiality rules, critical information, including the name of the cybercriminal chat, was removed. One of the Telegram cybercriminal chats in Cyrillic was chosen as the target. As a result, a folder containing 27 HTML pages with text was created. After processing this data using the system parser, a directory with 503 user-specific TXT files was generated which is shown in Fig. 8.

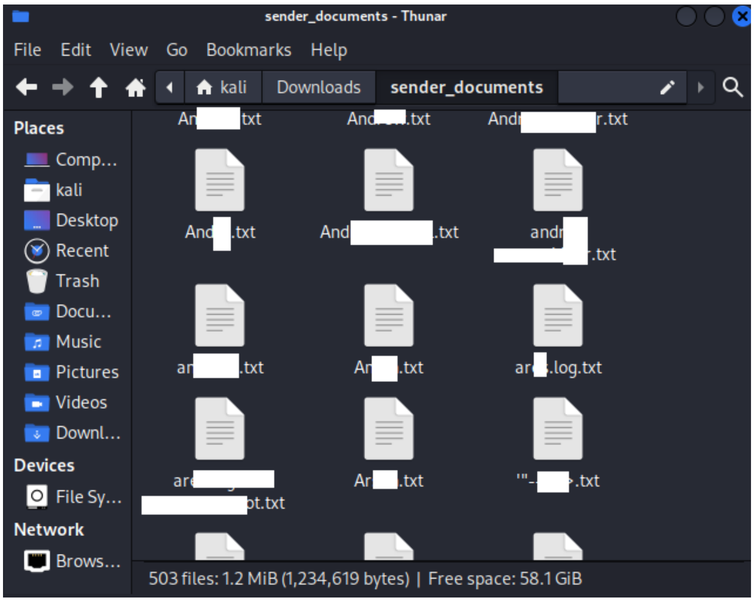
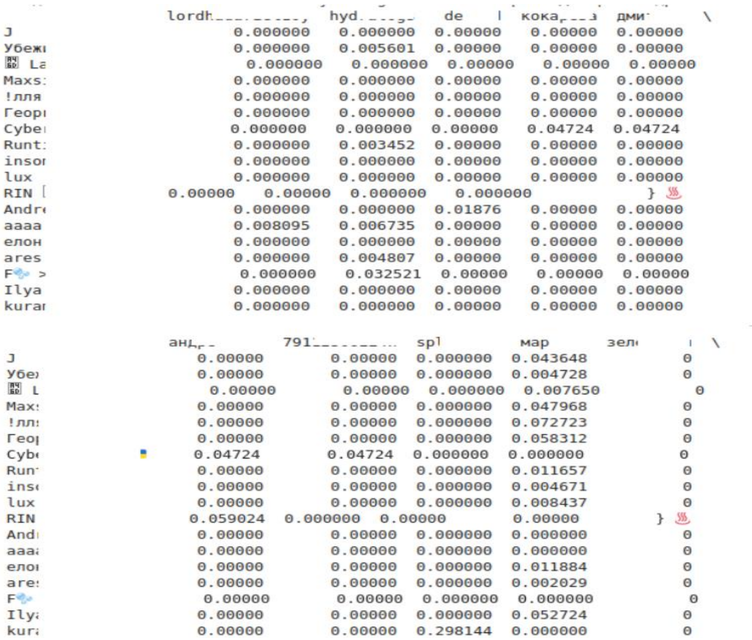


Fig. 8. An example of Telegram cybercriminals chat information gathering

Following the application of the ID deanonymization system, 10 main IDs appearing in the majority of TXT files were identified. The next step involved establishing the connection between the IDs and users by using the Similarity Search System. Fig. 9 presents only the results of applying this system.

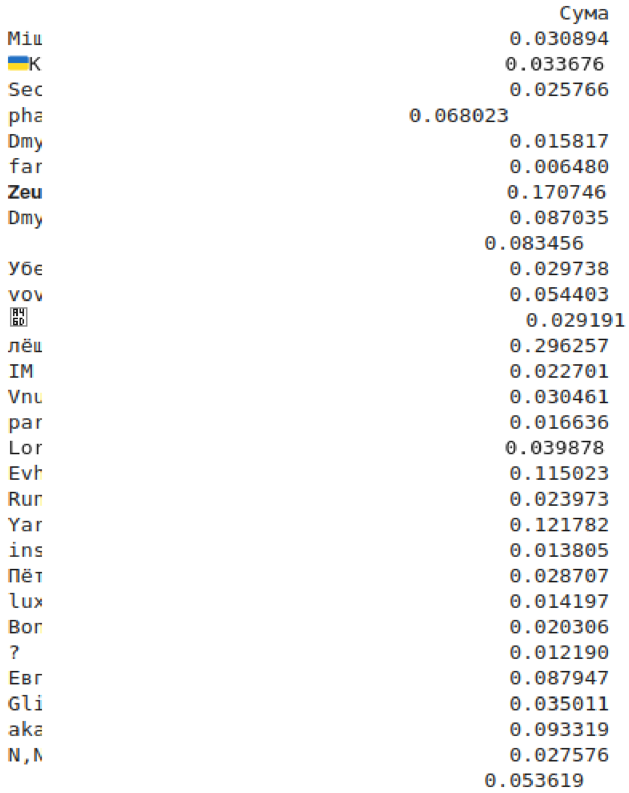


```
              lordh           hyd        de       кока       дми       \
J               0.000000   0.000000  0.00000   0.00000   0.00000
Убежі           0.000000   0.005601  0.00000   0.00000   0.00000
 La              0.000000   0.000000  0.00000   0.00000   0.00000
Maxs:           0.000000   0.000000  0.00000   0.00000   0.00000
!лля            0.000000   0.000000  0.00000   0.00000   0.00000
Геор            0.000000   0.000000  0.00000   0.00000   0.00000
Cybe           0.000000   0.000000  0.00000   0.04724   0.04724
Runt:           0.000000   0.003452  0.00000   0.00000   0.00000
insor           0.000000   0.000000  0.00000   0.00000   0.00000
lux             0.000000   0.000000  0.00000   0.00000   0.00000
RIN [     0.00000   0.00000   0.000000   0.000000             }
Andr            0.000000   0.000000  0.01876   0.00000   0.00000
aaaa            0.008095   0.006735  0.00000   0.00000   0.00000
елон            0.000000   0.000000  0.00000   0.00000   0.00000
ares            0.000000   0.004807  0.00000   0.00000   0.00000
F >              0.000000   0.032521  0.00000   0.00000   0.00000
Ilya            0.000000   0.000000  0.00000   0.00000   0.00000
kurar           0.000000   0.000000  0.00000   0.00000   0.00000

              анд         791        spl        мар        зел       \
J            0.00000      0.00000  0.000000  0.043648          0
Убе          0.00000      0.00000  0.000000  0.004728          0
 L            0.00000      0.00000  0.000000  0.007650          0
Max:         0.00000      0.00000  0.000000  0.047968          0
!лл:         0.00000      0.00000  0.000000  0.072723          0
Гео          0.00000      0.00000  0.000000  0.058312          0
Cyb          0.04724      0.04724  0.000000  0.000000          0
Run          0.00000      0.00000  0.000000  0.011657          0
ins          0.00000      0.00000  0.000000  0.004671          0
lux          0.00000      0.00000  0.000000  0.008437          0
RIN         0.059024  0.000000  0.00000      0.00000         }
And          0.00000      0.00000  0.000000  0.000000          0
aaa:         0.00000      0.00000  0.000000  0.000000          0
елоі         0.00000      0.00000  0.000000  0.011884          0
are:         0.00000      0.00000  0.000000  0.002029          0
F             0.00000      0.00000  0.000000  0.000000          0
Ilya         0.00000      0.00000  0.000000  0.052724          0
kur:         0.00000      0.00000  0.298144  0.000000          0
```

Fig. 9. Result of ID clarification

This figure demonstrates that the result was successful and the required ID was correctly identified. The next step is to define the users' text features. The input text used for the search was the same as in the previous section. Fig. 10 illustrates the example of applying the system.



```
                                    Сума
Miш                             0.030894
🇺🇦K                             0.033676
Sec                             0.025766
pha                     0.068023
Dmy                             0.015817
far                             0.006480
Zeu                             0.170746
Dmy                             0.087035
                              0.083456
Убє                             0.029738
vov                             0.054403
□                                 0.029191
лёш                             0.296257
IM                              0.022701
Vnu                             0.030461
par                             0.016636
Lor                             0.039878
Evh                             0.115023
Rur                             0.023973
Yar                             0.121782
ins                             0.013805
Пёт                             0.028707
lux                             0.014197
Bor                             0.020306
?                               0.012190
Евг                             0.087947
Gli                             0.035011
aka                             0.093319
N,N                             0.027576
                              0.053619
```

Fig. 10. Result of determining the user's text feature

Analysis of the system's performance

To define the velocity of the whole system, the analysis of speed was conducted. The execution time for the main operational steps of the system was measured. Table 3 presents the results of the system's execution time.

Table 3

Proposed system execution time

Operation	**Operation time for test chat (seconds)**	**Operation time for cybercriminal chat (seconds)**
Parser information processing system	1.55	53.24
ID deanonymization system	57.63	66.32
ID clarification	2.37	3.71
Text features definition	8.28	443.12

This table shows the high speed of the system operation for the test chat. The most time was taken by the ID deanonymization system due to the use of a large wordlist. The results indicate a dependency between the number of users in the Parser information processing system and the Text features definition.

DISCUSSION

Main findings

The experiments with a test chat have shown that the system for deanonymization of identifiers and defining text features is functional. The analysis revealed that the identifier definition stage takes the most time. This is because this stage utilizes the wordlist, resulting in longer execution times. The larger the wordlist, the more time this stage will require.

The experiment with a cybercriminal chat shows that the proposed system can succeed; however, it required 34 times more time to process the gathered information and 53 times more time to define users' text features compared to the test chat. Additionally, the cybercriminal chat had 503 users, 45 times more than the 11 users in the test chat. For stages related to identifiers, the execution time showed only a small difference, indicating that it does not significantly depend on the number of users.

The provided system can succeed in the defined tasks; however, it has limitations, as described in the limitations section. Also, with a higher number of users, the complexity of manually monitoring the results increases, and the possibility of implementing AI [25] should be considered.

Comparison with previous studies

Compared with previous studies described in the literature review section, the proposed system uses information retrieval techniques, which provide the possibility to estimate the probabilities of users' common features in the text and perform deanonymization. Because a different dataset was used, the proposed system cannot be directly compared with previous studies [8], [10] – [14]; however, based on the results, it can be considered one of the deanonymization methods.

CONCLUSIONS AND PROSPECTS FOR FURTHER RESEARCH

As the chats in Cyrillic were used for the experiments, the availability of wordlists in open access and digital format is rather limited. This limitation leads to some restrictions in experiments. The experiments showed that the similarity search system, using the vector model of IR and TF-IDF, depends on the consistency of all user messages. It also relies on the user input query for finding text features. If there are enough meaningful messages, the deanonymization of cybercriminals is possible.

Measures of operation speed were performed, and the results show high velocity for the test chat and moderate velocity for the cybercriminal chat. The results depend on several factors, such as the wordlist, dataset size, and the amount of user messages.

The experiments conducted show the efficiency of the proposed approach and the feasibility of using the system for analyzing messages from sources used by cybercriminals. The proposed solutions make it possible to establish related texts and affiliation to a specific author. Provided that a real name and nickname relation exists in cyberspace, the system allows author deanonymization. For the analysis of modern threats developed by cybercriminals, deanonymization is not as relevant as it is important to determine the affiliation of texts to one author.

The proposed solutions can be used in the work of security services that collect up-to-date information on the state of the latest threats of a certain type, for critical infrastructure facilities of a specific sector. Analysis of information from hacker chats can provide an understanding of popular trends and intentions in the cybercriminal environment in relation to critical infrastructure facilities.

The next research can include investigating messages using not only common dictionary wordlists but also hacker jargon wordlists, enabling the system to identify specialized terms and slang commonly used in cybercriminal communications, thereby enhancing its ability to detect hidden identifiers and patterns specific to this context. Additionally, future efforts could focus on implementing language-specific rules to adjust patterns for specific languages or platforms, ensuring greater accuracy across diverse linguistic datasets. For more complex cases, machine learning models, such as NLP models trained on identifier-rich datasets, could supplement regex-based approaches to improve detection. API integration with platforms like Twitter or Instagram could also be explored, enabling real-time validation of handles and other identifiers. This multi-faceted approach ensures flexibility and precision in extracting nicknames and related identifiers from user-generated text, broadening the system's applicability in diverse cybersecurity scenarios.

REFERENCES (TRANSLATED AND TRANSLITERATED

1. Nandan, A. B. (2021). Cybercrimes and Its Alarming Escalation during Recent Times: An International Legal Perspective. *International Journal of Law Management & Humanities, 4(4), 2413*.
2. Takey, Y. S., Tatikayala, S. G., Samavedam, S. S., Lakshmi Eswari, P. R., & Patil, M. U. (2021). Real Time early Multi Stage Attack Detection. *IEEE Xplore.* https://doi.org/10.1109/ICACCS51430.2021.9441956
3. Teendifferent. (2022). *Information Gathering In Cyber Security: Definition, Types, Tools & Techniques.* Medium. https://medium.com/@teendifferent/information-gathering-in-cyber-security-definition-types-tools-techniques-ae59cb394bf6
4. Chawki, M. (2010). Anonymity in cyberspace: finding the balance between privacy and security. *International Journal of Technology Transfer and Commercialisation, 9(3), 183*. https://doi.org/10.1504/ijttc.2010.030209
5. Chawki, M. & Wahab M. S. A. (2006). Identity Theft in Cyberspace: Issues and Solutions. *Lex Electronica, 11(1).*
6. Gröndahl, T., & Asokan, N. (2019). Text Analysis in Adversarial Settings. *ACM Computing Surveys, 52(3),* 1–36. https://doi.org/10.1145/3310331
7. Manning, C. D., Raghavan, P., & Schütze, H. (2008). Introduction to Information Retrieval. *Cambridge: Cambridge University Press.*
8. Rüdian, S., Pinkwart, N. & Liu, Z. (2018). I know who you are: Deanonymization using Facebook Likes. *Workshops der INFORMATIK*, 109–118.
9. Seorld. (2024). *PHP Facebook-Crawler*. Seorld.com. https://seorld.com/blog/social-media/facebook
10. Simioni, M., Gladyshev, P., Habibnia, B., & Nunes de Souza, P. R. (2021). Monitoring an anonymity network: Toward the deanonymization of hidden services. *DFRWS APAC*, 1–8
11. Boldyrikhin, N. V., Altunin, F. A., Svizhenko, A. A., Sosnovsky, I. A., & Yengibaryan, I. A. (2021). Deanonymization of users based on correlation analysis. *Journal of Physics: Conference Series, 2131(2), 022083*. https://doi.org/10.1088/1742-6596/2131/2/022083
12. Beato, F., Conti, M., & Preneel, B. (2013). Friend in the Middle (FiM): Tackling de-anonymization in social networks. *IEEE International Conference on Pervasive Computing and Communications Workshops (PERCOM Workshops).* https://doi.org/10.1109/percomw.2013.6529495
13. Peng W., Li F., Zou X., & Wu J. (2014). A Two-Stage Deanonymization Attack against Anonymized Social Networks. *IEEE Transactions on Computers, 63(2),* 290–303. https://doi.org/10.1109/tc.2012.202
14. Jiang, H., Yu, J., Cheng, X., Zhang, C., Gong, B., & Yu, H. (2022). Structure-Attribute-Based Social Network Deanonymization With Spectral Graph Partitioning. *IEEE Transactions on Computational Social Systems, 9(3),* 902–913. https://doi.org/10.1109/tcss.2021.3082901
15. Miller, M. (2022). *TF-IDF: Is It A Google Ranking Factor?* Search Engine Journal. https://www.searchenginejournal.com/ranking-factors/tf-idf/

16. Chalyi, O. (2023). Information Retrieval as A Way to Search for Common Features in The Text. *XXIV International R&D Online Conference for Students and Emerging Researchers "Science and Technology of the XXI Century", 1(57),* 16–18.

17. Richardson, L. (2024). *Beautiful Soup Documentation.* Crummy.com. https://www.crummy.com/software/BeautifulSoup/bs4/doc/

18. Guido Van Rossum, & Drake, F. L. (2011). *The Python language reference manual: for Python version 3.2*. Network Theory Ltd.

19. Scikit, L. (2024). *TfidfVectorizer — scikit-learn 0.20.3 documentation. Scikit-learn.org.* https://scikit-learn.org/stable/modules/generated/sklearn.feature_extraction.text.TfidfVectorizer.html

20. *Bakustarver*. (2024). GitHub. https://github.com/bakustarver/ukr-dictionaries-list-opensource

21. Bilodid, I. C. (2024). *Dictionary of the Ukrainian language in 11 volumes.* ukrlit.org. http://ukrlit.org/slovnyk/slovnyk_ukrainskoi_movy_v_11_tomakh

22. *Brown-uk.* (2024). GitHub. https://github.com/brown-uk/dict_uk/blob/master/data/dict/names-anim.lst

23. Raymond, E. (2024). *the Jargon File.* Netmeister. https://www.netmeister.org/news/jargon.html

24. Raymond, E. (2024). *The Original Hacker's Dictionary*. Netmeister https://www.dourish.com/goodies/jargon.html

25. Chalyi, O. (2024). An Evaluation of General-Purpose AI Chatbots: A Comprehensive Comparative Analysis. *InfoScience Trends, 1(1),* 52–66. https://doi.org/10.61186/ist.202401.01.07

Чалий Олексій Віталійович
магістр кібербезпеки, магістр обчислення,
лаборант інституту соціальних наук та прикладної інформатики
Каунісівський факультет Вільнюський університет, Каунас, Литва
ORCID ID: 0009-0006-3536-9715
oleksii.chalyi@knf.vu.lt

Стьопочкіна Ірина Валеріївна
к.т.н., доцент, доцент фізико технічного інституту
Національний технічний університет України «Київський
політехнічний інститут імені Ігоря Сікорського», Київ, Україна
ORCID ID: 0000-0002-0346-0390
i.stopochkina@kpi.ua

ДЕАНОНІМІЗАЦІЯ ТА ПОШУК ІНФОРМАЦІЇ В ЗАДАЧАХ РАННЬОГО ЗНАХОДЖЕННЯ ПОТЕНЦІЙНИХ АТАК НА КРИТИЧНУ ІНФРАСТРУКТУРУ

Анотація. Інформація про кібератаки, які зловмисники планують здійснити щодо об’єктів критичної інфраструктури, частково поширюється на зловмисних інформаційних сайтах. Дослідження інформаційних матеріалів та їх аналіз може дати розуміння етапів планування атак та їх запобігання. Частиною цієї проблеми є надання інструментів пошуку та аналізу інформації для виявлення лінгвістичних закономірностей, схожості в текстових даних, які здатні деанонімізувати кіберзлочинців та встановлювати взаємозв’язки між відкритими даними. У цій роботі запропоновано нову модель та відповідний прототип системи, що базується на моделі векторного простору та алгоритмі TF-IDF. Система призначена для аналізу загальнодоступних текстових даних (як в Інтернеті, так і в даркнеті) і відрізняється ймовірнісним підходом до аналізу ідентифікаторів автора інформації. Запропонована система також фокусується на виявленні прихованих зв’язків між анонімними акаунтами шляхом аналізу унікальних стилістичних і мовних особливостей. Вона використовує ці риси для відстеження шаблонів у спілкуванні, виявляючи приховані асоціації між кіберзлочинцями. Експерименти, проведені на основі аналізу реальних чатів, у тому числі чатів кіберзлочинців, демонструють потенціал системи для виявлення ідентифікаторів та визначення стилістичних особливостей. За наявності достатньо повного набору даних і списку цільових слів можна проаналізувати етапи підготовки атаки, зловмисників або групи, які беруть у ній участь. Результати дослідження підкреслюють важливість інтеграції передових методів лінгвістичного аналізу з ймовірнісними моделями для розширення можливостей розслідування кіберзагроз, що еволюціонують.

Ключові слова: пошук інформації; кібербезпека; деанонімізація; модель векторного поля; критична інфраструктура; кіберзлочинність; tf-idf алгоритм.

СПИСОК ВИКОРИСТАНИХ ДЖЕРЕЛ

1. Nandan, A. B. (2021). Cybercrimes and Its Alarming Escalation during Recent Times: An International Legal Perspective. *International Journal of Law Management & Humanities, 4(4), 2413*.
2. Takey, Y. S., Tatikayala, S. G., Samavedam, S. S., Lakshmi Eswari, P. R., & Patil, M. U. (2021). Real Time early Multi Stage Attack Detection. *IEEE Xplore.* https://doi.org/10.1109/ICACCS51430.2021.9441956
3. Teendifferent. (2022). *Information Gathering In Cyber Security: Definition, Types, Tools & Techniques.* Medium. https://medium.com/@teendifferent/information-gathering-in-cyber-security-definition-types-tools-techniques-ae59cb394bf6
4. Chawki, M. (2010). Anonymity in cyberspace: finding the balance between privacy and security. *International Journal of Technology Transfer and Commercialisation, 9(3), 183*. https://doi.org/10.1504/ijttc.2010.030209
5. Chawki, M. & Wahab M. S. A. (2006). Identity Theft in Cyberspace: Issues and Solutions. *Lex Electronica, 11(1).*
6. Gröndahl, T., & Asokan, N. (2019). Text Analysis in Adversarial Settings. *ACM Computing Surveys, 52(3),* 1–36. https://doi.org/10.1145/3310331

7. Manning, C. D., Raghavan, P., & Schütze, H. (2008). Introduction to Information Retrieval. *Cambridge: Cambridge University Press.*
8. Rüdian, S., Pinkwart, N. & Liu, Z. (2018). I know who you are: Deanonymization using Facebook Likes. *Workshops der INFORMATIK*, 109–118.
9. Seorld. (2024). *PHP Facebook-Crawler*. Seorld.com. https://seorld.com/blog/social-media/facebook
10. Simioni, M., Gladyshev, P., Habibnia, B., & Nunes de Souza, P. R. (2021). Monitoring an anonymity network: Toward the deanonymization of hidden services. *DFRWS APAC*, 1–8
11. Boldyrikhin, N. V., Altunin, F. A., Svizhenko, A. A., Sosnovsky, I. A., & Yengibaryan, I. A. (2021). Deanonymization of users based on correlation analysis. *Journal of Physics: Conference Series, 2131(2), 022083*. https://doi.org/10.1088/1742-6596/2131/2/022083
12. Beato, F., Conti, M., & Preneel, B. (2013). Friend in the Middle (FiM): Tackling de-anonymization in social networks. *IEEE International Conference on Pervasive Computing and Communications Workshops (PERCOM Workshops).* https://doi.org/10.1109/percomw.2013.6529495
13. Peng W., Li F., Zou X., & Wu J. (2014). A Two-Stage Deanonymization Attack against Anonymized Social Networks. *IEEE Transactions on Computers, 63(2),* 290–303. https://doi.org/10.1109/tc.2012.202
14. Jiang, H., Yu, J., Cheng, X., Zhang, C., Gong, B., & Yu, H. (2022). Structure-Attribute-Based Social Network Deanonymization With Spectral Graph Partitioning. *IEEE Transactions on Computational Social Systems, 9(3),* 902–913. https://doi.org/10.1109/tcss.2021.3082901
15. Miller, M. (2022). *TF-IDF: Is It A Google Ranking Factor?* Search Engine Journal. https://www.searchenginejournal.com/ranking-factors/tf-idf/
16. Chalyi, O. (2023). Information Retrieval as A Way to Search for Common Features in The Text. *XXIV International R&D Online Conference for Students and Emerging Researchers "Science and Technology of the XXI Century", 1(57),* 16–18.
17. Richardson, L. (2024). *Beautiful Soup Documentation.* Crummy.com. https://www.crummy.com/software/BeautifulSoup/bs4/doc/
18. Guido Van Rossum, & Drake, F. L. (2011). *The Python language reference manual: for Python version 3.2*. Network Theory Ltd.
19. Scikit, L. (2024). *TfidfVectorizer — scikit-learn 0.20.3 documentation. Scikit-learn.org.* https://scikit-learn.org/stable/modules/generated/sklearn.feature_extraction.text.TfidfVectorizer.html
20. *Bakustarver*. (2024). GitHub. https://github.com/bakustarver/ukr-dictionaries-list-opensource
21. Bilodid, I. C. (2024). *Dictionary of the Ukrainian language in 11 volumes.* ukrlit.org. http://ukrlit.org/slovnyk/slovnyk_ukrainskoi_movy_v_11_tomakh
22. *Brown-uk.* (2024). GitHub. https://github.com/brown-uk/dict_uk/blob/master/data/dict/names-anim.lst
23. Raymond, E. (2024). *the Jargon File.* Netmeister. https://www.netmeister.org/news/jargon.html
24. Raymond, E. (2024). *The Original Hacker's Dictionary*. Netmeister https://www.dourish.com/goodies/jargon.html
25. Chalyi, O. (2024). An Evaluation of General-Purpose AI Chatbots: A Comprehensive Comparative Analysis. *InfoScience Trends, 1(1),* 52–66. https://doi.org/10.61186/ist.202401.01.07