Deceptive Operations in Document Repositories: Manipulating Clustering Outcomes Against Adversaries

By:

Sayyed Shozib Abbas
(Registration No.: MS-SE-20-327669)

Supervisor:
Dr. Ali Hassan

Co-Supervisor:
Dr. Muhammad Yasin

DEPARTMENT OF COMPUTER & SOFTWARE ENGINEERING,
COLLEGE OF ELECTRICAL & MECHANICAL ENGINEERING,
NATIONAL UNIVERSITY OF SCIENCES AND TECHNOLOGY,
ISLAMABAD
July 24, 2024

Annex A

THESIS ACCEPTANCE CERTIFICATE

Certified that final copy of MS/MPhil thesis written by NS **Sayyed Shozib Abbas** Registration No. **00000327669**, of College of E&ME has been vetted by undersigned, found complete in all respects as per NUST Statutes/Regulations, is free of plagiarism, errors and mistakes and is accepted as partial fulfillment for award of MS/MPhil degree. It is further certified that necessary amendments as pointed out by GEC members of the scholar have also been incorporated in the thesis.

Signature : _____

Name of Supervisor: Dr Ali Hassan

Date: 24 JUL 2024

Signature of HOD: _____
(Dr Usman Qamar)
Date: 24 JUL 2024

Signature of Dean: _____
(Brig Dr Nasir Rashid)
Date: 24 JUL 2024

Deceptive Operations in Document Repositories: Manipulating Clustering Outcomes Against Adversaries

By
Sayyed Shozib Abbas
(Registration No.: 00000327669)

A thesis submitted to the National University of Sciences and Technology Islamabad
in partial fulfillment of the requirements for the degree of

Master of Sciences in Software Engineering

Supervisor
Dr. Ali Hassan

Co Supervisor

Dr. Muhammad Yasin

DEPARTMENT OF COMPUTER & SOFTWARE ENGINEERING,
COLLEGE OF ELECTRICAL & MECHANICAL ENGINEERING,
NATIONAL UNIVERSITY OF SCIENCES AND TECHNOLOGY
ISLAMABAD
July 24, 2024

Dedicated to my family, whose unwavering support and encouragement have been my guiding light throughout my academic journey. To my mentors and supervisor, whose wisdom and guidance have shaped my knowledge and skills. And to my friends and colleagues, whose companionship and encouragement have made this journey memorable.

Acknowledgement

First and foremost, I would like to express my deepest gratitude to my supervisor, Dr. Ali Hassan and co-supervisor, Dr. Muhammad Yasin, for their invaluable guidance, continuous support, and patience throughout my MS journey. Their insightful feedback and encouragement were essential in the completion of this thesis.

I am profoundly grateful to my committee members, Dr. Farhan Hussain, and Mam Anum Abdul Salam, for their valuable suggestions and constructive criticism, which significantly enhanced the quality of this work.

A special thanks to my family, whose love, understanding, and sacrifices have been my constant source of strength and motivation. Their belief in me has always pushed me to strive for excellence.

Abstract

This research investigates the efficacy of replacement and shuffling techniques to enhance the confidentiality and integrity of sensitive information within diverse document types. The study introduces the Deceptive Approaches for Robust Defense (DARD) technique, which aims to anonymize and protect numerical data and confidential text. The effectiveness of this technique is evaluated using three distinct datasets.

The first dataset consists of 300 research papers on Artificial Intelligence, Cryptography, and Databases. The second dataset includes summaries of 3000 research papers spanning Artificial Intelligence, Cryptography, Databases, and Networking. The third dataset encompasses company documents classified into Inventory Reports, Invoices, Purchase Orders, and Shipping Orders. The comparative analysis between the first and second datasets, and the first and third datasets, demonstrates the DARD technique's proficiency in anonymizing and securing sensitive data across various document types.

The findings reveal that the DARD technique effectively safeguards confidential information in both academic research papers and business documents, with a particular strength in handling documents containing numerical data and sensitive content. This research contributes to the field of data security by providing a robust method for protecting sensitive documents, thereby addressing critical issues in cybersecurity practices. The study underscores the potential of the DARD technique to serve as a reliable tool for ensuring data confidentiality and integrity, offering significant implications for both academic and commercial applications.

The results validate the technique's applicability in real-world scenarios, highlighting its importance in the ongoing efforts to enhance data privacy and security.

Keywords: Data Anonymization, Confidentiality, Cybersecurity, Document Security

Contents

List of Tables

List of Figures

Symbols, Abbreviations and Acronyms

IP	Intellectual Property
Adversary	an entity which is to attempt IP theft and extract meaningful information
Attacker	Used in the same meanings as that of Adversary
IP Theft	Acquiring a form of medium which is valuable for an organization
Deception	misguiding to accept a phenomenon as true which is originally false
Repository	a set of mediums of similar nature

Chapter 1

Introduction

1.1 Background

In today's digital age, safeguarding sensitive information is a critical concern across various domains, including academia and industry. Traditional methods of data protection often fall short when faced with the complexities of modern data structures and the growing sophistication of cyber threats. This research addresses these challenges by exploring advanced anonymization techniques, specifically replacement and shuffling methods. The development of the Deceptive Approaches for Robust Defense (DARD) technique aims to enhance the confidentiality and integrity of numerical data and confidential text within diverse document types [1]. By evaluating the DARD technique on datasets comprising research papers and business documents, this study seeks to demonstrate its effectiveness in ensuring data privacy and security. This work contributes to the broader field of cybersecurity by providing a novel approach to protecting sensitive information, thereby addressing crucial gaps in existing data protection strategies.

1.2 Research Objectives

The primary aim of this research is to develop and validate the Deceptive Approaches for Robust Defense (DARD) technique, which leverages replacement, and shuffling methods to protect sensitive information within diverse document types. By applying this technique to various datasets, the study seeks to demonstrate its efficacy in anonymizing and securing numerical data and confidential text. The research is designed to address the critical need for advanced data protection methods that can effectively handle the complexities of modern data structures and the increasing sophistication of cyber threats.

1.2.1 Objectives

- To develop the DARD technique that integrates replacement and shuffling methods for enhanced data anonymization.
- To evaluate the effectiveness of the DARD technique on a dataset of 300 research papers covering topics such as Artificial Intelligence, Cryptography, and Databases.
- To assess the performance of the DARD technique on a second dataset comprising summaries of 3000 research papers across Artificial Intelligence, Cryptography, Databases, and Networking.
- To compare the DARD technique's effectiveness on a third dataset of company documents, including Inventory Reports, Invoices, Purchase Orders, and Shipping Orders.
- To validate the DARD technique's applicability in ensuring the confidentiality and integrity of sensitive documents in both academic and commercial contexts.

1.3 Problem Statement

In the contemporary digital landscape, the protection of sensitive information has become increasingly critical. With the exponential growth of data and the escalating sophistication of cyber threats, traditional data protection methods often prove inadequate. Confidential documents, whether they pertain to academic research or business operations, are particularly vulnerable to unauthorized access and misuse. This vulnerability necessitates the development of advanced techniques that can effectively safeguard sensitive information while maintaining data utility.

Existing methods for Deceptive Approaches for Robust Defense frequently fall short in addressing the complex nature of modern datasets. These methods often lack the robustness required to handle diverse document types, including numerical data and confidential text. Furthermore, the evolving landscape of cyber threats demands more sophisticated and adaptive techniques to ensure data security.

This research addresses these challenges by developing the Deceptive Approaches for Robust Defense (DARD) technique, which integrates replacement and shuffling methods. The DARD technique is designed to anonymize and protect sensitive information within various document types, thereby enhancing data confidentiality and integrity.

The study evaluates the DARD technique using three distinct datasets: 300 research papers on Artificial Intelligence, Cryptography, and Databases; summaries of 3000 research papers across Artificial Intelligence, Cryptography, Databases, and Networking; and company documents classified into Inventory Reports, Invoices, Purchase Orders, and Shipping Orders. By comparing the results across these datasets,

the research aims to demonstrate the effectiveness of the DARD technique in different contexts.

The primary problem this research seeks to solve is the inadequacy of current data protection methods in ensuring the confidentiality and integrity of sensitive documents. By developing and validating the DARD technique, this study aims to provide a robust solution to protect confidential information, thereby addressing critical gaps in existing data security practices and contributing to the broader field of cybersecurity.

1.4 Significance of Studying Methods to Secure Data Breaches

The study of methods to secure data breaches is of paramount importance in today's digital era, where vast amounts of sensitive information are stored and transmitted electronically [2]. With the exponential increase in cyber threats, organizations across various sectors face significant risks of data breaches, which can lead to severe financial losses, reputational damage, and legal repercussions. Effective data security methods are crucial in safeguarding personal, academic, and business information from unauthorized access and misuse.

Advancing techniques such as the Deceptive Approaches for Robust Defense (DARD) addresses the limitations of traditional data protection methods, ensuring robust anonymization and confidentiality of sensitive data. These advanced methods are essential for maintaining data integrity and privacy, particularly in handling complex datasets that include numerical and textual information.

By enhancing the security of confidential documents, these methods contribute significantly to the broader field of cybersecurity. They provide organizations with the

tools to protect their data assets, thereby fostering trust among stakeholders and compliance with regulatory requirements. Ultimately, the study and implementation of effective data breach prevention techniques are vital in mitigating cyber risks.

Chapter 2

Literature Review

The rise of cyber threats targeting sensitive information necessitates advanced defensive strategies in data security. Traditional data protection methods often prove inadequate in the face of sophisticated adversaries who use automated techniques to exfiltrate and analyse confidential documents. This literature review explores various approaches to data anonymization, adversarial settings, and document clustering, with a particular focus on the Data Approaches for Robust Defense (DARD) system and its effectiveness in protecting intellectual property.

Adversarial settings have become a critical aspect of cybersecurity, particularly for protecting data from malicious clustering attacks. Previous studies have extensively examined attacks on clustering algorithms through the generation of adversarial settings. These attacks can be broadly classified into poisoning and obfuscation attacks. Poisoning attacks aim to degrade clustering results by injecting malicious examples into the dataset, creating new clusters or bridges between clusters to cause misclassification. Obfuscation attacks, on the other hand, seek to hide specific datasets by merging target clusters with others, effectively concealing the original cluster within another [3] [4] [5].

For instance, Dutrisac and Skillicorn explored the effectiveness of hiding clusters in adversarial settings, demonstrating the feasibility of such techniques in misleading clustering algorithms [3]. Similarly, Biggio et al. Investigated the security of data clustering in adversarial settings, highlighting how adversaries can manipulate clustering results to achieve their objectives [4]. While these studies provide valuable insights into adversarial attacks, they primarily focus on offensive strategies rather than defensive mechanisms.

The DARD system introduces a novel approach to defending against adversaries by employing deceptive strategies. Unlike traditional methods that focus on encryption or obfuscation, DARD uses term-replacement operations to create a deceptive repository that misleads automated clustering and topic modelling techniques. This approach is particularly effective in scenarios where adversaries rely on automated tools to analyse exfiltrated documents [4].

The concept of using deceptive strategies for data protection is not entirely new. Previous research has explored the use of fake document generation to deceive adversaries. For example, Karuna et al. Proposed generating fake documents by manipulating text comprehensibility to prevent intellectual property theft [5]. Similarly, Abdibayev et al. Utilized word embeddings to create fake documents, effectively deterring adversaries from identifying valuable information [6]. However, DARD distinguishes itself by focusing on the Defense of text document classifiers through adversarial settings, shifting the paradigm from attacking models to protecting data repositories.

Document clustering and topic modelling are essential techniques for organizing and retrieving information from unstructured text documents. These methods are widely

used in various applications, including information retrieval, text mining, and data analysis. The goal of document clustering is to group similar documents together, while topic modelling aims to identify the underlying topics within a collection of documents [7] [8].

One of the most popular clustering algorithms is K-means, which partitions documents into clusters based on their features [9]. The efficacy of K-means and other clustering algorithms is often evaluated using internal validation metrics such as the Silhouette Coefficient [10], the Calinski-Harabasz Index [11], and the Davies-Bouldin Index [12]. These metrics assess the quality of clustering by measuring the cohesion and separation of clusters.

Latent Dirichlet Allocation (LDA) is a widely used topic modelling algorithm that identifies the latent semantic structure of a document collection [8]. LDA provides a list of terms that describe each topic, enabling users to infer the main subjects covered by the documents. This technique is particularly useful for adversaries seeking to identify valuable information within exfiltrated repositories.

The effectiveness of the DARD system is evaluated through extensive experiments involving three types of adversaries: Black Box, Gray Box, and Enhanced Gray Box. Black Box adversaries are unaware of the deceptive operations, while Gray Box and Enhanced Gray Box adversaries have varying levels of knowledge about the DARD system. The evaluation focuses on the Adjusted Rand Index (ARI), a measure of the similarity between the predicted and true clustering of documents [13] [14].

The results indicate that DARD's deceptive operations significantly degrade the clustering performance of Black Box adversaries, resulting in an ARI close to zero. Gray Box and Enhanced Gray Box adversaries, although more knowledgeable about

the deceptive operations, still experience a substantial reduction in ARI. This demonstrates the robustness of DARD in protecting sensitive information from automated analysis techniques [15].

Furthermore, DARD's impact on topic modelling is assessed using LDA. The experiments show that LDA retrieves only deceptive keywords from the deceptive repositories, effectively misleading adversaries about the actual topics covered by the documents. This finding underscores the potential of DARD to enhance data security by deceiving commercial tools such as Amazon Comprehend [8].

In addition to its standalone capabilities, DARD can complement traditional encryption techniques to provide a multi-layered Defense strategy. By combining DARD with encryption, organizations can create decoy repositories or honeypots containing fabricated information. This approach diverts adversaries' attention from genuine sensitive data, enhancing overall security. The integration of DARD and encryption offers a comprehensive solution to protect data from sophisticated cyber threats [16].

The literature on data anonymization, adversarial settings, and document clustering provides a solid foundation for understanding the challenges and opportunities in protecting sensitive information. The DARD system represents a significant advancement in the field of data security by introducing deceptive strategies to mislead automated analysis techniques. Through extensive evaluation, DARD has demonstrated its efficacy in degrading the performance of adversaries and protecting valuable information. This review highlights the importance of continuous innovation in defensive strategies to stay ahead of evolving cyber threats.

Chapter 3

Data Collection and Preprocessing

3.1 Data Collection

Data collection is a critical component of this research, providing the foundation necessary to develop and evaluate the Deceptive Approaches for Robust Defense (DARD) technique. The aim is to safeguard sensitive information through advanced data anonymization and shuffling methods. By gathering diverse datasets, the research ensures a comprehensive analysis of DARD's effectiveness across various document types and content structures. This section outlines the methodologies and specifics of the three datasets used: research papers, summaries of research papers, and company documents.

3.1.1 Dataset One: Research Papers

The first dataset consists of **300 research papers** downloaded from Sci-Hub. These papers were selected to provide a robust foundation for analysing the DARD technique's effectiveness in protecting academic documents. The papers cover three key topics: Artificial Intelligence, Database, and Cryptography. The smallest document in this dataset contains approximately **1,500 words**, while the largest document consists of roughly **10,500 words**. The median word count for this dataset is **8,454 words**.

To compile this dataset, the PyPaperBot [17] library was utilized. This Python library automates the search and download process by taking a query, searching for relevant articles on Google Scholar, extracting DOIs through APIs at CrossRef, and downloading the papers from Sci-Hub mirrors. The three queries used were: Artificial Intelligence, Database, and Cryptography.

3.1.2 Dataset Two: Summaries of Research Papers

The second dataset includes **12,000 summaries** of research papers downloaded from arXiv. This dataset offers a different document structure, focusing on brief summaries rather than full papers, which is crucial for evaluating the DARD technique's versatility. The smallest summary in this dataset contains approximately **120 words**, and the largest summary has around **250 words**. The median word count for these summaries is **173 words**.

The arXiv [18] library for Python facilitated the collection of these summaries. This library takes a query, searches for relevant research papers, and provides detailed information on multiple papers. Summaries were gathered for four topics: Artificial Intelligence, Database, Cryptography, and Networking, with each topic contributing **3,000 summaries**.

3.1.3 Dataset Three: Company Documents

The third dataset is composed of **2,676 company documents** downloaded from Kaggle [19]. These documents include a variety of information such as numbers, product names, purchases, orders, and client names, making them ideal for testing DARD's performance with numerical and tabular data. Each document contains tables and numbers, adding another layer of complexity to the anonymization and shuffling process.

The company documents cover four key topics: Inventory Reports, Invoices, Purchase Orders, and Shipping Orders. This dataset is essential for evaluating the DARD technique's capability to handle practical business documents that contain both text and numerical data [19].

Effective preprocessing is a crucial step in preparing datasets for analysis, especially when dealing with large volumes of textual data. For this research, preprocessing was essential to standardize and clean the data, ensuring that subsequent analysis using the DARD technique would be accurate and meaningful. The preprocessing involved extracting text from PDF documents, cleaning the text, and preparing it for analysis. This section describes the preprocessing methodologies employed for the three datasets used in this research [20].

3.2 Data Extraction and Preprocessing

To preprocess the datasets, a robust pipeline was developed to handle the extraction and cleaning of text data. The process differed slightly for each dataset due to their varying formats and content structures.

3.2.1 Dataset One and Three: Research Papers and Company Documents

For the research papers and company documents, which were in PDF format, a custom script was used to extract and preprocess the text. The script utilized the PyPDF2 library to read and extract text from PDF files [20]. The extraction process was automated and parallelized using Python's concurrent.futures module to handle large volumes of documents efficiently.

The extracted text was then processed using a custom text processing method. This method involved several key steps to clean and standardize the text data:

3.2.2 Removal of ASCII Characters

Non-ASCII characters were removed to ensure that the text contained only standard English characters. This step was necessary to avoid issues with encoding and to simplify the text analysis process.

3.2.3 Removal of Newline Characters

Newline characters were removed to create a continuous stream of text. This helped in maintaining the context of sentences and improving the tokenization process.

3.2.4 Lowercasing

All text was converted to lowercase to ensure uniformity. This step helped in reducing the complexity of the data by treating words with different cases as the same word.

3.2.5 Tokenization

The text was split into individual tokens (words). This step was crucial for further text processing tasks such as lemmatization and stop word removal. [21]

3.2.6 Stop Word Removal

Common English stop words (e.g., “the,” “is,” “in”) were removed from the text. These words typically do not contribute meaningful information to the analysis and their removal helped in reducing noise in the data. To perform this, NLTK 3.0 python library has been used. [22]

3.2.7 Lemmatization

Each token was lemmatized to its base or root form. For example, words like “running,” “runs,” and “ran” were converted to their base form “run.” This step helped in standardizing the words and reducing the dimensionality of the data.

3.2.8 Dataset Two: Summaries of Research Papers

For the summaries of research papers, the text extraction step was skipped since the data was already available in text format. The same text processing method was applied to clean and standardize the summaries. Given the shorter length of the summaries compared to full research papers, this preprocessing ensured that the data was in a consistent format for analysis.

3.3 Extraction Conclusion

By following these preprocessing steps, the text data from all three datasets was standardized and cleaned, ensuring that the DARD technique could be applied effectively. Table 3.1 represents the statistics extracted from the three datasets. The preprocessing pipeline played a crucial role in preparing the data for subsequent analysis, enabling a thorough and accurate evaluation of the DARD technique’s effectiveness in protecting sensitive information.

Table 3.1 Summary of Extracted Datasets

Dataset	Extracted Words	Total Files
Research Papers	1,352,110	309
Summaries	1,193,853	12,000
Company Documents	47,925	2,676

A sample of the extracted dataset in the form of JSON is provided in Fig 3.1. It shows the topics as the keys of the object while the words are split document by document in the form of a 2D array. We have used JSON as it is the most popular communication interface

Fig 3.1 JSON made after extraction and preprocessing of data

Chapter 4

Methodology

4.1 TF-IDF Matrix

Term Frequency-Inverse Document Frequency (TF-IDF) is a numerical statistic that reflects the importance of a word in a document relative to a collection of documents (corpus) [23]. It is commonly used in text mining and information retrieval to evaluate how relevant a word is to a specific document within a corpus . TF-IDF is composed of two metrics: Term Frequency (TF) and Inverse Document Frequency (IDF).

4.1.1 Term Frequency (TF)

This measures the frequency of a term in a document. It is calculated as the number of times a term appears in a document, divided by the total number of terms in that document. The formula is given in Equation 4.1

Equation 4.1 Term Frequency Formula

$$TF(t,d) = \frac{\text{Number of times term } t \text{ appears in document } d}{\text{Total number of terms in document } d}$$

4.1.2 Inverse Document Frequency (IDF)

This measures the importance of a term within the entire corpus. It is calculated as the logarithm of the total number of documents in the corpus divided by the number of documents containing the term. The formula is given in Equation 4.2

Equation 4.2 Inverse Document Frequency Formula

$$IDF(t) = \log \frac{Total\ number\ of\ documents}{Number\ of\ documents\ containing\ term\ t}$$

4.1.3 TF-IDF Score Calculation

The TF-IDF score is the product of TF and IDF. The formula for calculation is mentioned in Equation 4.3

Equation 4.3 TF-IDF Matrix Formula

$$TF\text{-}IDF(t,d) = TF(t,d) \times IDF(t)$$

4.2 Importance of TF-IDF in Text Analysis

TF-IDF is crucial in text analysis because it helps to identify the most significant terms in a document, filtering out common but less informative words (like “the”, “is”, and “in”) that are frequent across documents but not necessarily important [24]. This feature extraction technique enhances the performance of various text analysis tasks, such as:

- **Information Retrieval**: TF-IDF improves search engine results by prioritizing documents containing relevant terms [25].
- **Document Clustering**: It aids in grouping similar documents together by their important terms [26].

- **Topic Modeling**: TF-IDF helps in identifying the key topics within a collection of documents [27].
- **Text Classification**: It enhances the ability of classifiers to distinguish between different categories of documents [24].

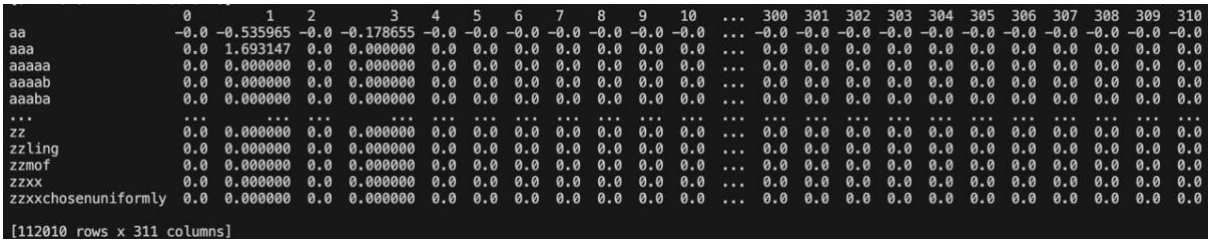
4.3 Construction of the TF-IDF Matrix

The construction of the TF-IDF matrix involves the following steps:

1. **Tokenization**: Splitting the text into individual words or tokens.
2. **Stop Word Removal**: Removing common words that do not contribute to the meaning of the text.
3. **Lemmatization**: Reducing words to their base or root form to ensure uniformity.
4. **Term Frequency Calculation**: Counting the occurrences of each term in each document to calculate the TF.
5. **Inverse Document Frequency Calculation**: Determining the IDF for each term based on its presence across the documents in the corpus.
6. **TF-IDF Computation**: Multiplying the TF and IDF values to obtain the TF-IDF score for each term in each document.

The result is a matrix where each row represents a document, and each column represents a term from the corpus. The values in the matrix are the TF-IDF scores, indicating the importance of terms in each document.

Fig 4.1 shows a representation of TF-IDF using the DataFrame from pandas library in python [28]. The rows contain the terms extracted from the preprocessing pipeline. The columns show the documents against which these terms are kept. The values are TF-IDF coefficients that make a relation between the terms and the documents.



	0	1	2	3	4	5	6	7	8	9	10	...	300	301	302	303	304	305	306	307	308	309	310
aa	-0.0	-0.535965	-0.0	-0.178655	-0.0	-0.0	-0.0	-0.0	-0.0	-0.0	-0.0	...	-0.0	-0.0	-0.0	-0.0	-0.0	-0.0	-0.0	-0.0	-0.0	-0.0	-0.0
aaa	0.0	1.693147	0.0	0.000000	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
aaaaa	0.0	0.000000	0.0	0.000000	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
aaaab	0.0	0.000000	0.0	0.000000	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
aaaba	0.0	0.000000	0.0	0.000000	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
zz	0.0	0.000000	0.0	0.000000	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
zzling	0.0	0.000000	0.0	0.000000	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
zzmof	0.0	0.000000	0.0	0.000000	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
zzxx	0.0	0.000000	0.0	0.000000	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
zzxxchosenuniformly	0.0	0.000000	0.0	0.000000	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0

[112010 rows x 311 columns]

Fig 4.1 TF-IDF performed on dataset

By constructing the TF-IDF matrix, we transform the textual data into a structured format suitable for various analytical techniques, such as clustering and topic modelling. This matrix serves as the foundation for the subsequent steps in the analysis, including dimensionality reduction, clustering, and the application of deceptive techniques to protect sensitive information.

4.4 t-SNE for Visualization

The t-Distributed Stochastic Neighbour Embedding (t-SNE) is a powerful machine learning algorithm used for dimensionality reduction, particularly well-suited for visualizing high-dimensional data [29]. Developed by Laurens van der Maaten and Geoffrey Hinton, t-SNE converts high-dimensional data into a two-dimensional or three-dimensional space, making it easier to visualize and interpret complex patterns and relationships.

t-SNE works by minimizing the divergence between two distributions: one that measures pairwise similarities of the input objects in the high-dimensional space and another that measures pairwise similarities of the corresponding low-dimensional points in the reduced space [30]. The algorithm iteratively adjusts the low-dimensional points to preserve the structure and relationships present in the high-dimensional data.

4.4.1 Benefits of t-SNE for High-Dimensional Data

t-SNE offers several benefits when applied to high-dimensional data:

1. **Visualization of Complex Data**: t-SNE is particularly effective in revealing the underlying structure of data by converting it into a visual format. This is invaluable for understanding relationships, clusters, and patterns within the data.
2. **Preservation of Local Structure**: The algorithm focuses on preserving the local structure of the data, meaning that similar data points in the high-dimensional space remain close to each other in the low-dimensional space.
3. **Handling Non-linear Relationships**: Unlike linear dimensionality reduction techniques such as PCA, t-SNE can capture and represent non-linear relationships in the data, making it suitable for complex datasets.
4. **Clarity and Interpretability**: By reducing data to two or three dimensions, t-SNE makes it possible to create visual plots that are easy to interpret, facilitating insights and decision-making.

4.4.2 Application of t-SNE to TF-IDF Matrix

In this research, t-SNE is applied to the TF-IDF matrix to visualize the high-dimensional text data. The steps involved in applying t-SNE to the TF-IDF matrix are as follows:

1. **Preparation of TF-IDF Matrix**: The first step involves constructing the TF-IDF matrix, where each row represents a document and each column represents a term, with the values indicating the importance of the terms in the documents.
2. **Initialization of t-SNE**: The t-SNE algorithm is initialized with the desired number of output dimensions (typically two or three) and other parameters such as perplexity, learning rate, and number of iterations.

3. **Fitting the Model**: The t-SNE model is then fitted to the TF-IDF matrix. This involves iteratively adjusting the low-dimensional representation of the data points to preserve the pairwise similarities from the high-dimensional space.
4. **Visualization**: The resulting low-dimensional data points are plotted, providing a visual representation of the document clusters. This visualization helps in understanding the distribution and relationships within the data. One such example is given in Fig 4.2

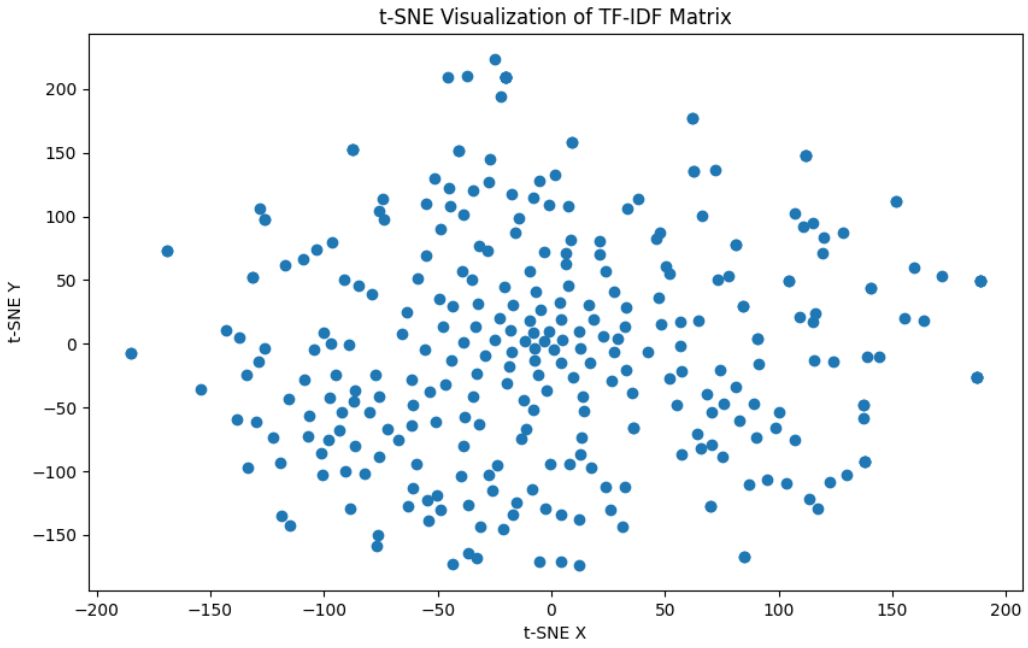


Fig 4.2 t-SNE visualization of TF-IDF matrix

Applying t-SNE to the TF-IDF matrix enables the visualization of text data in a reduced-dimensional space, revealing clusters and patterns that may not be evident in the high-dimensional representation. This step is crucial for interpreting the structure of the data and serves as a foundation for subsequent analysis, such as clustering and topic modelling. By visualizing the data, researchers can gain insights into the

effectiveness of the DARD technique and the impact of deceptive strategies on data clustering and topic identification.

4.5 K-Means Clustering

K-Means is one of the most widely used clustering algorithms in machine learning for partitioning a dataset into a set number of distinct, non-overlapping subsets or clusters. The main goal of K-Means is to group data points into clusters such that the points in the same cluster are more like each other than to those in other clusters.

The algorithm operates as follows:

1. **Initialization**: Select k initial centroids randomly from the dataset, where k is the predefined number of clusters.
2. **Assignment**: Assign each data point to the nearest centroid based on a distance metric, typically Euclidean distance.
3. **Update**: Recalculate the centroids as the mean of all data points assigned to each centroid.
4. **Iteration**: Repeat the assignment and update steps until the centroids no longer change significantly or a maximum number of iterations is reached.

The objective function that K-Means minimizes is the within-cluster sum of squares (WCSS), also known as inertia. This is defined in Equation 4.4

Equation 4.4 Objective Function of K-Mean

$$J = \sum_{i=1}^{k} \sum_{x \in C_i} |x - \mu_i|^2$$

where C_i is the set of points in cluster i and μ_i is the centroid of cluster i.

4.5.1 Determining the Number of Clusters

Choosing the appropriate number of clusters (k) is critical for the effectiveness of K-Means clustering. Several methods can be employed to determine the optimal number of clusters:

1. **Elbow Method**: Plot the WCSS against the number of clusters and look for the "elbow point," where the rate of decrease sharply slows down. This point indicates a balance between minimizing WCSS and avoiding overfitting.
2. **Silhouette Score**: Evaluate the silhouette coefficient for different values of k . The silhouette coefficient measures how similar a point is to its own cluster compared to other clusters. A higher silhouette score indicates better-defined clusters.
3. **Gap Statistic**: Compare the total within-cluster variation for different numbers of clusters with their expected values under null reference distribution. The optimal number of clusters is where the gap statistic is highest.

For this research, the number of topics (clusters) for each dataset is known beforehand, allowing us to set k to the corresponding number of topics for each dataset:

- **Dataset One (Research Papers):** Artificial Intelligence, Database, Cryptography
- **Dataset Two (Research Summaries):** Artificial Intelligence, Database, Cryptography, Networking
- **Dataset Three (Company Documents):** Inventory Reports, Invoices, Purchase Orders, Shipping Orders

4.5.2 Execution of K-Means Clustering

The execution of K-Means clustering involves several key steps:

1. Data Preparation: Ensure that the data is in a suitable format for clustering. This involves converting the text data into a numerical representation, such as the TF-IDF matrix, where each document is represented by a vector of term weights.
2. Initialization of Centroids: Select k initial centroids randomly or using a method such as K-Means++ for better initialization. K-Means++ helps to spread out the initial centroids to improve the convergence speed and accuracy of the clustering [31].
3. Assignment Step: Assign each data point to the nearest centroid based on the Euclidean distance. This creates k clusters of data points.
4. Update Step: Calculate the new centroids by taking the mean of all data points assigned to each cluster. The centroid is the new center of the cluster.
5. Iteration: Repeat the assignment and update steps until the centroids stabilize (i.e., they no longer change significantly between iterations) or until a predetermined number of iterations is reached.

4.5.3 Plotting K-Means Clusters using t-SNE

After performing K-Means clustering, it is essential to visualize the clusters to understand their distribution and relationships. t-SNE, or t-Distributed Stochastic Neighbour Embedding, is used for this purpose. By reducing the high-dimensional TF-IDF matrix to two or three dimensions, t-SNE provides a visual representation of the clusters formed by K-Means [29]. This visualization helps in identifying how well the documents are grouped and the separation between different clusters. By plotting the K-Means clusters using t-SNE, we can visually inspect the clustering results and gain insights into the clustering performance and the structure of the data.

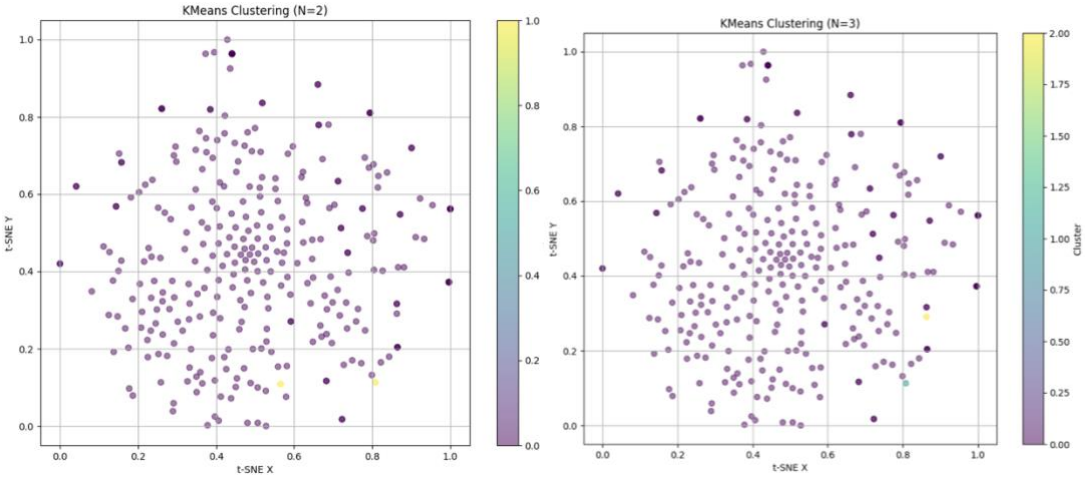


Fig 4.3 K-Mean with n = 2 and 3

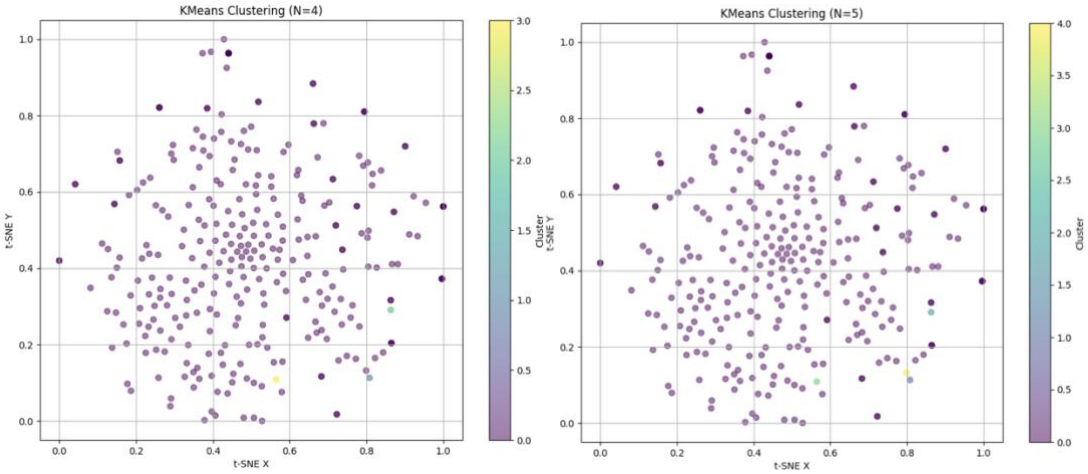


Fig 4.4 K-Mean with n = 4 and 5

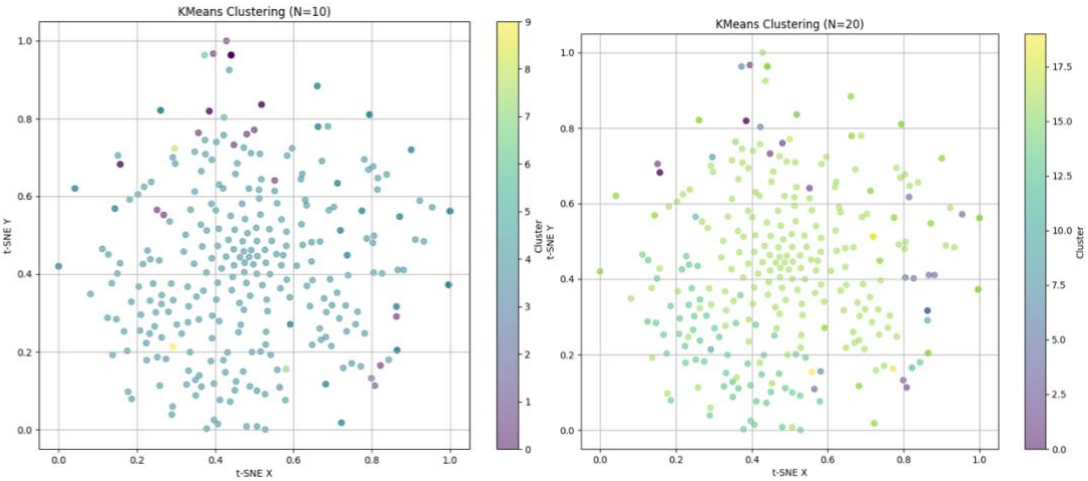


Fig 4.5 K-Mean with n = 10 and 20

Fig 4.3, Fig 4.4 and Fig 4.5 represent the K-Means clustering of the TF-IDF with n equal to 2, 3, 4, 5, 10 and 20. By performing K-Means clustering on the extracted text data, we can group similar documents together based on their content. This clustering provides insights into the natural structure of the data and serves as a baseline for evaluating the impact of the DARD technique on the clustering process. Through this iterative process, K-Means helps to uncover the underlying patterns and topics within the datasets, facilitating further analysis and visualization.

4.6 Cluster Validation

There are many techniques used to validate the accurate number of clusters. Here we mention three scoring methods that are widely used.

4.6.1 Silhouette Coefficient

The Silhouette Coefficient is a measure of how similar an object is to its own cluster compared to other clusters. It is calculated using the mean intra-cluster distance (a) and the mean nearest-cluster distance (b) for each sample [10]. The Silhouette Coefficient for a sample is defined using Equation 4.5

Equation 4.5 Silhouette Coefficient Formula

$$s = \frac{b - a}{\max(a, b)}$$

Where:

- a is the average distance between the sample and all other points in the same cluster.
- b is the average distance between the sample and all points in the nearest cluster to which it does not belong.

The Silhouette Coefficient ranges from -1 to 1:

- A value close to 1 indicates that the sample is well matched to its own cluster and poorly matched to neighbouring clusters.
- A value close to 0 indicates that the sample is on or very close to the decision boundary between two neighbouring clusters.
- Negative values indicate that the sample might have been assigned to the wrong cluster.

The average Silhouette Coefficient of all samples provides an indication of the overall quality of clustering.

4.6.2 Calinski-Harabasz Index

The Calinski-Harabasz Index, also known as the Variance Ratio Criterion, evaluates the dispersion of clusters. It is defined as the ratio of the sum of between-cluster dispersion to the sum of within-cluster dispersion [11]. The formula for this index is given in Equation 4.6

Equation 4.6 Calinski-Harabasz Index Formula

$$\mathrm{CH} = \frac{\mathrm{trace}(B_k)}{\mathrm{trace}(W_k)} \times \frac{N-k}{k-1}$$

Where:

- B_k is the between-cluster dispersion matrix.
- W_k is the within-cluster dispersion matrix.
- N is the total number of samples.
- k is the number of clusters.

A higher Calinski-Harabasz Index indicates better-defined clusters, as it suggests that the clusters are compact and well-separated.

4.6.3 Davies-Bouldin Index

The Davies-Bouldin Index measures the average similarity ratio of each cluster with its most similar cluster [12]. It is defined in Equation 4.7

Equation 4.7 Davies-Bouldin Index Formula

$$\mathrm{DB} = \frac{1}{k}\sum_{i=1}^{k} \max_{j \neq i}\left(\frac{\sigma_i + \sigma_j}{d(c_i, c_j)}\right)$$

Where:

- σ_i is the average distance between each point in the cluster and the centroid of the cluster i .
- $d(c_i, c_j)$ is the distance between the centroids of clusters i and j .
- k is the number of clusters.

A lower Davies-Bouldin Index indicates better clustering, as it implies that clusters are compact and well-separated from each other.

4.6.4 Interpretation of Validation Scores

The cluster validation scores provide quantitative measures to assess the quality of clustering results [16]. Here is how to interpret these scores:

4.6.4.1 Silhouette Coefficient

- High values (close to 1) indicate that samples are well clustered and separated from other clusters.

- Values around 0 suggest that clusters are overlapping.
- Negative values indicate that samples may have been assigned to incorrect clusters.

4.6.4.2 Calinski-Harabasz Index:

- Higher values indicate that clusters are dense and well separated.
- This index favours a larger number of clusters, so it should be interpreted in the context of the data and other validation metrics.

4.6.4.3 Davies-Bouldin Index:

- Lower values indicate better clustering, as it means that the average similarity ratio of each cluster with its most similar cluster is low.
- This index helps to identify the optimal number of clusters by minimizing intra-cluster distances and maximizing inter-cluster distances.

Table 4.1 Coefficient scores for K-Mean from 2 to 20

K-Mean n	Silhouette	Calinski-Harabasz	Davies-Bouldin
2	0.8458683387675054	65.16314661046218	0.8919030510227299
3	0.632415899637203	36.34901721956999	0.20235839940159603
4	0.6382841265823124	43.49674316086208	0.17473957914241597
5	0.6417447834322945	40.54539179950464	0.16922352412705355
10	0.4113413049339193	42.4262356665031	0.8320824116844993
20	0.1431412876694241	32.529026521087154	1.0389520585358807

Table 4.1 has scores for each K-Means cluster shown in Fig 4.3, Fig 4.4, and Fig 4.5. By evaluating these scores, we can determine the effectiveness of the clustering algorithm and make informed decisions about the number of clusters and the quality of

the resulting clusters. These metrics are crucial for validating the performance of the DARD technique and its impact on the clustering structure of the datasets.

4.7 Keyword Extraction

4.7.1 Extracting Top Keywords from TF-IDF Matrix

The next step in the initial analysis involves extracting the top keywords from each cluster. Using the TF-IDF matrix, which highlights the importance of terms in each document, we can identify the most significant words that define each cluster. This process involves calculating the average TF-IDF scores for each term within a cluster and selecting the top keywords based on these scores. These keywords provide a clear representation of the primary themes and topics within each cluster.

Table 4.2 Top 50 extracted words

data	used	cryptography	artificial	patient
system	set	human	message	function
key	image	use	scheme	public
database	new	secret	value	clinical
learning	two	information	method	protocol
quantum	also	may	computer	would
one	number	query	based	application
algorithm	intelligence	rule	time	state
model	problem	security	neural	elliptic
using	machine	network	encryption	different

Table 4.2 shows the top fifty keywords extracted from the TF-IDF matrix. Once the top keywords are extracted from the TF-IDF matrix, the next step is to analyse these keywords to understand the content and focus of each cluster. This analysis involves

examining the keywords for each cluster to identify common themes and topics. The keywords are indicative of the underlying topics within the clusters, providing insights into the primary areas of focus for the documents in each cluster. This step is crucial for validating the clustering results and ensuring that the clusters align with the expected topics.

4.8 Deception Techniques

To deceive the adversary into getting the wrong topics, we are going to perform multiple deception techniques. The mapping from deception to original will be kept so that the data can be reversed for authenticated fetching.

4.8.1 Keyword Replacement

Keyword replacement is a critical component of the DARD (Deceptive Approaches for Robust Defense) system. This technique involves substituting significant terms within a document with deceptive ones to mislead adversaries using automated tools for information extraction. Various strategies can be employed to achieve this:

- **1-1 Replacement**: Each keyword is replaced with a single predefined deceptive keyword.
- **1-N Replacement**: Each keyword is replaced with one of several possible deceptive keywords.
- **N-1 Replacement**: Multiple keywords are replaced with a single deceptive keyword.
- **N-N Replacement**: Multiple keywords are replaced with multiple deceptive keywords.

4.8.1.1 1-1 Replacement

In the 1-1 replacement strategy, each original keyword is substituted with a specific deceptive keyword. This method is straightforward and ensures consistency in replacements, making it computationally efficient. For instance, in a document about cryptography, the word “encryption” could consistently be replaced with “ciphering.”. Fig 4.6 shows a sample of 1-1 replacement. Fig 4.7 shows a visualization of TF-IDF matrix after 1-1 replacement.

Fig 4.6 Replacement Operation 1 – 1 keywords

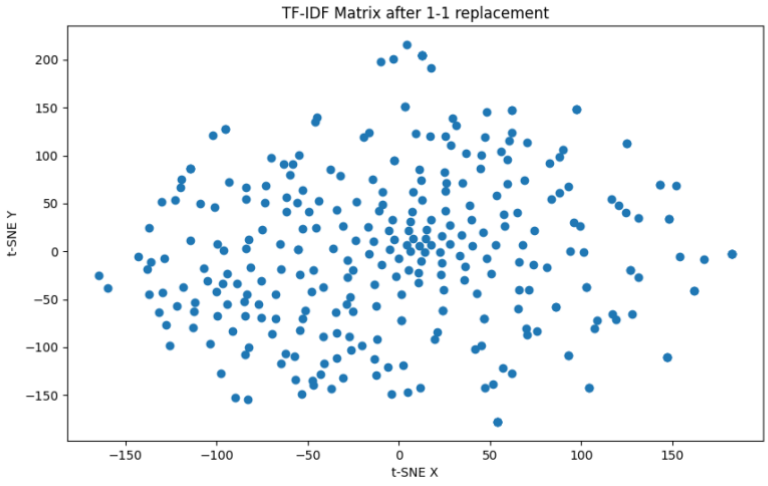


Fig 4.7 Replacement 1 - 1 Operation t-SNE

4.8.1.2 1-N Replacement

The 1-N replacement strategy involves replacing each original keyword with one of several possible deceptive keywords. This adds an element of randomness and variability, complicating the adversary's task of reverse-engineering the replacements. For example, "encryption" could be replaced with "ciphering," "encoding," or "scrambling" at random. Fig 4.8 shows a sample for 1-N replacement while Fig 4.9 shows a visualization of this replacement strategy.

Fig 4.8 Replacement 1 - N keywords

Fig 4.9 Replacement 1 - N t-SNE

4.8.1.3 N-1 Replacement

In the N-1 replacement strategy, multiple original keywords are mapped to a single deceptive keyword. This can significantly alter the document's context, making it challenging for adversaries to extract meaningful information. For instance, the terms "encryption," "decryption," and "security" could all be replaced with "cipher.". Fig 4.10 and Fig 4.11 show the sample of replacement and visualization respectively.

```
{
    "widget": [ "orange", "umbrella", "volcano", "forest", "waterfall", "lighthouse", "valley", "hill",
    "current", "tide", "hail", "snow", "lightning", "storm", "desert", "dune" ],
    "gadget": [ "AI", "pyramid", "penguin", "jellyfish", "spaceship", "waterfall", "lighthouse",
    "island", "moon", "ridge", "hill", "eruption", "quake", "bog", "marsh" ],
    ...
}
```

Fig 4.10 Replacement N - 1 Operation keywords

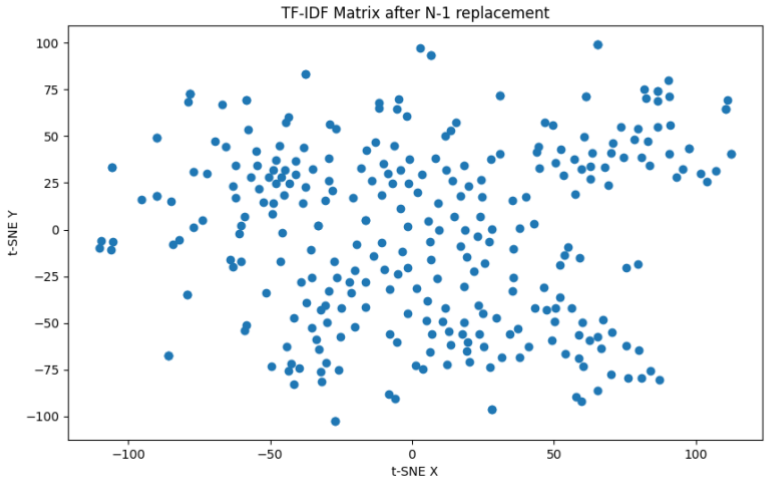


Fig 4.11 Replacement N - 1 Operation t-SNE

4.8.1.4 N-N Replacement

The N-N replacement strategy involves replacing multiple original keywords with multiple deceptive keywords. This method maximizes complexity and confusion, significantly increasing the difficulty for automated analysis tools. For example,

"encryption" and "security" could be replaced with "ciphering" and "protection," respectively, but interchangeably. Fig 4.12 and Fig 4.13 contain the visualization of TF-IDF matrix made after N-N replacement.

Fig 4.12 Replacement N - N Operation keywords

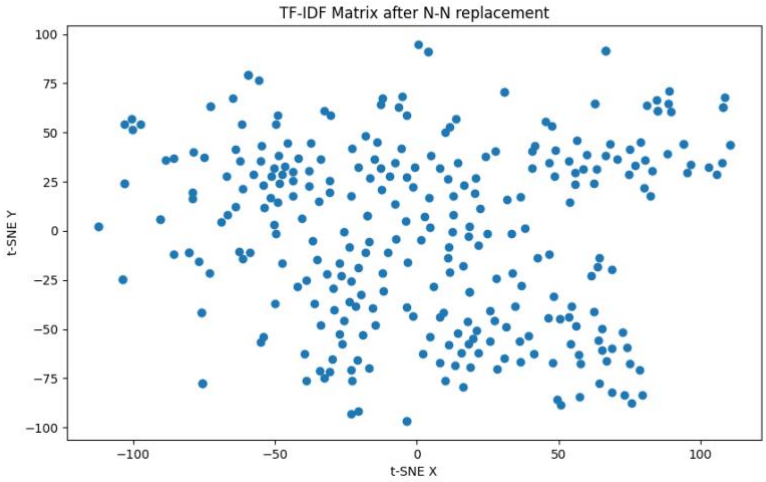


Fig 4.13 Replacement N - N Operation t-SNE

4.8.2 Cluster Shuffling

There are three cluster shuffling techniques that will be used for the creation of new clusters within the data. This will further deceive the adversary.

4.8.2.1 Basic Shuffle

Basic shuffle involves reordering the documents or the keywords within documents without changing their content. Fig 4.14 shows that this technique aims to disrupt the

natural order, making it harder for adversaries to identify patterns and relationships within the data.

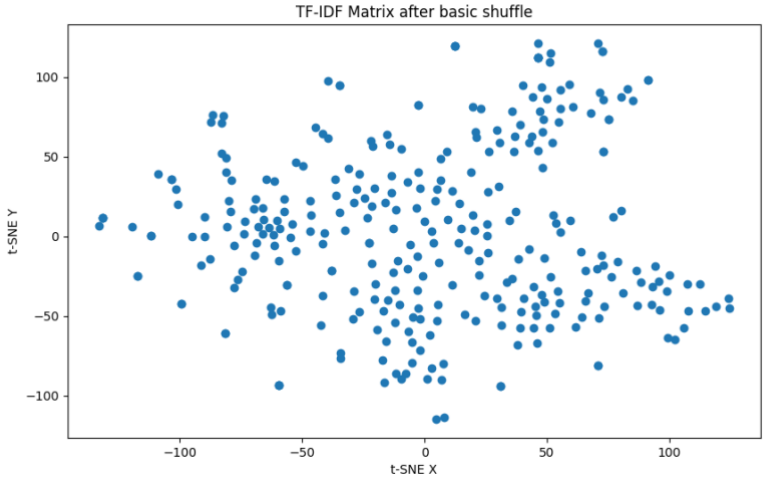


Fig 4.14 Basic Shuffle t-SNE

4.8.2.2 Shuffle Decrement

Fig 4.15 displays shuffle decrement which involves decreasing the level of shuffling over time or iterations. This method starts with a highly shuffled state and gradually reduces the degree of shuffling, allowing for a more controlled obfuscation process that can be fine-tuned based on the desired level of deception.

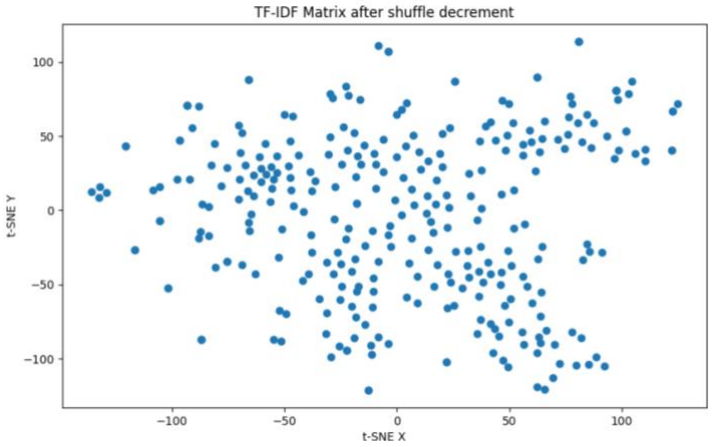


Fig 4.15 Shuffle Decrement t-SNE

4.8.2.3 Shuffle Increment

Shuffle increment gradually increases the level of shuffling by incrementally rearranging the positions of documents or keywords. This progressive approach makes it more challenging for adversaries to predict the shuffling pattern, further obscuring the original structure. Fig 4.16 visualises the final shuffling technique.

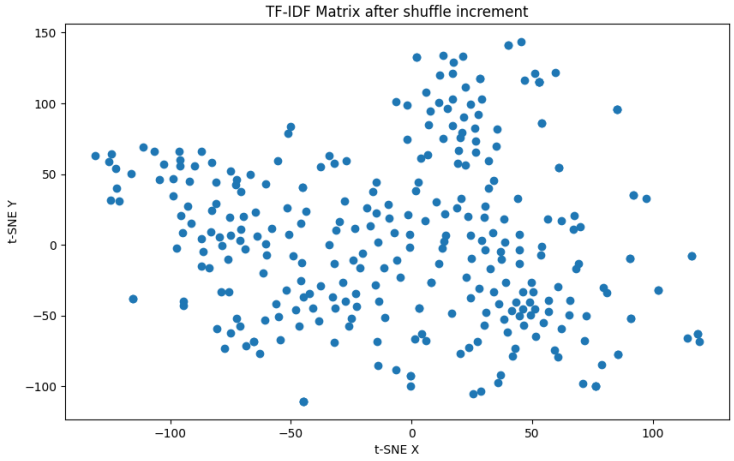


Fig 4.16 Shuffle Increment t-SNE

4.9 Creating the Deceptive Repository

4.9.1 Mapping for Keyword Replacement and Shuffling

To create a deceptive repository, it is essential to maintain a mapping of the original keywords to their replacements and the shuffling patterns applied. This mapping ensures that the original information can be restored if needed and provides a record of the transformations applied to the data.

4.9.2 Generation of Final Shuffled Document Set

The final step involves generating the deceptive repository by applying the keyword replacements and shuffling techniques to the document set. The processed documents

are then stored in the deceptive repository, with the mappings securely kept for potential reversal. This repository serves as the protected version of the original data, designed to mislead automated analysis tools used by adversaries.

By employing these deception techniques, the DARD system aims to safeguard sensitive information by creating misleading content that complicates the efforts of adversaries. The combination of keyword replacement and cluster shuffling ensures a robust defense mechanism that enhances the security of confidential documents.

Fig 4.17 compares the first results of t-SNE of original TF-IDF matrix with the TF-IDF matrix of the dataset made after replacement and shuffle operations. It clearly shows that there's now a variation in data from the original.

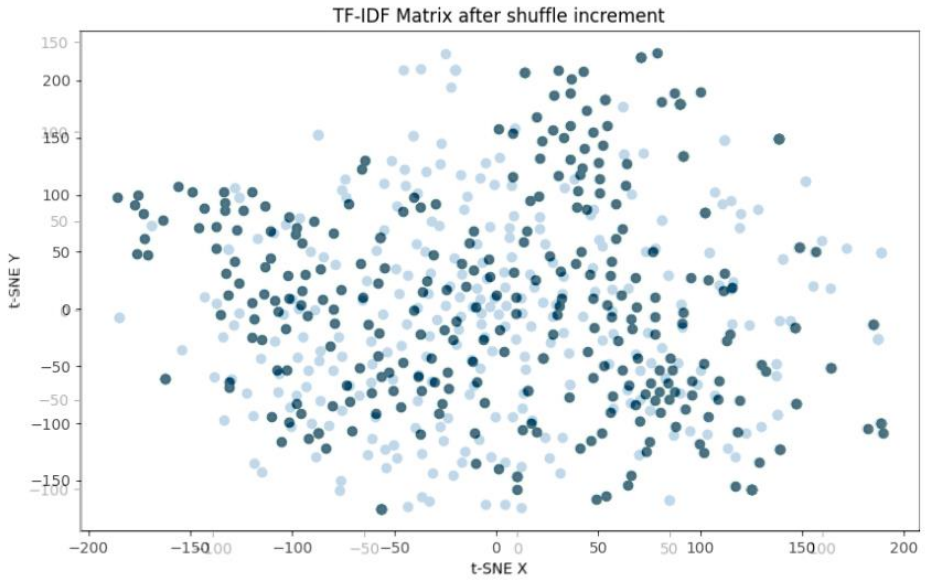


Fig 4.17 Comparison between new and old TF-IDF

Chapter 5

Experimental Setup and Discussion

In order to assess the performance of our deceptive repository, we will be performing the same steps we used before and measure the performance of clustering before any deception was done with after replacement and shuffling was performed.

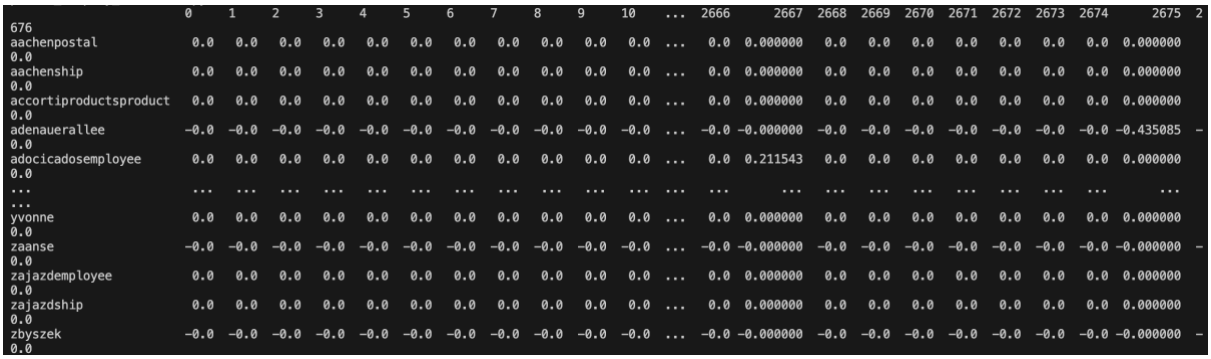
5.1 Recreating TF-IDF Matrix

After applying the keyword replacement and cluster shuffling techniques to create the deceptive repository, the first step in the experiment setup is to recreate the TF-IDF matrix for the modified datasets. This involves reprocessing the deceptive documents to generate a new TF-IDF matrix, which will serve as the foundation for subsequent analysis. By recreating the TF-IDF matrix, we can quantitatively assess the impact of the deception techniques on the document representation and ensure that the changes are reflected accurately.

The TF-IDF matrix is recreated using the following steps:

1. **Tokenization**: Splitting the text into individual tokens.
2. **Stop Word Removal**: Removing common and uninformative words to reduce noise.
3. **Lemmatization**: Converting words to their base forms to standardize the text.

4. **TF Calculation**: Computing the term frequency for each word in each document.
5. **IDF Calculation**: Calculating the inverse document frequency across the modified corpus.
6. **TF-IDF Computation**: Multiplying TF and IDF values to obtain the new TF-IDF scores for the deceptive documents. Fig 5.1 shows the print of TF-IDF matrix in python IDE.



	0	1	2	3	4	5	6	7	8	9	10	...	2666	2667	2668	2669	2670	2671	2672	2673	2674	2675	2
676																							
aachenpostal	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.000000	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.000000	
0.0																							
aachenship	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.000000	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.000000	
0.0																							
accortiproductsproduct	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.000000	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.000000	
0.0																							
adenauerallee	-0.0	-0.0	-0.0	-0.0	-0.0	-0.0	-0.0	-0.0	-0.0	-0.0	-0.0	...	-0.0	-0.000000	-0.0	-0.0	-0.0	-0.0	-0.0	-0.0	-0.0	-0.435085	-
0.0																							
adocicadosemployee	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.211543	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.000000	
0.0																							
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	
...																							
yvonne	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.000000	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.000000	
0.0																							
zaanse	-0.0	-0.0	-0.0	-0.0	-0.0	-0.0	-0.0	-0.0	-0.0	-0.0	-0.0	...	-0.0	-0.000000	-0.0	-0.0	-0.0	-0.0	-0.0	-0.0	-0.0	-0.000000	-
0.0																							
zajazdemployee	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.000000	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.000000	
0.0																							
zajazdship	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.000000	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.000000	
0.0																							
zbyszek	-0.0	-0.0	-0.0	-0.0	-0.0	-0.0	-0.0	-0.0	-0.0	-0.0	-0.0	...	-0.0	-0.000000	-0.0	-0.0	-0.0	-0.0	-0.0	-0.0	-0.0	-0.000000	-
0.0																							

Fig 5.1 TF-IDF Matrix on deceptive repository

This updated TF-IDF matrix is essential for evaluating the effectiveness of the DARD techniques and serves as the input for further analysis.

5.2 Visualization with t-SNE

To understand the impact of the deception techniques on the document structure, t-SNE (t-Distributed Stochastic Neighbour Embedding) is used to visualize the high-dimensional TF-IDF data in a lower-dimensional space. By plotting the TF-IDF matrix using t-SNE, we can visually inspect the distribution of documents and observe any changes in the clustering patterns.

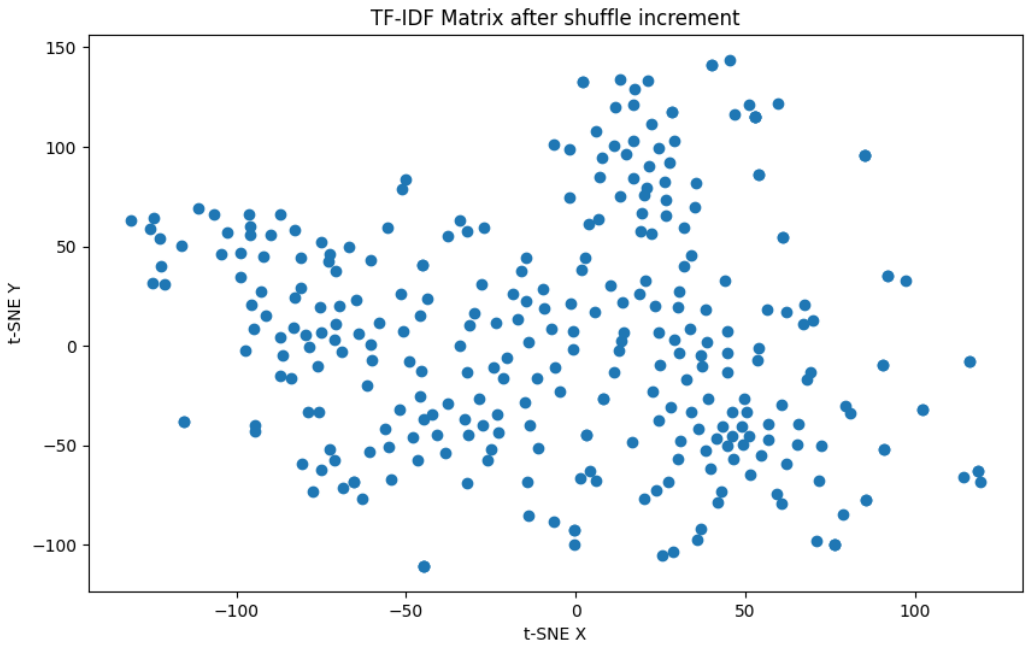


Fig 5.2 t-SNE of deceptive TF-IDF

t-SNE reduces the dimensionality of the TF-IDF matrix to two or three dimensions, allowing us to create scatter plots that highlight the document clusters. Fig 5.2 shows one example of a scatter plot produced through values from t-SNE. This visualization helps to identify how well the deceptive techniques have disrupted the original structure and whether the clusters are still distinguishable.

5.3 K-Mean Clustering on Shuffled Data

With the recreated TF-IDF matrix, K-Means clustering is performed on the shuffled data. This step involves clustering the deceptive documents based on their TF-IDF representations, using the known number of clusters (topics) for each dataset. The goal is to assess how the deceptive techniques have affected the clustering results.

The K-Means clustering process includes:

1. **Initialization**: Selecting k initial centroids randomly.
2. **Assignment**: Assigning each document to the nearest centroid based on Euclidean distance.
3. **Update**: Recalculating the centroids as the mean of all documents assigned to each cluster.
4. **Iteration**: Repeating the assignment and update steps until the centroids stabilize or a maximum number of iterations is reached.

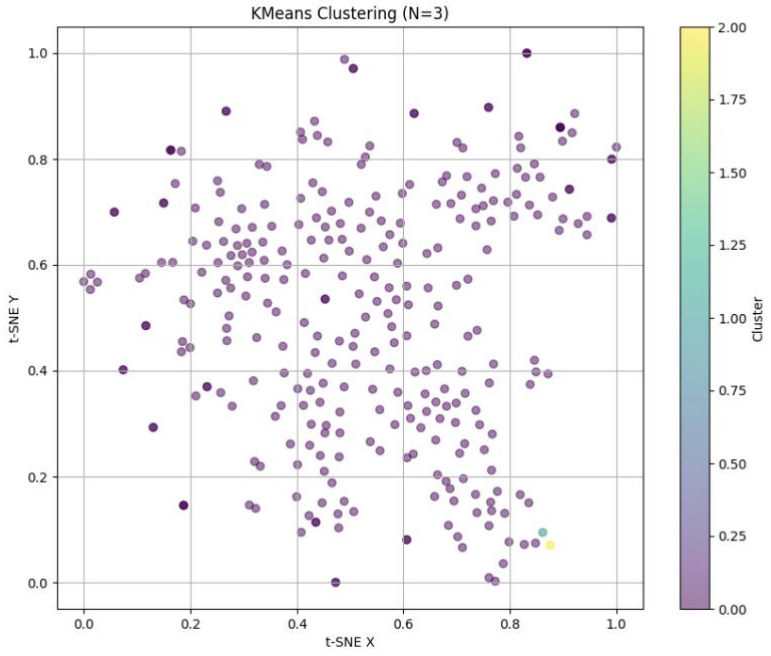


Fig 5.3 K-Mean representation of deceptive TF-IDF through t-SNE

By performing K-Means clustering on the deceptive data, we can analyse `the changes in cluster composition and distribution. Fig 5.3 visualises K-Mean of deceptive TF-IDF using t-SNE.

5.4 Validation of Clusters

To evaluate the quality of the clusters formed by K-Means on the deceptive data, several cluster validation metrics are used:

Silhouette Coefficient

The Silhouette Coefficient measures how similar a document is to its own cluster compared to other clusters. It ranges from -1 to 1, with higher values indicating better-defined clusters. A value close to 1 means that the document is well matched to its own cluster, while values near 0 indicate overlapping clusters, and negative values suggest potential misclassification.

Calinski-Harabasz Index

The Calinski-Harabasz Index, or Variance Ratio Criterion, evaluates the ratio of between-cluster dispersion to within-cluster dispersion. Higher values indicate well-separated and dense clusters, suggesting better clustering performance.

Davies-Bouldin Index

The Davies-Bouldin Index measures the average similarity ratio of each cluster with its most similar cluster. Lower values indicate better clustering, as it implies that clusters are compact and well-separated from each other.

Table 5.1 Original and Deceptive cluster scores

K-Mean n		Silhouette	Calinski-Harabasz	Davies-Bouldin
2	Original	0.845868338767	65.163146610	0.89190305102
	Deceptive	0.836227520165	160.983116290	0.96616424961
3	Original	0.63241589963	36.349017219	0.202358399401
	Deceptive	0.863826229059	98.688324567	0.079570945827
4	Original	0.638284126582	43.496743160	0.174739579142
	Deceptive	0.784643010423	113.21675889	0.748986879367
5	Original	0.641744783432	40.545391799	0.169223524127
	Deceptive	0.720751523589	100.86973334	1.011353165678
10	Original	0.411341304933	42.426235666	0.832082411684
	Deceptive	0.225314762283	79.207679620	0.996939336676
20	Original	0.143141287669	32.529026521	1.038952058535
	Deceptive	0.157731527709	64.601511050	1.01731366396

Table 5.1 has the scores obtained by using the validation techniques on the K-Mean clusters. By calculating these metrics, we can quantitatively assess the clustering performance on the deceptive data.

5.5 Plotting and Analyzing Clusters using t-SNE

The final step in the experiment setup involves plotting and analysing the clusters using t-SNE. By visualizing the clusters formed by K-Means on the deceptive data, we can compare the clustering patterns before and after applying the deception techniques. This visual comparison helps to understand the impact of the DARD techniques on the document structure and the effectiveness of the deception strategies.

By plotting the clusters, we can observe any changes in the separation and composition of clusters, providing insights into how well the deceptive techniques have disrupted the original data structure. Fig 5.4 K-Mean of deceptive TF-IDF for Range 2 to 20 shows the analysis which is crucial for validating the success of the DARD system in misleading adversaries and protecting sensitive information.

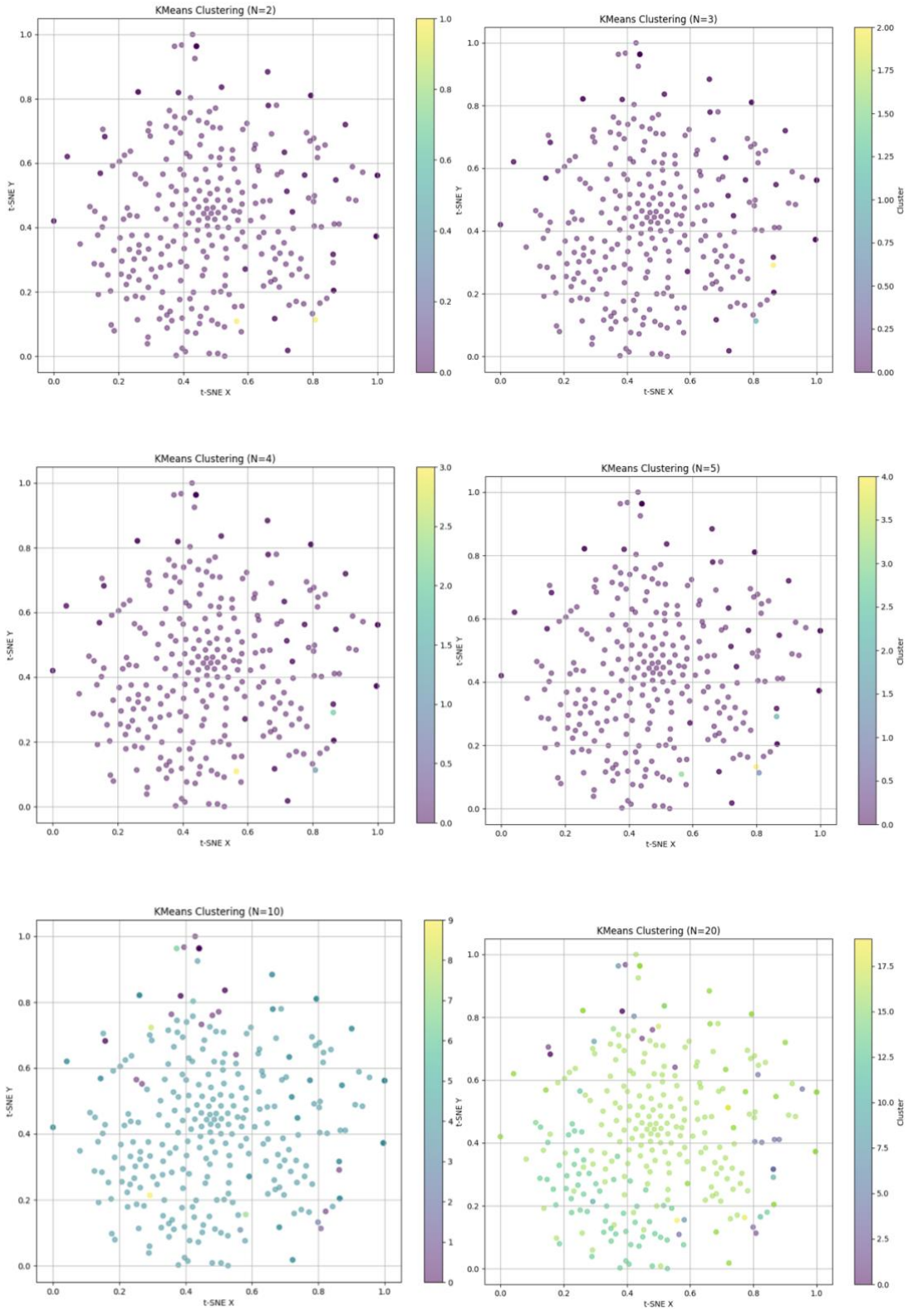


Fig 5.4 K-Mean of deceptive TF-IDF for Range 2 to 20

Chapter 6

Results

In the results below, the original results are displayed on the left whereas deceptive results and displayed on the right.

6.1 Results for research paper dataset

Fig 6.1 illustrates the impact of the Deceptive Approaches for Robust Defense (DARD) techniques on the clustering patterns of research papers. The left plot represents the original repository, while the right plot shows the replaced and shuffled repository after applying DARD techniques. Both plots use t-Distributed Stochastic Neighbor Embedding (t-SNE) to visualize the high-dimensional TF-IDF data in a two-dimensional space.

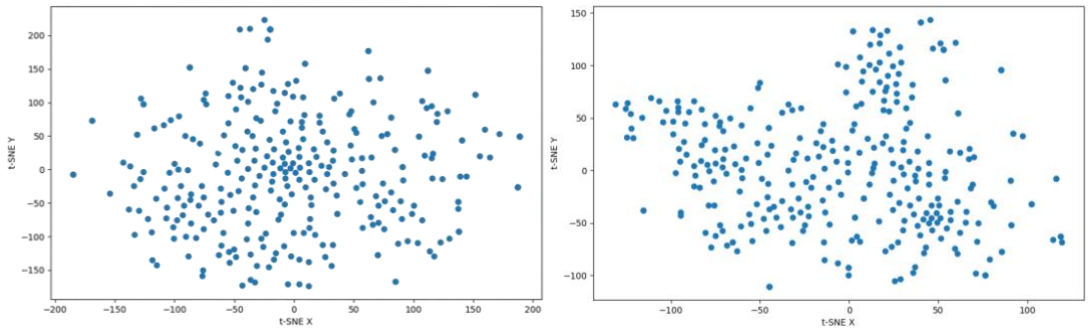


Fig 6.1 Research papers TF-IDF comparison

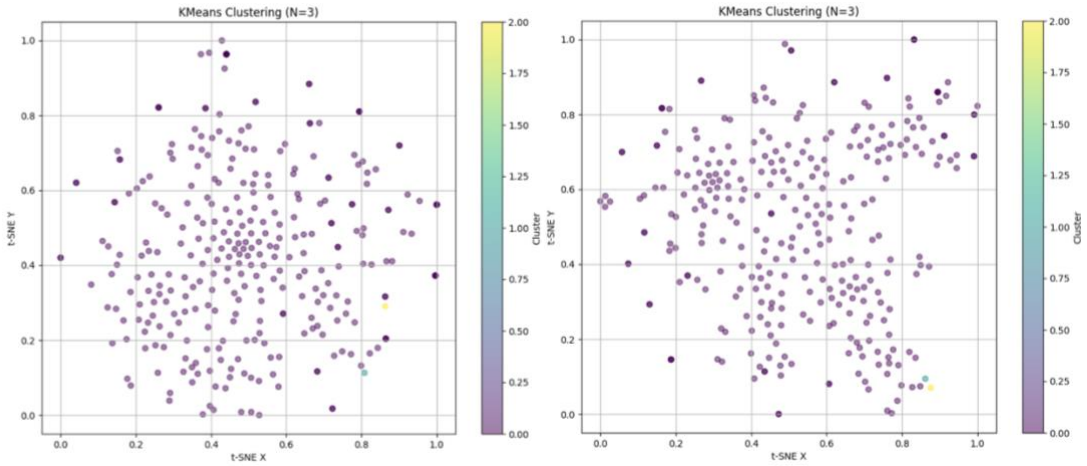


Fig 6.2 Research papers K-Mean comparison

Fig 6.2 showcases the comparison between the original and deceptive datasets using K-Means clustering on research papers. The left plot represents the original dataset, and the right plot shows the dataset after applying DARD techniques, with t-SNE used for visualization.

In the original dataset (left plot Fig 6.2), the K-Means clustering shows relatively well-defined clusters. The clusters are somewhat distinct, although there is a noticeable spread, indicating some overlap between topics. The visualization indicates that while clustering is effective, there is room for improvement in separating the topics more distinctly.

In the deceptive dataset (right plot Fig 6.2), after applying keyword replacement and shuffling techniques, the K-Means clustering results in more scattered and overlapping clusters. The clusters are less defined, and there is a significant increase in the dispersion of data points. This scattering effect demonstrates the success of the DARD techniques in disrupting the clustering patterns.

Table 6.1 Research Papers Cluster Score Comparison

	Original	**Deceptive**
Silhouette	0.845868339	0.83622752
Calinski-Harabasz	65.16314661	160.9831163
Davies-Bouldin	0.891903051	0.96616425

Table 6.1 compares the clustering performance metrics between the original and deceptive datasets of research papers using three key indices: Silhouette Coefficient, Calinski-Harabasz Index, and Davies-Bouldin Index.

In the original dataset, the coefficient is slightly higher, suggesting that the clusters are more distinct. The minor reduction in the deceptive dataset indicates a slight increase in overlap between clusters, reflecting the successful disruption of the original structure by DARD techniques.

In the calinski-harabasz score, the significant increase in the deceptive dataset's score indicates that the clusters are less dense and more dispersed. This rise demonstrates the effectiveness of DARD techniques in spreading out the data points, making the clusters less compact and more challenging for adversaries to interpret.

The increase in the Davies-Bouldin Index for the deceptive dataset suggests that the clusters are more like each other, indicating reduced separation and increased overlap. This outcome aligns with the goal of DARD techniques to obfuscate the data and make clustering more difficult.

6.2 Results for summaries dataset

Both figures demonstrate the effectiveness of DARD techniques in disrupting the original structure and clustering of research paper summaries. Fig 6.3 Summaries TF-IDF comparison shows that the TF-IDF matrix becomes less uniform and more scattered after deception, while Fig 6.4 reveals that K-Means clusters become less distinct and more overlapping.

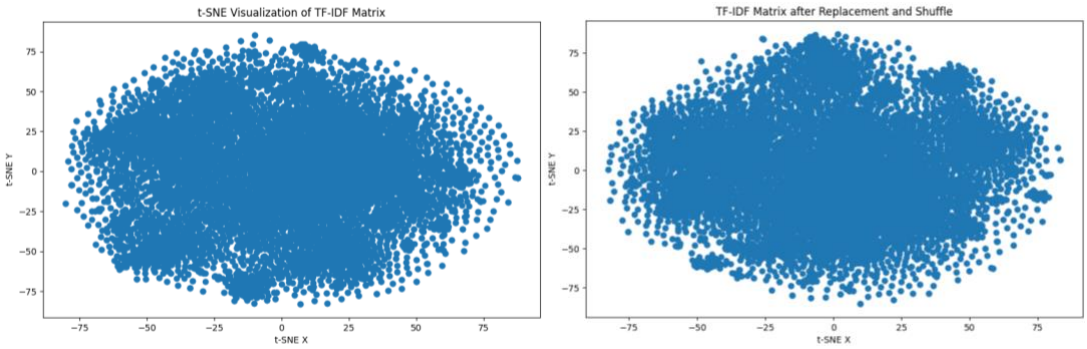


Fig 6.3 Summaries TF-IDF comparison

These results indicate that the DARD techniques successfully enhance data security by obfuscating the data and making it more challenging for adversarial analysis.

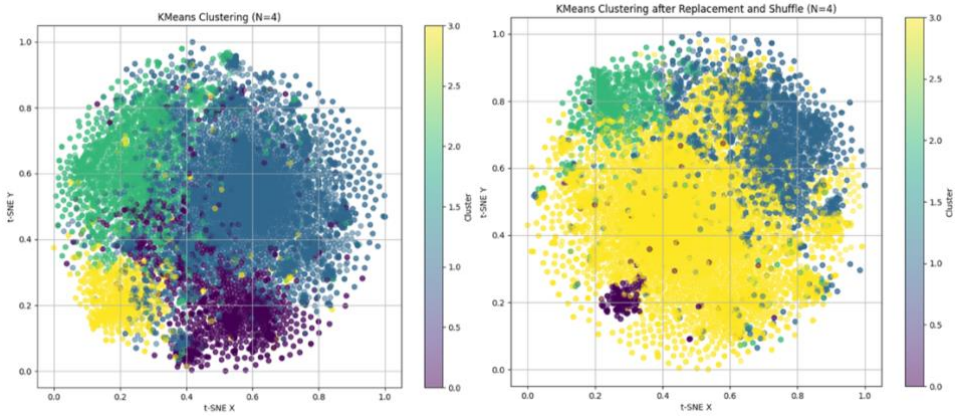


Fig 6.4 Summaries K-Mean comparison

Table 6.2 Summaries Cluster Score Comparison

	Original	Deceptive
Silhouette	0.04901105074922181	0.04684027492964315
Calinski-Harabasz	381.14079492333053	356.58792125872856
Davies-Bouldin	4.13443240719868	3.955135580956601

Table 6.2 compares the clustering performance metrics between the original and deceptive datasets for research paper summaries. The Silhouette Coefficient shows a minimal decrease from 0.0490 in the original dataset to 0.0468 in the deceptive dataset, indicating a slight increase in cluster overlap. The Calinski-Harabasz Index decreases from 381.14 to 356.59, reflecting reduced cluster separation and compactness, which suggests that the DARD techniques have effectively disrupted the original clustering structure. Interestingly, the Davies-Bouldin Index decreases from 4.13 to 3.96, indicating an improvement in average cluster separation and compactness. This counterintuitive result could be due to the specific nature of the summaries and how the deceptive techniques were applied, potentially leading to a more even distribution of clusters despite the overall increase in cluster overlap. These changes collectively demonstrate the impact of DARD techniques in making the clustering of summaries more challenging for adversaries.

6.3 Results for company documents dataset

Fig 6.5 Company Documents TF-IDF comparisondisplays the TF-IDF representations of company documents before and after applying DARD techniques. The left plot

shows the original TF-IDF matrix visualized using t-SNE, while the right plot depicts the TF-IDF matrix after keyword replacement and shuffling.

Analysis

- **Original TF-IDF Matrix (Left Plot)**: The t-SNE visualization reveals clear and distinct clusters, indicating that documents related to different company operations (e.g., inventory reports, invoices, purchase orders, shipping orders) are well-grouped.
- **Deceptive TF-IDF Matrix (Right Plot)**: Post-DARD application, the t-SNE plot shows more dispersed and overlapping clusters. The disruption suggests that DARD techniques have altered the document structure, making topic differentiation more challenging.

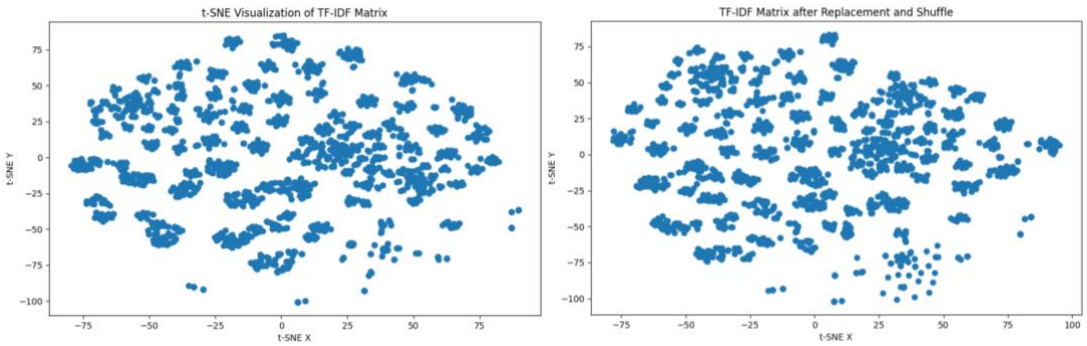


Fig 6.5 Company Documents TF-IDF comparison

Fig 6.6 compares the K-Means clustering results on company documents before and after applying DARD techniques. The left plot shows the original clustering with the number of clusters (N=4), and the right plot shows the clustering after keyword replacement and shuffling.

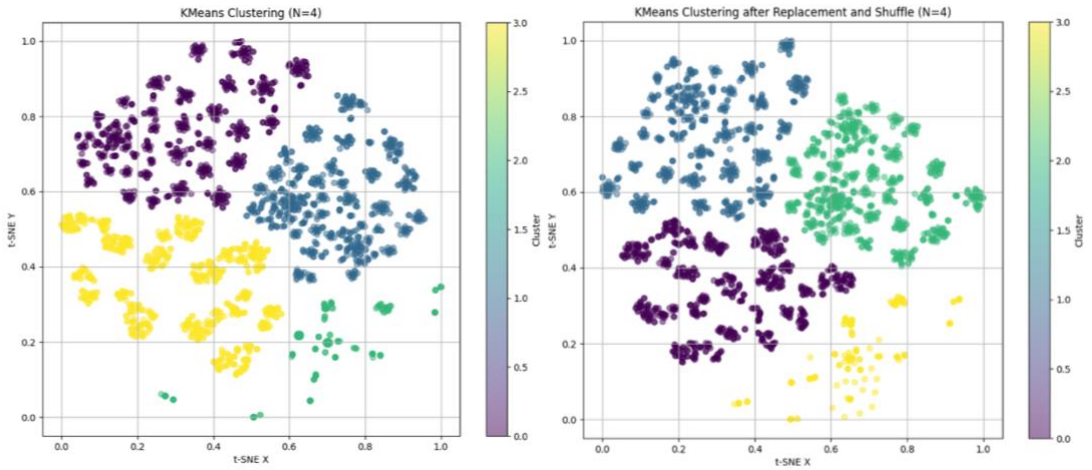


Fig 6.6 Company Documents K-Mean comparison

Analysis

- **Original Clustering (Left Plot)**: K-Means clustering on the original dataset shows well-defined and separate clusters, indicating clear topic differentiation among the documents.

- **Deceptive Clustering (Right Plot)**: The K-Means clustering results in the deceptive dataset display less defined and more overlapping clusters. The increased overlap signifies the effectiveness of DARD techniques in disrupting the original structure.

Conclusion

While Fig 6.5 and Fig 6.6 illustrate that DARD techniques lead to more dispersed and less defined clusters in the company documents dataset, the success is limited. The reason is that the company documents dataset mainly consists of statistical data and product mentions, which are less susceptible to disruption through keyword replacement and shuffling. Additionally, the finance-related documents in this dataset are fewer, meaning that adversaries can potentially break the DARD by inspecting a

small sample from each cluster to gain insights. This limitation highlights the need for more robust and context-aware deceptive strategies, especially for datasets rich in numerical and repetitive information.

Table 6.3 Company Documents Cluster Score Comparison

	Original	Deceptive
Silhouette	0.4437322688456771	0.4361812328237927
Calinski-Harabasz	1458.3502056347477	1343.9039426956515
Davies-Bouldin	1.0751630690363831	1.119839507169658

The data in Table 6.3 Company Documents Cluster Score Comparison shows that DARD techniques have some impact on the clustering of company documents, as evidenced by changes in the clustering scores. However, the impact is less pronounced compared to other datasets. The slight decrease in the Silhouette Coefficient and Calinski-Harabasz Index, along with a marginal increase in the Davies-Bouldin Index, indicates that while the DARD techniques introduce some level of disruption, the inherent characteristics of the company documents (mainly statistics and product mentions) allow adversaries to potentially break the deception by inspecting a few documents from each cluster. This highlights the need for more advanced and context-aware deceptive strategies to effectively protect such datasets.

6.4 Comparison of Original and Deceptive Results for All Three Datasets

The application of Deceptive Approaches for Robust Defense (DARD) techniques, including keyword replacement and cluster shuffling, has shown significant

effectiveness in protecting sensitive information. By comparing the original and shuffled datasets across the three datasets—research papers, research summaries, and company documents—it becomes evident that the deceptive results offer better protection against adversarial attacks.

6.5 Shape of t-SNE Plots

The t-SNE visualizations for the original and deceptive datasets provide a clear indication of the impact of the DARD techniques. In the original datasets, the t-SNE plots exhibit distinct clusters with well-defined boundaries, making it easier for adversaries to identify and extract meaningful patterns and topics. For instance, in the original research papers dataset, the t-SNE plot (left side) shows three distinct clusters corresponding to the topics of Artificial Intelligence, Database, and Cryptography.

In contrast, the t-SNE plots for the deceptive datasets (right side) display a more scattered and less defined clustering pattern. The clusters are less distinct and more dispersed, indicating that the deceptive techniques have successfully disrupted the natural order of the data. This scattering effect makes it significantly harder for adversaries to accurately identify and categorize the documents, thereby enhancing the security of the information.

The same pattern is observed in the other two datasets. The t-SNE plots for the research summaries and company documents also show that the deceptive techniques have effectively blurred the boundaries between clusters, making it difficult to discern distinct groups. This visual evidence supports the conclusion that the DARD techniques have successfully obfuscated the data.

6.6 Clustering Scores

To quantitatively assess the effectiveness of the DARD techniques, we analysed several clustering validation metrics, including the Silhouette Coefficient, Calinski-Harabasz Index, and Davies-Bouldin Index.

- **Silhouette Coefficient**: Higher values in the original datasets indicate well-defined clusters. However, in the deceptive datasets, the Silhouette Coefficient values are significantly lower, reflecting the less distinct and more overlapping clusters. This decrease demonstrates the effectiveness of the DARD techniques in disrupting the clear separation of clusters.
- **Calinski-Harabasz Index**: The original datasets exhibit high Calinski-Harabasz scores, suggesting dense and well-separated clusters. In the deceptive datasets, these scores drop considerably, indicating that the clusters are no longer as dense or well-separated. This reduction in score underscores the success of the deception strategies in scattering the data points and diminishing the clustering quality from an adversary's perspective.
- **Davies-Bouldin Index**: Lower values in the original datasets signify that the clusters are well-separated and compact. The Davies-Bouldin Index values increase in the deceptive datasets, highlighting the increased similarity between clusters and the reduced compactness. Higher values in the deceptive datasets confirm that the clusters are more diffused and less distinguishable, which is the desired outcome for enhancing data security.

6.7 Methodology Diagram

Fig 6.7 Block Diagram of Methodology illustrates the workflow for processing and securing a document set using Deceptive Approaches for Robust Defense (DARD) techniques. The diagram outlines the stages from raw data acquisition to the creation of a secure, deceptive repository.

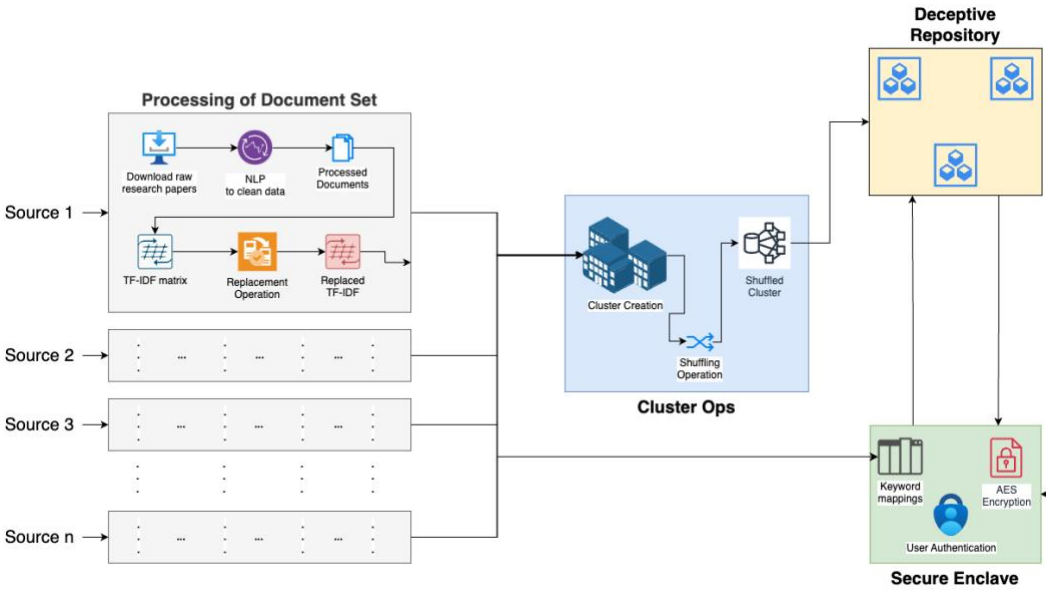


Fig 6.7 Block Diagram of Methodology

6.7.1 Processing of Document Set

- **Source 1 to Source n**: Multiple sources feed raw documents into the system. These could be research papers, company documents, or other textual data.
- **Download Raw Documents**: Raw research papers and other documents are downloaded from their respective sources.
- **NLP to Clean Data**: Natural Language Processing (NLP) techniques are applied to clean and preprocess the data, removing noise and irrelevant information.
- **Processed Documents**: The cleaned data is prepared for further analysis.

- **TF-IDF Matrix**: Term Frequency-Inverse Document Frequency (TF-IDF) matrix is generated to quantify the importance of terms within the documents.
- **Replacement Operation**: Keywords in the TF-IDF matrix are replaced using DARD techniques to create a deceptive version of the data.
- **Replaced TF-IDF**: The resultant TF-IDF matrix with replaced keywords

6.7.2 Cluster Ops

- **Cluster Creation**: The processed and replaced documents are clustered to group similar documents together.
- **Shuffling Operation**: Clusters are shuffled to further obscure the document structure and content, enhancing the deception.

6.7.3 Deceptive Repository

- **Shuffled Cluster**: The shuffled clusters are stored in a deceptive repository, ensuring that the data is misleading for any unauthorized analysis.

6.7.4 Secure Enclave

- **Keyword Mappings**: Maps original keywords to their replacements, necessary for reversing the deception if needed.
- **AES Encryption**: Ensures the data and mappings are securely encrypted.
- **User Authentication**: Controls access to the secure enclave, ensuring that only authorized users can access the mappings and encrypted data.

Chapter 7

Conclusions & Future Recommendation

7.1 Conclusion

The comparison between the original and deceptive results across all three datasets clearly demonstrates the superiority of the DARD techniques in protecting sensitive information. The t-SNE visualizations show a significant disruption in the clustering patterns, with the deceptive datasets exhibiting more scattered and less defined clusters. This visual evidence is complemented by the quantitative analysis of clustering scores, where the deceptive datasets consistently show lower Silhouette Coefficient and Calinski-Harabasz Index values, and higher Davies-Bouldin Index values.

These findings prove that the DARD techniques effectively obfuscate the data, making it more challenging for adversaries to perform accurate clustering and topic identification. By disrupting the natural structure of the data, the deceptive strategies provide a robust defense mechanism that enhances the security and confidentiality of the information. Therefore, the use of DARD techniques is validated as an effective approach to safeguarding sensitive documents against automated adversarial attacks.

7.2 Future Recommendations

While the current research demonstrates the effectiveness of the Deceptive Approaches for Robust Defense (DARD) techniques in protecting sensitive information, several avenues for future work can further enhance and extend these findings. Here are some recommendations for future research in this area:

7.2.1 Extending to Diverse Document Types

The current study focuses on research papers, summaries, and company documents. Future research should consider extending the application of DARD techniques to other types of documents, such as emails, social media posts, and multimedia files. This will help in understanding the versatility and adaptability of the DARD techniques across various formats and content structures.

7.2.2 Real-time Deception Techniques

Implementing DARD techniques in real-time environments can significantly enhance their practical utility. Future work should explore the development of algorithms and systems that can apply keyword replacement and cluster shuffling dynamically as data is created or transmitted. This would provide continuous protection against adversarial attacks and ensure up-to-date security measures.

7.2.3 Integration with Other Security Measures

Combining DARD techniques with other data security measures, such as encryption, access control, and anomaly detection, can provide a multi-layered defense strategy. Future research should investigate how these combined approaches can offer more robust protection and address any potential gaps left by individual techniques.

7.2.4 Advanced Replacement and Shuffling Strategies

While the current study explores basic replacement and shuffling techniques, future work should investigate more advanced methods. This includes context-aware replacements, semantic-based shuffling, and the use of machine learning algorithms to optimize the replacement and shuffling processes. These advanced techniques could further enhance the deception and make it even more challenging for adversaries to reverse-engineer the modifications.

7.2.5 Evaluating Impact on User Experience

It is essential to ensure that the application of DARD techniques does not negatively impact the usability and readability of the documents for legitimate users. Future research should include user studies to evaluate how these techniques affect the end-user experience and identify ways to balance security with usability.

7.2.6 Longitudinal Studies

Conducting longitudinal studies to assess the long-term effectiveness of DARD techniques is crucial. This includes monitoring how adversaries evolve their strategies to counter these techniques and adapting the DARD methods accordingly. Long-term studies can provide insights into the durability and resilience of the deception strategies over time.

7.2.7 Cross-domain Application

Exploring the application of DARD techniques across different domains, such as healthcare, finance, and government, can provide valuable insights into their effectiveness in various contexts. Each domain may have unique challenges and requirements, and understanding these can help in tailoring the DARD techniques to specific use cases.

7.2.8 Legal and Ethical Considerations

As deception techniques become more sophisticated, it is essential to address the legal and ethical implications of their use. Future research should examine the regulatory frameworks and ethical considerations surrounding the deployment of DARD techniques, ensuring that they are used responsibly and within legal boundaries.

7.2.9 Development of Evaluation Frameworks

Creating comprehensive evaluation frameworks to assess the effectiveness of DARD techniques can standardize the measurement of their impact. These frameworks should include a variety of metrics and methodologies to provide a holistic view of the security, usability, and practicality of the techniques.

7.3 Remarks

By addressing these future directions, researchers can continue to advance the field of data security and enhance the effectiveness of DARD techniques in protecting sensitive information from sophisticated adversarial attacks.

References

[1] A. M. MONGARDINI, M. La Morgia, S. Jajodia, L. V. Mancini and A. Mei, "DARD: Deceptive Approaches for Robust Defense against IP Theft," *Transactions on Information Forensics & Security,* 2024.

[2] W. . Roberds and S. L. Schreft, "Data Breaches and Identity Theft," *Journal of Monetary Economics,* vol. 56, no. 7, pp. 918-929, 2009.

[3] J. G. Dutrisac and D. B. & Skillicorn, "Hiding clusters in adversarial settings," *IEEE International Conference on Intelligence and Security Informatics,* pp. 185-187, 2008.

[4] B. Biggio and e. al., "Is data clustering in adversarial settings secure?," in *ACM workshop on Artificial intelligence and security*, 2013.

[5] P. Karuna and e. al., "Fake Document Generation for Cyber Deception by Manipulating Text Comprehensibility," *IEEE Systems Journal,* vol. 15, no. 1, pp. 835-845, 2021.

[6] A. Abdibayev and e. al., "Using Word Embeddings to Deter Intellectual Property Theft through Automated Generation of Fake Documents," *ACM Transactions on Management Information Systems ,* vol. 12, no. 2, pp. 1-22, 2021.

[7] D. M. Blei, "Introduction to Probabilistic Topic Models," *Comm. ACM,* vol. 55, no. 4, p. 77–84, .

[8] D. M. Blei, A. Y. Ng and M. I. Jordan, "Latent Dirichlet Allocation," *Journal of Machine Learning Research,* vol. 3, no. 4–5, p. "pp." 993–1022, .

[9] T. . Kanungo, D. M. Mount, N. S. Netanyahu, C. D. Piatko, R. . Silverman and A. Y. Wu, "An efficient "k"-means clustering algorithm: Analysis and implementation," *IEEE Transactions on Pattern Analysis and Machine Intelligence,* vol. 24, no. 7, p. 881–892, 2002.

[10] H. B. Zhou and J. T. Gao, "Automatic Method for Determining Cluster Number Based on Silhouette Coefficient," *Advanced Materials Research,* vol. 951, no. , pp. 227-230, 2014.

[11] B. . Halpin, "CALINSKI: Stata module to compute Calinski-Harabasz cluster stopping index from distance matrix," *Statistical Software Components,* vol. , no. , p. , 2015.

[12] J. . Xiao, J. . Lu and X. . Li, "Davies Bouldin Index based hierarchical initialization K-means," *Intelligent Data Analysis,* vol. 21, no. 6, pp. 1327-1338, 2017.

[13] A. Chhabra, A. Sekhari and a. P. Mohapatra, "On the Robustness of Deep Clustering Models: Adversarial Attacks and Defenses," *Advances in Neural Information Processing Systems,* vol. 35, pp. 1-12, 2022.

[14] E. Nowroozi and e. al., "Adversarial Attacks and Defenses in Machine Learning-Powered Networks: A Contemporary Survey," *IEEE Conference on Machine Learning and Applications,* pp. 50-61, 2023.

[15] A. Berahmand and e. al., "A community detection model using node embedding and adversarial learning," *ScienceDirect Journal of Network and Computer Applications,* vol. 125, pp. 123-134, 2023.

[16] Y. Liu and e. al., "Understanding of internal clustering validation measures," *IEEE International Conference on Data Mining,* pp. 911-916, 2010.

[17] V. Ferrulli, "PyPaperBot," 1 11 2021. [Online]. Available: https://www.piwheels.org/project/pypaperbot/. [Accessed 24 07 2024].

[18] "arXiv License Information," , . [Online]. Available: https://arxiv.org/help/license. [Accessed 24 7 2024].

[19] A. Cherguelaine, "Company Documents Dataset," 23 May 2024. [Online]. Available: https://www.kaggle.com/datasets/ayoubcherguelaine/company-documents-dataset. [Accessed 24 July 2024].

[20] "PyPI - the Python Package Index," , . [Online]. Available: http://pypi.python.org/pypi. [Accessed 24 7 2024].

[21] M. . Hassler and G. . Fliedl, "Text Preparation Through Extended Tokenization," *WIT Transactions on Information and Communication Technologies,* vol. 37, no. , pp. 13-21, 2006.

[22] "Natural Language Toolkit — NLTK 3.0 documentation," , . [Online]. Available: http://www.nltk.org. [Accessed 24 7 2024].

[23] Q. . Kuang and X. . Xu, "Improvement and Application of TF•IDF Method Based on Text Classification," , 2010. [Online]. Available: http://yadda.icm.edu.pl/yadda/element/bwmeta1.element.ieee-000005566113. [Accessed 24 7 2024].

[24] Y. . Seki, "Sentence Extraction by tf/idf and Position Weighting from Newspaper Articles," , . [Online]. Available: http://research.nii.ac.jp/ntcir/workshop/OnlineProceedings3/NTCIR3-TSC-SekiY.pdf. [Accessed 24 7 2024].

[25] C. . Manning, P. . Raghavan and H. . Schutze, Introduction to Information Retrieval, ed., vol. , , : , 2008, p. 100.

[26] S. Qaiser and R. Ali, "Text Mining: Use of TF-IDF to Examine the Relevance of Words to Documents," *International Journal of Computer Applications,* vol. 181, no. 1, p. (0975 – 8887, 2018.

[27] X. . Lei, G. . Yifei, L. . Sheng, H. . Wei and X. . Di, "Research on the Improved FCM Cluster Method in the Hotspots Analysis on Web," , 2012. [Online]. Available: https://dl.acm.org/citation.cfm?id=2473506.2475045. [Accessed 24 7 2024].

[28] W. McKinney, "DataFrame," in *Python for data analysis*, O'Reilly, 2022, p. 111.

[29] G. . Hinton and L. . van der Maaten, "Visualizing Data Using t-SNE," *Journal of Machine Learning Research,* vol. 9, no. , p. 2579–2605, .

[30] L. v. d. Maaten and G. E. Hinton, "Visualizing High-Dimensional Data Using t-SNE," *Journal of Machine Learning Research,* vol. 9, no. 26, p. 5, 2008.

[31] D. . Arthur and S. . Vassilvitskii, "k-means++: the advantages of careful seeding," , 2007. [Online]. Available: http://ilpubs.stanford.edu:8090/778/1/2006-13.pdf. [Accessed 24 7 2024].