DISC: A Dataset for Information Security Classification

Elijah Bass[a], Massimiliano Albanese[b] and Marcos Zampieri[c]
Center for Secure Information Systems, George Mason University, Fairfax, U.S.A.

Keywords: Information Security, Information Protection, Security Classification, Artificial Intelligence, Datasets.

Abstract: Research in information security classification has traditionally relied on carefully curated datasets. However, the sensitive nature of the classified information contained in such documents poses challenges in terms of accessibility and reproducibility. Existing data sources often lack openly available resources for automated data collection and quality review processes, making it difficult to facilitate reproducible research. Additionally, datasets constructed from declassified information, though valuable, are not readily available to the public, and their creation methods remain poorly documented, rendering them non-reproducible. This paper addresses these challenges by introducing DISC, a dataset and framework, driven by artificial intelligence principles, for information security classification. This process aims to streamline all the stages of dataset creation, from preprocessing of raw documents to annotation. By enabling reproducibility and augmentation, this approach enhances the utility of available document collections for information security classification research and allows researchers to create new datasets in a principled way.

1 INTRODUCTION

In the realm of information security, *classification* refers to the process of categorizing and protecting information based on its sensitivity, confidentiality, or criticality (NIST, 2004). Information classification is an essential aspect of information security and data governance, enabling organizations to adeptly manage information while ensuring the protection of sensitive data and compliance with regulatory standards. Classification systems are designed to assist organizations in developing robust classification strategies that align with their security requirements and facilitate efficient data management. Government agencies, private sector entities, and industry-specific users (e.g., healthcare, payment card industry) leverage information classification systems as integral components of their information protection planning. Government agencies rely on information classification systems to control access to national security information. Industry-specific users adhere to classification protocols dictated by regulatory and standards requirements, such as the Health Insurance Portability and Accountability Act (HIPAA) for safeguarding medical industry information in the United States and the Payment Card Industry Data Security Standard (PCI DSS) for protecting data used in credit card transactions.

Information classification stands as a critical aspect of government and military operations, with each country maintaining a unique classification system tailored to its specific needs. In the United States, Executive Order (EO) 13526[1] mandates a uniform system for classifying, safeguarding, and declassifying national security information within the government and military (White House, 2009). Government officials with the role of Original Classification Authorities (OCAs) determine the classification level based on the assessment of potential damage to national security due to unauthorized disclosure. Security Classification Guides detail and manage the sensitivity of information (Information Security Oversight Office, 2018). Section 1.2 of EO 13526 prescribes three distinct levels of information classification labels, namely CONFIDENTIAL, SECRET, and TOP SECRET (White House, 2009). The CONFIDENTIAL classification is assigned to information where unauthorized disclosure is expected to cause damage to national security, while the SECRET classification is reserved for information whose unauthorized disclosure

[a] https://orcid.org/0009-0001-8680-1815
[b] https://orcid.org/0000-0002-2675-5810
[c] https://orcid.org/0000-0002-2346-3847

[1] Once issued, executive orders remain in force until they are canceled or revoked by the current president, expire, or are adjudicated as unlawful.

Bass, E., Albanese, M. and Zampieri, M.
DISC: A Dataset for Information Security Classification.
DOI: 10.5220/0012763400003767

In *Proceedings of the 21st International Conference on Security and Cryptography (SECRYPT 2024)*, pages 175-185
ISBN: 978-989-758-709-2; ISSN: 2184-7711

is expected to cause serious harm to national security. The TOP SECRET classification is applied when the unauthorized disclosure is anticipated to cause exceptionally grave damage to national security. Information that doesn't meet the sensitivity requirements outlined in EO 13526 Section 1.2 is designated as UNCLASSIFIED.

The advancement of information security classification research has historically been hindered by limited shareability of curated datasets and resources required to automate dataset curation and assess quality. Previous information security efforts have focused on researching classification algorithms. Usable datasets and frameworks to replicate research are not yet available. The sensitive nature of information used in security classification precludes the sharing of most datasets. The need for manual review of data extracted from the few available document repositories has been demonstrated as essential for ensuring data quality and accuracy (Engelstad et al., 2015b; Brown and Charlebois, 2010; Engelstad et al., 2015a). This is particularly evident in correcting errors resulting from optical character recognition (OCR) applied to legacy documents, and in preserving the integrity of document information.

To tackle these challenges, we propose a framework for automating the creation of datasets to serve as training data to support the development of advanced information security classification systems. We illustrate the use of this framework by creating DISC, the first reproducible information security dataset, which we make publicly available. The proposed approach ensures the integrity and quality of the data, thus reducing the need for manual review, by harnessing the power of AI-based methods and Large Language Models (LLMs). We report the results of initial experiments on information classification using the DISC dataset and discuss future research directions.

The remainder of this paper is organized as follows: Section 2 describes datasets curated in prior research. Section 3 introduces a formal model for representing documents and supporting granular classification. Section 4 outlines the methodology for collecting, processing, and storing information from a declassified document database to create a shareable and reusable dataset. Section 5 describes areas of research that can benefit from the proposed methods and DISC whereas Section 6 demonstrates the utility of DISC on information security classification. Finally, Section 7 provides concluding remarks and identifies future research directions.

2 RELATED WORK

The sensitivity of CLASSIFIED information and its potential security impact from public release precludes the availability of readily accessible public datasets. As a result, information classification datasets have been curated using publicly released, declassified, or leaked information sources. In response to the evolving need for diverse datasets, an emerging research field focuses on generating document datasets, leveraging advancements in deep learning to augment existing capabilities and resources.

In the United States, information must undergo reclassification or the declassification process upon the expiration of its original classification period (White House, 2009). Publicly released government documents that have completed the declassification process serve as raw data for creating a realistic classification dataset. The Digital National Security Archive (DNSA) stands out as a comprehensive collection, comprising over 100,000 declassified records that document historic United States policy decisions[2]. This archive contains electronic copies of both previously UNCLASSIFIED and UNCLASSIFIED government documents related to U.S. foreign policy from the post-World War II era to the present day. Notably, research highlighted in (Engelstad et al., 2015b; Brown and Charlebois, 2010; Engelstad et al., 2015a) utilized specific DNSA domains, encompassing a mix of CLASSIFIED and UNCLASSIFIED documents, including:

- Afghanistan: The Making of U.S. Policy, 1973-1990.
- China and the United States: From Hostility to Engagement, 1960-1998.
- The Philippines: U.S. Policy during the Marcos Years, 1965-1986.

Information made publicly available through data breaches and data leaks can aid in the creation of difficult-to-obtain research datasets. However, researchers must conscientiously identify and address the ethical concerns associated with the process through which the information was released into the public domain before utilizing leaked data (Boustead and Herr, 2020).

WikiLeaks, founded in 2006, is an organization known for collecting and publishing sensitive leaked documents. Notably, WikiLeaks published confidential United States diplomatic cables spanning from 2003 to 2010. These diplomatic cables, originating from U.S. Embassies and Consulates worldwide,

[2]https://nsarchive.gwu.edu/publications-collections.

were presented on the WikiLeaks website in a static HTML format, organized by embassies, and sorted by their respective origination dates. In the research conducted in (Alzhrani et al., 2016), the leaked diplomatic cables were processed, leading to the creation of datasets for the Baghdad, London, Berlin, and Damascus embassies. The content of these datasets was classified at the paragraph level, with a total of 21,718 paragraphs labeled as UNCLASSIFIED and 9,291 paragraphs labeled as SECRET.

The challenge of insufficient datasets for training and validating sensitive information classification models is a common hurdle and may be addressed using data augmentation. The research in (Jadli et al., 2020) introduced the use of deep convolutional adversarial networks (DCGAN) to generate synthetic document images from an existing scanned document dataset. These generated images were combined with original images in a dataset to train an image classifier using a convolutional neural network for document classification. The accuracy of the resulting model, trained on the augmented dataset, was then compared to a model trained on a similar-sized dataset consisting solely of original documents. Remarkably, the constructed model performed comparably to the model trained with authentic data, achieving accuracy of 90% and 91%, respectively.

The research in (Chakraborty et al., 2021) focuses on generating additional documents by replacing concepts with semantically related concepts selected from an ontology. The results yield, for every real document d, a set $Fake(d)$ of fake documents that closely resemble d. Another approach demonstrated in (Whitham, 2017) employs a rule-based and preset template-based method for textual information generation. This technique involves parsing text to identify part-of-speech tags, selecting an original document as a template, and utilizing word transposition and substitution based on part-of-speech tagging and n-grams.

In varying degrees, previous research efforts have described the processes utilized to create datasets for training and testing information security classifier algorithms. However, efforts in contacting prior researchers to obtain datasets to recreate prior research results has been unsuccessful, confirming the challenges in accessing quality data for information classification research. To the author's knowledge, DISC is the first openly available and reusable dataset for information security classification[3].

[3]The dataset and associated resources will be made publicly available upon publication of this paper.

3 CORPUS FORMAL MODEL

The availability of high-quality and well-curated datasets, often referred to as corpora, is essential for advancing the state-of-the-art in information security classification. A suitable corpus serves as the foundation upon which researchers can build and evaluate models, algorithms, and methodologies to address key challenges and issues within the domain of information security classification. This section describes a formal model for representing documents and supporting fine-grained classification of document elements (paragraphs and sentences), whereas Section 4 provides a comprehensive description of the corpus utilized in our research, encompassing both the data collection methodology and relevant information regarding the composition, characteristics, and structure of the corpus.

In the field of information security, the information security classification of documents based on their sensitivity and confidentiality requirements plays a pivotal role in safeguarding valuable data from unauthorized access and disclosure. However, despite the critical importance of this task, existing information security classification literature often lacks a formalized information model capable of describing documents at different levels of granularity, such as paragraphs or sentences. This gap presents a significant challenge in efficiently and accurately predicting the security classification of different parts of a document and the classification of the information that could be derived by combining different parts of the same or different documents. This modeling capability is therefore critical for reasoning about inference attacks and preventing the under-classification of documents within a corpus.

3.1 Modeling Documents in a Corpus

A corpus of documents in an information system can be represented as a set:

$$D = \{d_1, d_2, ..., d_i, ...d_{m-1}, d_m\} \quad (1)$$

where $m = |D|$ is the number of documents in the corpus. A document d_i can then be represented as an ordered sequence of paragraphs:

$$d_i = \langle p_{i1}, p_{i2}, ..., p_{ij}, ...p_{im_i} \rangle \quad (2)$$

where m_i is the number of paragraphs in document d_i and p_{ij} is the j-th paragraph in document d_i.

The ability to accurately model and analyze paragraphs within a corpus holds significant importance for various information security applications ranging from information security categorization to threat detection. Paragraphs serve as fundamental units of

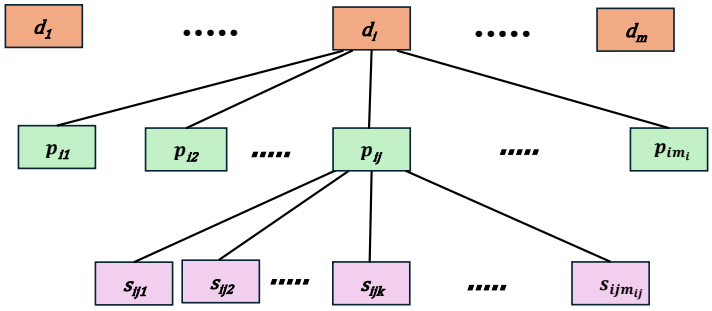


Figure 1: Hierarchical corpus model.

information within documents and include coherent and semantically meaningful segments of text. Understanding the structural and contextual attributes of paragraphs within a corpus is essential for extracting valuable insights, identifying patterns, and making informed decisions regarding information security policies and practices. This section focuses on the task of modeling paragraphs within a corpus, with a specific emphasis on its relevance to information security classification.

Finally, a paragraph p_{ij} can be represented as an ordered sequence of sentences:

$$p_{ij} = \langle s_{ij1}, s_{ij2}, ..., s_{ijk}, ..., s_{ijm_{ij}} \rangle \tag{3}$$

where m_{ij} is the number of sentences in paragraph p_{ij} and s_{ijk} is the k-th sentences in paragraph p_{ij}.

Considering that the number of documents in the corpus is m and m_i denotes the number of paragraphs in document d_i, the total number of paragraphs in the corpus is given by:

$$\sum_{i=1}^{m} m_i \tag{4}$$

Similarly, considering that the number of sentences in a paragraph p_{ij} is m_{ij}, the total number of sentences in the corpus is given by:

$$\sum_{i=1}^{m} \sum_{j=1}^{m_i} m_{ij} \tag{5}$$

This hierarchical model of a document corpus is illustrated in Figure 1.

3.2 Achieving Granular Classification

In the domain of information security classification, achieving granular classification at the level of individual sentences and paragraphs is crucial for effectively safeguarding sensitive data and ensuring compliance with security policies, while preserving access to information for those with sufficient clearance.

Traditional approaches to document classification often focus on categorizing entire documents or sections based on their overall sensitivity. There exists a growing recognition of the need for more granular classification techniques that can discern the varying levels of sensitivity present within a document's elements. This section explores the importance of achieving granular sentence and paragraph classification within the context of information security.

To model granular classification at the sentence and paragraph level, let $C = \{\text{UNCLASSIFIED, CONFIDENTIAL, SECRET, TOP SECRET}\}$ denote the set of security classifications and let P and S respectively denote the sets of all paragraphs and sentences in the corpus, which can be defined as follows.

$$P = \bigcup_{i=1}^{m} \left(\bigcup_{j=1}^{m_i} \{p_{ij}\} \right) \tag{6}$$

$$S = \bigcup_{i=1}^{m} \left(\bigcup_{j=1}^{m_i} \left(\bigcup_{k=1}^{m_{ij}} \{s_{ijk}\} \right) \right) \tag{7}$$

We can represent the sentence-level classification as a mapping $\delta_s : S \to C$ that associates a classification label $\delta_s(s)$ to each sentence $s \in S$. Similarly, we can represent the paragraph-level classification as a mapping $\delta_p : P \to C$ and the document-level classification as a mapping $\delta_d : D \to C$.

Given a paragraph $p_{ij} \in P$, one would intuitively assume that its classification is the highest classification among its sentences:

$$\delta_p(p_{ij}) = \max_{k \in (1, m_{ij})} \delta_s(s_{ijk}) \tag{8}$$

Similarly, given a document $d_i \in D$, one would intuitively assume that its classification is the highest classification among its paragraphs:

$$\delta_d(d_i) = \max_{j \in (1, m_i)} \delta_p(p_{ij}) \tag{9}$$

However, combining multiple pieces of information at one classification level may result in information that should be classified at a higher level. Thus, information classification must consider whether the aggregate of multiple units of text (sentences or paragraphs) should be assigned a higher classification

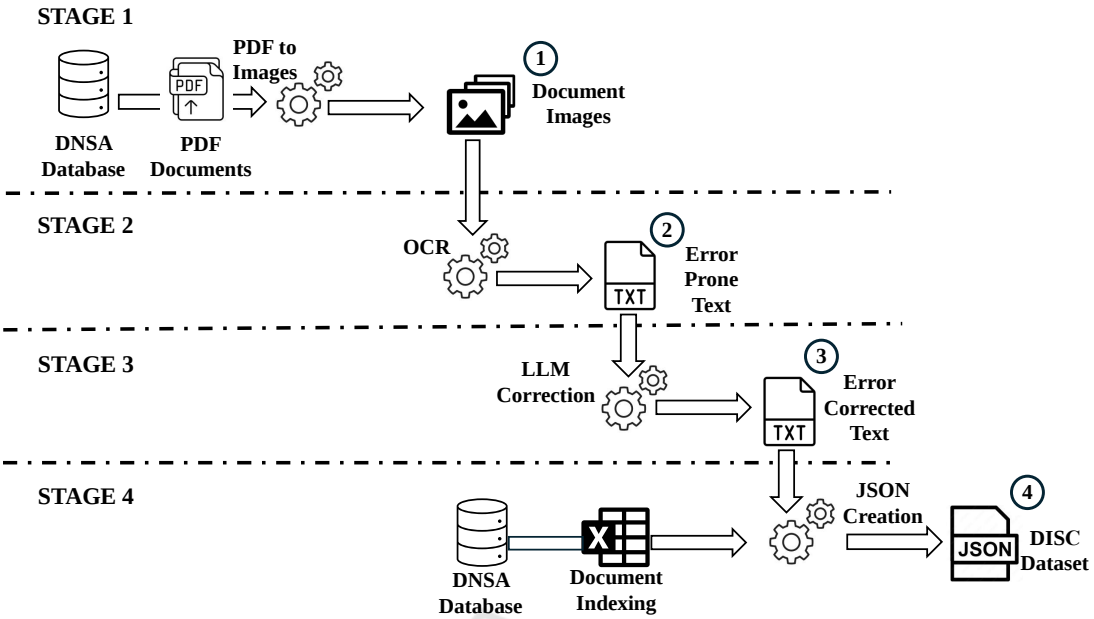


Figure 2: Dataset for Information Security Classification (DISC) processing chain.

level than the highest classification level of all the individual units. In other words Eqs. 8 and 9 must be rewritten to define lower bounds on the classification levels.

$$\delta_p(p_{ij}) \geq \max_{k \in (1, m_{ij})} \delta_s(s_{ijk}) \tag{10}$$

$$\delta_d(d_i) \geq \max_{j \in (1, m_i)} \delta_p(p_{ij}) \tag{11}$$

The formal information model introduced in this section forms a foundational framework for organizing and categorizing document content in a systematic manner. At its core, the formal information model delineates the hierarchical structure of documents, breaking them down into constituent paragraphs and sentences. Each level of granularity within the model is endowed with metadata attributes that capture relevant information about the content and potential security implications of the text. By leveraging this formalized representation of document content, the proposed method aims to streamline the process of security classification by automating key tasks such as document collection, OCR for text extraction, and subsequent analysis. By facilitating the utilization of advanced Natural Language Processing (NLP) techniques and machine learning algorithms, the proposed formal information model enables intelligent classification decisions based on the semantic understanding of document contents at the paragraph and sentence levels.

4 CORPUS DESCRIPTION

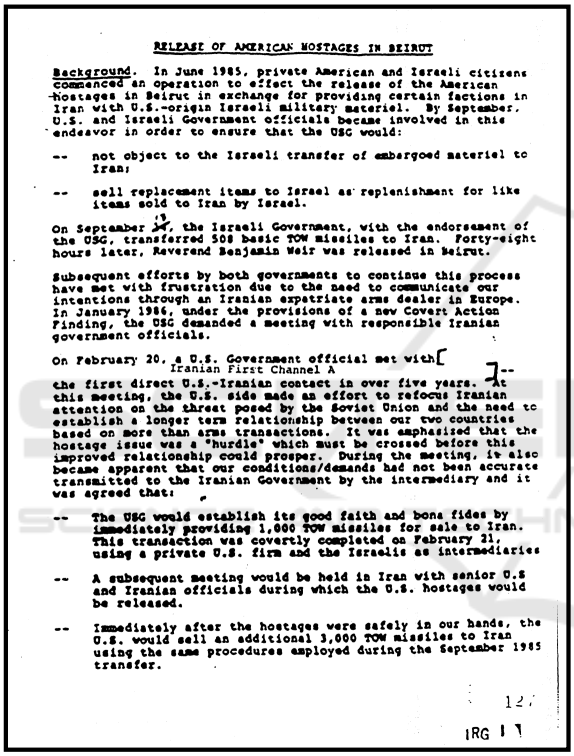
In this section, we first describe the methodology we employed to generate the DISC corpus from the original DNSA database, and then describe in detail the structure of the corpus itself.

4.1 Data Collection Methodology

The DISC corpus is built upon the documents contained within the DNSA database, and specifically the documents in the three topic areas outlined in Section 2. Out of the 2,459 documents available within those topic areas, documents deemed not relevant to information classifications were excluded. Specifically, the following documents were omitted: duplicate documents, documents classified as UNKNOWN, and documents marked as EXCISED (i.e., the classified text sections were redacted).

The DISC processing pipeline is illustrated in Figure 2. Stage 1 of the processing pipeline involves performing image processing on the pages of the PDF documents. The documents in the DNSA database consist of non-searchable PDF files containing scanned images of the original printed documents. The PDF files were searched and organized according to their initial classification (UNCLASSIFIED, CONFIDENTIAL, SECRET, or TOP SECRET) and the topic areas mentioned earlier. The initial classification of each document was determined by the author and classification originator, based on the procedures

in (White House, 2009). The documents within the DNSA database have undergone the declassification process, therefore the value of their reclassification metadata is set to UNCLASSIFIED. The document reclassification (UNCLASSIFIED) and DNSA database document search parameters (initial classification and domain) forms the metadata associated with the resultant documents during Stage 1 of the DISC processing pipeline. The Python `fitz` package was used to ingest individual PDF files and convert their pages to images. Figure 3 shows an example of a page from a PDF document in the DSNA database converted to an image in Stage 1.



RELEASE OF AMERICAN HOSTAGES IN BEIRUT

Background. In June 1985, private American and Israeli citizens commenced an operation to effect the release of the American hostages in Beirut in exchange for providing certain factions in Iran with U.S.-origin Israeli military materiel. By September, U.S. and Israeli Government officials became involved in this endeavor in order to ensure that the USG would:

-- not object to the Israeli transfer of embargoed materiel to Iran;

-- sell replacement items to Israel as replenishment for like items sold to Iran by Israel.

On September 14, the Israeli Government, with the endorsement of the USG, transferred 508 basic TOW missiles to Iran. Forty-eight hours later, Reverend Benjamin Weir was released in Beirut.

Subsequent efforts by both governments to continue this process have met with frustration due to the need to communicate our intentions through an Iranian expatriate arms dealer in Europe. In January 1986, under the provisions of a new Covert Action Finding, the USG demanded a meeting with responsible Iranian government officials.

On February 20, a U.S. Government official met with [Iranian First Channel A]. At the first direct U.S.-Iranian contact in over five years. At this meeting, the U.S. side made an effort to refocus Iranian attention on the threat posed by the Soviet Union and the need to establish a longer term relationship between our two countries based on more than arms transactions. It was emphasized that the hostage issue was a "hurdle" which must be crossed before this improved relationship could prosper. During the meeting, it also became apparent that our conditions/demands had not been accurate transmitted to the Iranian Government by the intermediary and it was agreed that:

-- The USG would establish its good faith and bona fides by immediately providing 1,000 TOW missiles for sale to Iran. This transaction was covertly completed on February 21, using a private U.S. firm and the Israelis as intermediaries.

-- A subsequent meeting would be held in Iran with senior U.S and Iranian officials during which the U.S. hostages would be released.

-- Immediately after the hostages were safely in our hands, the U.S. would sell an additional 3,000 TOW missiles to Iran using the same procedures employed during the September 1985 transfer.

IRG

Figure 3: Stage 1: Example of non-searchable image-based PDF document.

Stage 2 of the DISC processing pipeline uses an Optical Character Recognition (OCR) algorithm to convert the images of document pages into editable and searchable data. The OCR processing stage enables subsequent information manipulation, editing, searching, and analysis using information security classification algorithms. The Python `pytesseract` module was used to perform OCR processing on the images formed from the pages of individual documents. Figure 4 shows the output of OCR processing for the document image of Figure 3. In the interest of space, only the first few paragraphs of the document are shown in Figure 4.

RELEASE OF AMERICAN **WOSTAGES** In BEIRUT

Background. In June 1985, **privete** American and **Yereeli citeizens** Commenced **8m** Operation to **ecfect** the release of the **Americen +iostages** in Beirut in exchange for providing certain factions in Iran with **0.$**.-origan **Isreeli militery sateriel**. By September, U.S. and Israeli Government officials became involved in **thie** -endeavor in order to ensure that the **086** would:

– not object to the **Isrseli transfer of embergoed naterie! tc** Iran;

=< eell replacement items to **lerae)]** as replenishment for like items sold to **fran** by Iserael.

ih On September **3f**, the Israeli Government, with **che** endorsement of the **OSG**, transferred 508 **besic** TOW missiles to Iran. Forty-eight hours later, Reverend Benjamin Weir was **relesseé** in beirut.

Figure 4: Stage 2: OCR error-prone textual output for the document image in Figure 3. OCR errors in the Background section of the document errors are indicated in red text.

RELEASE Of AMERICAN **WOSTAGES** In BEIRUT

Background. In June 1985, **private** American and **Yereeli citizens** Commenced **8m** Operation to **effect** the release of the **American + hostages** in Beirut in exchange for providing actions in Iran with **0.$**.-origin **Israeli military materiel**. By September, U.S. and Israeli Government officials became involved in **the** -endeavor in order to ensure that the **086** would:

– not object to the **Israeli** transfer of **emerged nature! tc** Iran;

=< well replacement items to **were)]** as replenishment for like items sold to **fran** by **israel**.

ih) On September **3f**, the Israeli Government, with **che** endorsement of the **OS**, transferred 508 basic **OW** missiles to Iran. Forty-eight hours later, Reverend Benjamin Weir was **releaseé** in **bert**.

Figure 5: Results of the Python Speller package's `autocorrection` module on the OCR output in Figure 4. Uncorrected errors in the Background section of the document are indicated in red text. `Autocorrection` module corrections on the OCR output are in blue text.

Stage 3 of the DISC processing pipeline performs error correction on the OCR output. The Python Speller package's `autocorrection` module was investigated for accuracy in correcting the OCR processing errors. The primary motivation for investigating the Python Speller package is its ability to perform local processing, thus maintaining confidentiality of information, and offer cost-effectiveness and speed. The result of the `autocorrection` module on the information in Figure 4 is illustrated in Figure 5. This module was able to correct only words with errors involving one or two characters.

The performance of the LLM-based error correction methodology was compared against the Python Speller package. LLMs detect OCR errors by analyz-

Table 1: Stage 4: DNSA documents indexing information.

ID	Title	Abstract	Pub. Date	Authors
1679059219	Mujajedin Cross Border Cow Raid...	Soviet Union Armed Forces retaliate...	June 14, 1987	U.S. Consulate of Peshawar
1679076686	Afghanistan: Gailani Welcome Zahir Shah's...	Itihad-i Islami Baraye Azad-i Afghanistan...	May 18, 1987	U.S.Embassy of Pakistan
1679059734	Jamiat-i-Islami Comment on Mujahideen ...	Newspapers report that the Soviet Union ...	May 2, 1987	U.S. Consulate of Peshawar
1679077357	Soviet Reprisals in Northern Afghanistan	United States Embassy. Afghanistan officers ...	April 26, 1987	U.S. Embassy Afghanistan of Afghanistan
1679060091	Xinhua: Soviet, Afghan Planes Change...	United States Embassy. Afghanistan officers ...	April 21, 1987	U.S. Foreign Broadcast Information Service

ing inconsistencies or grammatical anomalies within the uncorrected text. Therefore, words that do not fit within the context of the document or are unlikely to occur based on language patterns can be identified as OCR errors and replaced with the mostly likely words. Previous related works outlined the need for manual visual reviews to ensure quality of the information. The LLM-based technique implemented in the DISC processing pipeline allows automation of the OCR error correction process and greatly reduces the need for manual information review. The LLM prompt utilized was meticulously crafted to correct OCR errors while preventing the generation of additional or alternative content. The result of the LLM-based OCR error correction on the information in Figure 4 is illustrated in Figure 6. The LLM-based OCR error correction greatly outperformed Python's `autocorrection` module, demonstrating the ability to understand the context and remediate poor OCR results. The accuracy of LLM-based error correction, compared to the Python Speller package, outweighs the additional overhead in terms of time and cost. Therefore, the LLM methodology was selected to perform Stage 3 of the DISC processing pipeline, as illustrated in Figure 2.

RELEASE OF AMERICAN **HOSTAGES** IN BEIRUT

Background. In June 1985, private American and **Israeli citizens** commenced **an** operation to **effect** the release of the American **hostages** in Beirut in exchange for providing certain factions in Iran with **U.S.**-origin **Israeli military material**. By September, U.S. and Israeli Government officials became involved in **this** endeavor in order to ensure that the **U.S.** would:

– not object to the **Israeli** transfer of **embargoed material to** Iran;

– **sell** replacement items to **Iran** as replenishment for like items sold to Iran by Israel.

On September **3**, **the** Israeli Government, with the endorsement of the **U.S.G,** transferred 508 basic TOW missiles to Iran. Forty-eight hours later, Reverend Benjamin Weir was **released** in **Beirut**.

Figure 6: Stage 3: LLM-based error free correction of the OCR output in Figure 4. Corrected errors in the Background section of the document are indicated in blue text.

Finally, Stage 4 of the DISC processing pipeline collects indexing information of DNSA documents for correlation to the LLM-enhanced OCR error-corrected textual information obtained from Stage 3 to create the DISC JSON entries. In addition to the image-based PDF copies of the original documents, the DNSA database user interface facilitates generating and exporting an Excel spreadsheet containing document metadata. This metadata includes indexing information and an abstract summarizing the document. The indexing information collected from DNSA consists of document title, classification date, database name, authors information, and assigned unique document identifier, as illustrated in Figure 1.

The final step in the DISC processing pipeline consists in correlating document abstract and indexing information with the textual information produced in Stage 3 for storage in the DISC's JSON format, as shown in Figure 7.

```
{
  "DocID": 7,
  "Title": "Release of American Hostages in
       Beirut ... ",
  "ORC-Text": "RELEASE OF AMERICAN WOSTAGES In
       BEIRUT ... ",
  "Text": "RELEASE OF AMERICAN HOSTAGES IN
       BEIRUT ... ",
  "Abstract": "Oliver L. North calls Soviet
       policy ... ",
  "Classification": [{
      "ClassID": "7_1",
      "Label": "Top Secret",
      "Date": "c. April 4, 1986"
    },
    {
      "ClassID": "7_2",
      "Label": "Unclassified"
    }
  ],
  "Database": "Digital National Security
       Archive",
  "Domain": "Afghanistan: The Making of U.S.
       Policy, 1973\u20131990",
  "Author": "United States. National Security
       Council",
  "StoreID": "1679059818"
}
```

Figure 7: Document indexing and textual information stored in a JSON record.

4.2 Corpus Description

Section 4.1 provided a comprehensive overview of the methodology utilized in the creation of the DISC corpus from its precursor, the DNSA database. The processing chain and information processing strategy elaborated upon in section 2 provides a repeatable process to recreate or augment the developed reusable DISC dataset. This section dives deeper into the DISC corpus, focusing on its information organization and presenting detailed statistics pertaining to the archived documents available in DISC.

The information within the DNSA not only consists of PDF documents comprised of textual images but also includes indexed information created during content curation, Indexes information describes the document content and provenance information. The information obtainable for each document in the DISC corpus is illustrated in Table 2.

Table 2: Attributes for documents in the DISC corpus.

Attribute	Description
DocID	Unique Document ID
Title	Document title
Abstract	Brief summary of the document
OCR-Text	Text extracted via OCR from the PDF document
Text	Text reconstructed via LLM from the OCR output
Classification	Document classification events (each document may have multiple classification events)
Database	References to the database the document was extracted from
Domain	Domain within the database the document was extracted from
Author	Authorship information
StoreID	Database-assigned unique document ID

Changes in information sensitivity, facilitate modeling of the time-decay of information sensitivity, or satisfy decisions to increase level of information confidentiality.

Table 3: Attributes for classification events.

Attribute	Description
ClassID	Unique classification event ID
Label	Classification level (UNCLASSIFIED, SECRET, or TOP SECRET)
Date	Classification date

The DISC consists of the following DNSA domain areas:

- AF, Afghanistan: The Making of U.S. Policy, 1973-1990.
- CH, China and the United States: From Hostility to Engagement, 1960-1998.
- PH, The Philippines: U.S. Policy during the Marcos Years, 1965-1986.

The documents statistics and original documentation classifications contained in DISC are listed in Table 4.

Table 4: Information classification statistics for documents in the DISC corpus.

Domain	UNCLASSIFIED	SECRET	TOP SECRET	TOTAL
AF	401	206	13	620
CH	151	666	132	949
PH	411	469	1	881
TOTAL	963	1,341	146	2,450

5 APPLICATIONS

DISC was developed primarily to support NLP algorithm research for reactive information security classification of existing documents e.g., security classification after information creation. However, DISC can be utilized for other research purposes such as proactive information control and access applications.

Proactive information classification involves the preemptive process of categorizing and labeling data or information based on its content, sensitivity, or other characteristics before it is created or shared. This approach aims to enhance information management and security by implementing automated systems or policies that identify sensitive information during its creation, sharing, or storage. Through proactive information classification, high-value information can be identified in real-time, enabling organizations to implement immediate data protection measures. The DISC dataset facilitates research in NLP techniques to enforce security controls, authorization permissions, and data access protection measures at the point of information creation.

LLMs exhibit proficiency in processing and comprehending intricate language found in textual documents, making them valuable for enhancing information classification tasks. Leveraging their knowledge and contextual understanding, LLMs can extract crucial information, identify topics, and conduct sentiment analysis. An analysis of LLMs performance in classifying public affairs documents was conducted

in (Peña et al., 2023). Four distinct Spanish LLMs were employed to classify up to 30 different topics in the dataset. The findings underscored the effectiveness of LLMs in performing information classification on domain-specific documents. DISC would provide the necessary dataset to assist in the research and training of LLMs to automate the information security classification process.

6 EXPERIMENTAL EVALUATION

This section demonstrates how DISC can be used for information security classification (Brown and Charlebois, 2010; Engelstad et al., 2015a; Engelstad et al., 2015b). We use Term Frequency-Inverse Document Frequency (TF-IDF) with multiple classifiers such as Naive Bayes (NB), Support Vector Machines (SVM), K-nearest Neighbor (KNN), and Gradient Boost (GB). A Bidirectional Encoder Representations from Transformers (BERT) classifier (Devlin et al., 2019) was utilized to compare performance with the TF-IDF classifiers. We present the results of the different models in Table 5 along with performance results for each individual model in Tables 6 to 10.

Table 5: Performance comparison of classifier models.

Classifier	UNCLASSIFIED		CLASSIFIED		
	Precision	Recall	Precision	Recall	F1 Score
NB	1.00	0.94	0.94	1.00	0.97
SVM	0.97	0.99	0.99	0.97	0.98
KNN	0.97	0.91	0.91	0.91	0.94
GB	0.91	0.98	0.97	0.90	0.94
BERT	0.97	0.99	0.99	0.97	0.98

The CLASSIFIED documents misclassified by BERT were examined to provide insight into how the algorithm was learning language structure. The BERT algorithm classified the text "Reports on Taiwan's air defense capability, including availability of surface-to-air missiles and fighter aircraft." as being CLASSIFIED while the DNSA Database labeled the information as being UNCLASSIFIED. However, military capabilities and vulnerabilities of countries are considered classified information. In fact, an Internet search of the Taiwan air defense topic confirmed that information on military capabilities and vulnerabilities is in fact classified information and was recently leaked within an on-line chat forum (Nakashima et al., 2023). This result confirms that the manual process of classifying data is in inconsistently applied within organizations and it is prone to human error.

Table 6: Confusion matrix for the TF-IDF/Naïve Bayes classifier.

	UNCLASSIFIED	CLASSIFIED
UNCLASSIFIED	291	19
CLASSIFIED	0	315

Table 7: Confusion matrix for the TF-IDF/SVM classifier.

	UNCLASSIFIED	CLASSIFIED
UNCLASSIFIED	307	3
CLASSIFIED	8	307

Table 8: Confusion matrix for the TF-IDF/KNN classifier.

	UNCLASSIFIED	CLASSIFIED
UNCLASSIFIED	281	29
CLASSIFIED	8	307

Table 9: Confusion matrix for the TF-IDF/Gradient Boost classifier.

	UNCLASSIFIED	CLASSIFIED
UNCLASSIFIED	303	7
CLASSIFIED	31	284

Table 10: Confusion matrix for the BERT classifier.

	UNCLASSIFIED	CLASSIFIED
UNCLASSIFIED	306	4
CLASSIFIED	8	307

In conclusion, the experimentation conducted utilizing DISC for information security classification research has demonstrated several key benefits of our approach. By making our dataset openly available to the research community and providing a detailed methodology for its creation, we have enabled other researchers to access and utilize the identical dataset for their own studies. Moreover, the method utilized for creating DISC is replicable and facilitates augmentation and expansion of the dataset to accommodate diverse research needs and objectives. One of the most significant advantages of DISC is the utilization of a LLM, which significantly reduces the need for manual quality review of the data to address optical character recognition errors. The utilization of a LLM not only streamlines the dataset creation process but also ensures the accuracy and reliability of the data, thereby enhancing the credibility and trustworthiness of research findings derived from the dataset. LLM error correction is especially suited for declassified document archives which often consist of poor quality historical handwritten, typewritten, or photocopied documents due to the time

```
{
  "$schema": "http://json-schema.org/draft-07/schema#",
  "title": "Generated schema for DISC",
  "type": "array",
  "items": {
    "type": "object",
    "properties": {
      "DocID": {"type": "number"},
      "Title": {"type": "string"},
      "OCRtext": {"type": "string"},
      "Text": {"type": "string"},
      "Abstract": {"type": "string"},
      "Classification": {
        "type": "array",
        "items": {
          "type": "object",
          "properties": {
            "ClassID": {"type": "string"},
            "Label": {"type": "string"},
            "Date": {"type": "string"}
          },
          "required": ["ClassID", "Label"]
        }
      },
      "Database": {"type": "string"},
      "Domain": {"type": "string"},
      "Author": {"type": "string"},
      "StoreID": {"type": "string"}
    },
    "required": ["DocID", "Title", "Text1", "Text2", "Abstract", "Classification", "Database", "Domain
        ", "Author", "StoreID"]
  }
}
```

Figure 8: JSON Schema for DISC.

period requirement of information declassification. Furthermore, the experimentation results demonstrate the practical utility of DISC in reproducing prior research approaches. The research efforts in (Engelstad et al., 2015b; Brown and Charlebois, 2010; Engelstad et al., 2015a) leverage documents from the DNSA. However, the researcher don't explicitly state document selection process or list the documents select. DISC stores the unique database assigned document in the StoreID field and allows duplication of research datasets. Thereby, exactly reproducing prior experimental results is facilitated by listing StoreID values to recreate a subset dataset from DISC. This not only validates the reliability and consistency of DISC but also highlights its utility to serve as a valuable resource for bench-marking, comparison, and validation of information security classification algorithms and methodologies.

7 CONCLUSIONS

This paper presented DISC, the first reproducible information security classification dataset publicly available to information security researchers and professionals. By providing detailed instructions on how to curate this dataset from the original raw data, we not only enable others to reproduce the dataset, but also offer a framework for curating additional datasets from similar data sources. DISC contains information about 2,450 documents from the DNSA database, including 963 OFFICIAL, 1,341 SECRET, and 146 TOP SECRET documents. As discussed in Section 5, DISC can facilitate several areas of future research in information security classification. In addition, DISC can be utilized in various research purposes beyond information security classification such as implementing information control and granting access decisions. Future work would ideally include more documents incorporating additional DNSA database domains into the dataset following the same method-

ology and annotation. DISC facilitates collaboration, reproducibility, and innovation in future research in mitigating information security cybersecurity challenges. Overall, DISC represents a significant contribution to the information security classification research community, offering an accessible, reliable, and scalable resource for advancing research in this critical domain.

Using the proposed framework, we are currently working to include more documents in DISC. While the models tested in this paper have proven to achieve high performance in this task, we intend to evaluate the performance of open-source LLMs for this task. Expanding the DISC dataset creation framework to encompass open-source LLMs offers the capability to uphold the confidentiality of sensitive data during the processing of private information. We intent to utilize an extended version of DISC to refine recently introduced LLMs like Falcon and Llama-2 in crafting decision-making processes for information security classification levels. This endeavor will furnish the community with a vital resource for preserving confidentiality in classifying highly sensitive information. Finally, the framework presented in this paper can be applied to other domains and languages. We encourage the community to pursue research with data from other repositories (e.g., industry data) as well as on documents in languages other than English.

REFERENCES

Alzhrani, K., Rudd, E. M., Boult, T. E., and Chow, C. E. (2016). Automated big text security classification. In *Proceedings of the 2016 IEEE Conference on Intelligence and Security Informatics (ISI 2016)*, pages 103–108, Tucson, AZ, USA. IEEE.

Boustead, A. E. and Herr, T. (2020). Analyzing the ethical implications of research using leaked data. *Political Science and Politics*, 53(3):505–509.

Brown, J. D. and Charlebois, D. (2010). Security classification using automated learning (scale): Optimizing statistical natural language processing techniques to assign security labels to unstructured text. Technical Memorandum 2010-215, Defence R&D Canada – Ottawa.

Chakraborty, T., Jajodia, S., Katz, J., Picariello, A., Sperli, G., and Subrahmanian, V. S. (2021). A fake online repository generation engine for cyber deception. *IEEE Transactions on Dependable and Secure Computing*, 18(2):518–533.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In Burstein, J., Doran, C., and Solorio, T., editors, *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4171–4186, Minneapolis, USA. Association for Computational Linguistics.

Engelstad, P. E., Hammer, H., Yazidi, A., and Bai, A. (2015a). Advanced classification lists (dirty word lists) for automatic security classification. In *Proceedings of the 2015 International Conference on Cyber-Enabled Distributed Computing and Knowledge Discovery*, pages 44–53.

Engelstad, P. E., Hammer, H. L., Kongsgård, K. W., Yazidi, A., Nordbotten, N. A., and Bai, A. (2015b). Automatic security classification with lasso. In *Proceedings of the 16th International Workshop on Information Security Applications (WISA 2015)*, volume 9503 of *Lecture Notes in Computer Science*. Springer.

Information Security Oversight Office (2018). Developing and using security classification guides.

Jadli, A., Hain, M., Chergui, A., and Jaize, A. (2020). DCGAN-based data augmentation for document classification. In *Proceedings of the 2nd IEEE International Conference on Electronics, Control, Optimization and Computer Science (ICECOCS 2020)*.

Nakashima, E., Shepherd, C., and Cadell, C. (2023). Taiwan highly vulnerable to Chinese air attack, leaked documents show. *Washington Post*.

NIST (2004). FIPS 199: Standards for security categorization of federal information and information systems. Federal Information Processing Standards Publication 199, National Institute of Standards and Technology.

Peña, A., Morales, A., Fierrez, J., Serna, I., Ortega-Garcia, J., Puente, Í., Córdova, J., and Córdova, G. (2023). Leveraging large language models for topic classification in the domain of public affairs. In Coustaty, M. and Fornés, A., editors, *Proceeding of the 17th International Conference on Document Analysis and Recognition (ICDAR 2023)*, pages 20–33. Springer.

White House (2009). Executive order 13526: Classified national security information.

Whitham, B. (2017). Automating the generation of enticing text content for high-interaction honeyfiles. In *Proceedings of the 50th Hawaii International Conference on System Sciences (HICSS 2017)*.

APPENDIX

The database documents and associated information are stored within DISC in a JSON (JavaScript Object Notation) data structure based on the JSON Schema illustrated in Figure 8.