Creation and Analysis of a Natural Language Understanding Dataset for DoD Cybersecurity Policies (CSIAC-DoDIN V1.0)

Ernesto Quevedo
School of Eng. & Computer Science
Dep. of Computer Science
Baylor University
Email: ernesto_quevedo1@baylor.edu

Ana Paula Arguelles
School of Eng. & Computer Science
Dep. of Computer Science
Baylor University
Email: ana_arguelles1@baylor.edu

Alejandro Rodriguez
School of Eng. & Computer Science
Dep. of Computer Science
Baylor University
Email: alejandro_rodriguez4@baylor.edu

Jorge Yero
School of Eng. & Comp. Sci.
Dep. of Computer Science
Baylor University
Email: jorge_yero1@baylor.edu

Dan Pienta
Haslam College of Business
Dep. Acc. & Infor. Manag.
University of Tennessee
Email: dpienta@utk.edu

Tomas Cerny
School of Eng. & Comp. Sci.
Dep. System & Industrial Eng.
University of Arizona
Email: tcerny@arizona.edu

Pablo Rivas , *Senior, IEEE*
School of Eng. & Comp. Sci.
Dep. of Computer Science
Baylor University
Email: pablo_rivas@baylor.edu

***Abstract*—Recent studies in Legal NLP showed the lack of structured data to train Deep Learning models in several tasks. With the increased importance of privacy policies in the current digital world, the research community released multiple datasets related to privacy policies in the last few years. However, other empirical studies have shown the lack of transferability between domain-specific language models in a legal subdomain to other more separate subdomains. With the focus on privacy policies, models are not tested on other policies. In this work, we release the CSIAC-DoDIN V1.0 dataset, focused on cybersecurity policies, responsibilities, and procedures of the organizations involved. This first version offers classic Legal NLP tasks such as several Multiclass Classification tasks and text co-occurrence. Furthermore, we also provide a baseline for this dataset and tasks with experiments using classic transformer-based language models such as BERT, RoBERTa, Legal-BERT, and PrivBERT.**

***Index Terms*—dataset, large language models , legal natural language processing, machine learning, deep learning**

I. Introduction

Cybersecurity policy documents enclose the policies, responsibilities, and procedures an individual or organization should do to protect digital assets like networks, devices, data, and systems from unauthorized access, modification, disclosure, and destruction. Finding or understanding a policy is one of many challenges. A big problem, like in the case of privacy policies [1], [2], is the length and language complexity of these documents. One also needs to comprehend the responsibilities and procedures that must be followed. Moreover, creating resources to classify and cluster an extensive set of cybersecurity policy documents requires a lot of manual work.

The recent advances in Legal NLP provide an opportunity to streamline the automatic comprehension of cybersecurity policies. High-level natural language understanding (NLU) applications could assist legal practitioners in creating these policies by providing an automated approach to classifying and grouping them. A fine-tuned large language model could also assist in the writing process and quality check of the entire document. The benefits affect legal practitioners and the organization and individuals impacted by the cybersecurity policy document. Legal NLP could provide automatic question-answering and information retrieval. A knowledge graph could be built if excellent performance is reached in named entity recognition and relation extraction. This is easy to tackle since these documents are usually well-structured and contain all references to other related documents.

Recent research in Legal NLP is mainly focused on privacy policies which is just a subdomain of the cybersecurity policies and the proper procedures to protect against malicious intentions that can provoke a big disaster. In addition, existing datasets do not include guidances, responsibilities, or procedures, just the policies to follow. Furthermore, the need for more structured data to train Deep Learning models [3] is a general problem in Legal NLP, and the field of policies does not escape this problem. This paper introduces a new dataset named Cyber Security and Information Systems Analysis Center Department of Defense Information Networks (CSIAC-DoDIN) V1.0, made up of Cybersecurity-Related Policies and Issuances developed by the DoD Deputy CIO for Cybersecurity [4].[1] To the best of our knowledge, it is the first of its kind includes policies, guidelines, strategies, responsibilities, and procedures. Since our dataset is about cybersecurity policies, the language domain is vaster than the one provided by privacy policies. It could give the means to train and obtain better models.

The dataset is publicly available here [5]:

https://doi.org/10.6084/m9.figshare.22800185.v1

[1] https://dodiac.dtic.mil/dod-cybersecurity-policy-chart/

The dataset highlights the DoD's robust cybersecurity policies, owing to its strategic significance and sensitive global operations. These policies offer insights into premier cybersecurity best practices and methodologies. The chart also incorporates standards from the National Institute of Standards and Technology (NIST) and ISO, which are influential across industries. Many companies look to the DoD as a model for developing their cybersecurity and privacy policies. The paper clarifies that these policies aren't exclusive to the DoD.

Using this dataset, we created a set of Multiclass-Classification tasks and Text-Co-Occurrence. Also, we provide a baseline with four classic transformer-based language models such as BERT, RoBERTa, Legal-BERT, and PrivBERT, on these tasks, which allowed the comparison of general and domain-specific pre-trained language models. This will serve as the starting point for future works using this dataset.

The contributions of this paper can be summarized as:

1) Introduced a new curated dataset conformed to Cybersecurity-Related Policies and Issuances developed by the DoD Deputy CIO for Cybersecurity.
2) Provide a baseline with four classic transformer-based language models such as BERT, RoBERTa, Legal-BERT, and PrivBERT, applied to the Multiclass-Classification and Text Co-Occurrence tasks obtained from the dataset.
3) Provide open access to the dataset and code to train and evaluate the baselines, making it easier to build upon this work and test new custom models or pre-trained ones.[2]

This paper is structured as follows. First, we present the related works. Second, we describe our dataset creation methodology and provide descriptive statistics and the Legal NLP tasks we created. Next, we offer the experiments performed and the results obtained from them. After that, we present a discussion and future work section, followed by the conclusions.

II. Related Work

In the last few years, there has been rapid growth in legal text processing and analysis. Multiple datasets, benchmarks, and tasks have been created [1], [2], [6]–[10]. However, they are focused on something other than cybersecurity policies.

Among the legal subdomains that have gained popularity is the privacy policies subdomain. Privacy policies are popular datasets like [6], which produces a comprehensive labeled list of privacy policy documents and categories for classification. In addition, works like [9] curated a dataset of 1,071,488 English language privacy policies spanning over two decades and over 130,000 distinct websites. Furthermore, future works have reused these datasets to create other NLP tasks besides Multiclass Classification, like Question Answering [2], and evaluated big language models like Bidirectional Encoder Representations from Transformers (BERT) [11] on them.

A recent work [1] introduced a benchmark dataset for general language understanding in privacy policies called PrivacyGLUE that contains several of the existent datasets of privacy policies. Furthermore, the authors also provided an analysis and comparison of the performance of multiple transformer-based language models like BERT, RoBERTa [12], Legal-BERT [13] and PrivBERT [14].

In a more generalized legal domain, the research [10] produced a Benchmark Dataset for legal language understanding in English called LexGLUE that has multiple legal language datasets. The authors also analyzed and compared multiple transformer-based language models in each of the designed NLU tasks. The transformer-based language models evaluated are BERT, RoBERTa, DeBERTa [15], Longformer [16], BigBird [17], Legal-BERT, CaseLaw-BERT [18].

The research on both benchmarks [1], [10] showed the disparity among legal subdomains and how a domain-specific model like Legal-BERT or PrivBERT could be outperformed by general-purpose models in some tasks and datasets. This increases the importance of having multiple curated and labeled datasets on different subdomains of the legal field, besides the paucity of legal domain datasets in recent literature [3], [19].

Privacy policies are gaining all popularity, and multiple datasets are being created when any other type of policies and similar documents are being ignored. Therefore, we decided to introduce a new dataset focused on cybersecurity policies with this work. In addition, we include many documents and guidance related to cybersecurity and provide a baseline in our dataset for further research.

III. Dataset

This section describes the knowledge base used to build these datasets, the annotation scheme and structure of the corpus, the Legal NLP tasks created, and the description of the different versions of the dataset according to the Legal NLP task.

A. Knowledge Base Description

As the knowledge base for building this dataset, we used a chart that clusters and classifies the Cybersecurity-Related Policies and Issuances developed by the DoD Deputy CIO for Cybersecurity.[3] The chart organizes the cybersecurity policies and guidance documents by Strategic Goal and Office of Primary Responsibility. The chart captures many applicable policies with a grand organizational scheme. This organizational scheme provides the distribution in separate clusters of each document and its content. Leveraging this ensures our dataset encompasses a broad spectrum of policies from a globally influential entity in cybersecurity. The meticulous upkeep of this knowledge base assures the policies we have included are current and pertinent. Furthermore, the DoD policies on the Chart also include policies from the National Institute of Standards and Technology (NIST), which is widely applied by industry. The Chart's design includes specific and comprehensive DoD and "best practice" industry policies, such as the NIST, and ISO-type policies for organizations to use for security and privacy. Often, the NIST serves as a general benchmark for industry to judge policy.

Additionally, the DoD Cybersecurity Policies have a great global relevance. The DoD holds immense strategic significance

[2] https://github.com/Fidac/DoD-Document-Classification.git

[3] https://dodiac.dtic.mil/dod-cybersecurity-policy-chart/

TABLE I: Subclusters corresponding to each cluster and its description

Subcluster	Outer Cluster	Description
Lead and Govern	Organize	Provide vision and follow through to set the enterprise direction, foster a culture of accountability, and provide insight and oversight for the enterprise.
Design for the Fight	Organize	Deliver, synchronize and integrate capabilities across the organization in time by shaping capabilities, engineering for the entire enterprise, leveraging technology, investing for success, and balancing risk.
Develop the Workforce	Organize	Provide a learning continuum to recruit, retain, and educate qualified professionals while keeping capabilities current through education and training, proper structure of the workforce, and the cultivation of awareness of initiatives.
Partner for Strength	Organize	Leverage the unique capabilities of partners from various areas such as intra-goverment, academia, cybersecurity and IT industry, defense industry, and international/global partners.
Secure Data in Transit	Enable	Provide robust, state-of-the-art cryptographic products and key management services for secure data transmission.
Manage Access	Enable	Provide secure, authenticated access to authorized users for proper visibility, configuration, connection, and allocation of resources through managed identity credentials, privileges, and resources.
Assure Information Sharing	Enable	Allow for secure and seamless information flow and management across security domains by assuring publishing, discovery, and collaboration.
Understand the Battlespace	Anticipate	Align and leverage information from audits, sensors, forensics, and incident management across an enterprise through knowing adversaries, networks, and consequences.
Prevent and Delay Attackers and Prevent Attackers from Staying	Anticipate	Leverage knowledge of networks, vulnerabilities, and adversaries to harden systems, defend perimeters, and assess defenses. Lower adversarial capabilities through detecting, diagnosing, eliminating, preventing, and constraining attacks.
Develop and Maintain Trust	Prepare	Guarantee integrity and availability of systems by assuring use, engineering for survivability, and maintaining integrity.
Strengthen Cyber Readiness	Prepare	Harden response procedures by linking units across the enterprise, stress testing response procedures, identifying critical assets, and improving continuity planning.
Sustain Missions	Prepare	Enable enterprise mission with limited interruption during an attack by assessment for fighting through adverse events, sustaining critical systems during degradation, and rapidly restoring systems to a trusted state.

in cybersecurity, setting benchmarks for various entities. While our dataset might not capture the entire global breadth of all world policies, it provides in-depth insights into a sophisticated cybersecurity framework with wide-scale use in the industry. This dataset facilitates comparative research, enabling an analysis of DoD policies against other frameworks and spotlighting unique attributes and alignment areas.

B. Annotation Scheme

The annotation scheme for each document and policy inside is based on the organization provided in the chart. In 2009, the Department of Defense (DOD) developed clusters directed at the goal the policy was designed to achieve as part of its information assurance strategy and ensure proper unit-level capabilities. This scheme is made up of seven outer clusters and twelve subclusters. Each outer cluster is one class the document belongs to, and the subcluster is another. Table I shows the subclusters corresponding to each cluster and its description. The seven outer clusters are described as follows:

Organize: Policies in this cluster relate to how enterprise units (e.g., departments) can organize for unity and purpose. These policies provide guidance to ensure unit capabilities are designed, organized, and managed so that capabilities are synergistic, flexible, and dynamic in responding to the demands of any event and can support the shared objectives of the whole enterprise.

Enable: Policies in this cluster relate to the access of information. These policies provide guidance to ensure that information is available to authorized parties but also protected from adversaries. The policies are designed so that all units have proper visibility, control, and management of information assets in a secure manner.

Anticipate: Policies in this cluster relate to how enterprise units can anticipate and prevent attacks on data and networks. These policies provide guidance on how to stop attacks outside the network perimeter of an enterprise but also allow for perimeter flexibility and maneuverability when needed. Additionally, the policies outline guidance for secondary defense if a network perimeter is breached.

Prepare: Policies in this cluster relate to how enterprise units prepare and operate during a successful data breach or cyber-attack. The policies provide guidance for increasing system resiliency by ensuring cyber assets self-monitor, self-attest, and self-repair. The policies note that units that experience a cyber-attack have assurances that the enterprise remains functional or has a plan in the event of complete system degradation.

Authorities: This cluster of policies outlines which agency or authority (e.g., Department of Defense, United States Coast Guard) the policy is applicable. Additionally, this outlines the acting authority of policy that is to be followed in the event of multiple entities from different enterprises working collectively.

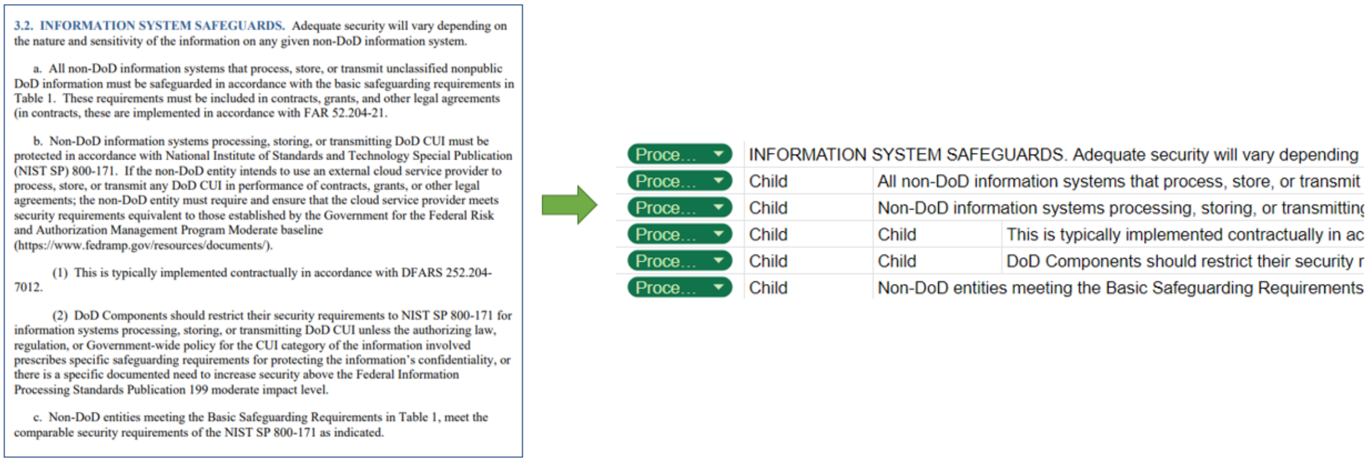
National/Federal: This cluster of policies relates to national and federal policies that provide guidance for actions for entities at that level.

Operational/Subordinate Policy: This cluster relates to policies that are for specific entities that might be used in coordination with a superseding policy.

C. Extraction and Annotation process

Each annotator got assigned a set of outer clusters and followed the following protocol:

- Access the link on the chart and find the correct document.
- Once the document is found, determine if the document is guidance, strategy, or policy.
- If the document is guidance or strategy, extract the document PDF.
- If the document is a policy, then extract the PDF and also go through the document and extract every policy, responsibility, and procedure. Also, extract the general Purpose, Scope and Applicability of the document.
- If the link wasn't accessible or there were other reasons not to access the PDF, report and ignore that document.



3.2. INFORMATION SYSTEM SAFEGUARDS. Adequate security will vary depending on the nature and sensitivity of the information on any given non-DoD information system.

a. All non-DoD information systems that process, store, or transmit unclassified nonpublic DoD information must be safeguarded in accordance with the basic safeguarding requirements in Table 1. These requirements must be included in contracts, grants, and other legal agreements (in contracts, these are implemented in accordance with FAR 52.204-21.

b. Non-DoD information systems processing, storing, or transmitting DoD CUI must be protected in accordance with National Institute of Standards and Technology Special Publication (NIST SP) 800-171. If the non-DoD entity intends to use an external cloud service provider to process, store, or transmit any DoD CUI in performance of contracts, grants, or other legal agreements; the non-DoD entity must require and ensure that the cloud service provider meets security requirements equivalent to those established by the Government for the Federal Risk and Authorization Management Program Moderate baseline (https://www.fedramp.gov/resources/documents/).

(1) This is typically implemented contractually in accordance with DFARS 252.204-7012.

(2) DoD Components should restrict their security requirements to NIST SP 800-171 for information systems processing, storing, or transmitting DoD CUI unless the authorizing law, regulation, or Government-wide policy for the CUI category of the information involved prescribes specific safeguarding requirements for protecting the information's confidentiality, or there is a specific documented need to increase security above the Federal Information Processing Standards Publication 199 moderate impact level.

c. Non-DoD entities meeting the Basic Safeguarding Requirements in Table 1, meet the comparable security requirements of the NIST SP 800-171 as indicated.

Proce... ▾	INFORMATION SYSTEM SAFEGUARDS. Adequate security will vary depending		
Proce... ▾	Child	All non-DoD information systems that process, store, or transmit	
Proce... ▾	Child	Non-DoD information systems processing, storing, or transmittin	
Proce... ▾	Child	Child	This is typically implemented contractually in ac
Proce... ▾	Child	Child	DoD Components should restrict their security r
Proce... ▾	Child	Non-DoD entities meeting the Basic Safeguarding Requirements	

Fig. 1: example of a set of procedures on the left extracted and annotated on the excel file on the right.

After analyzing the documents with a legal expert, we noticed that the policy documents were structured with the following sections: Purpose, Authority, Scope, Policies, Responsibilities, Procedures, Definitions, and References. With further discussion with a legal expert, the type of text containing the policy documents' essence is enclosed in the Purpose, Scope, Policies, Responsibilities, and Procedures section. Therefore, in this first version, we focused the extraction on those sections. However, if deemed useful, the full document is also included in the dataset. Therefore, every item extracted from a document is classified by type as a policy, responsibility, or procedure. As a better description of each class, we can say:

Policy: A policy is designed to set parameters for decision-makers while allowing for flexibility for decision-makers.

Responsibility: Responsibilities within the policy designate which organization members are accountable to ensure the policy is adhered to.

Procedure: A procedure within the policy provides step-by-step instructions for performing a routine task.

The annotator would search for the Policies, Responsibilities, and Procedures sections if the document were a policy. Then they would iterate through each item in those sections and put them in an Excel file with the fields of `Id`, `Cluster`, `Classification`, `Purpose`, `Scope and Applicability`, `Type`, `Text`, and several columns with the name `Child_Level#`.

The values in the Cluster column correspond to which cluster the document belongs to in the chart. The Classification column will have the name of the subcluster the document belongs to, if any. The type column would have the values of Policy, Responsibility, or Procedure, according to which section of the document the text was extracted.

We decided to keep the structure of the text in our dataset. This is the use of the Child_Level# columns. Several policies, responsibilities, and procedures usually have subitems and can end with a tree form of a few levels of depth. Therefore, we represented these trees in Excel instead of just putting in our dataset the final text that will concatenate the text from the root with the text of every node in the path to a leaf. Fig. 1 shows how the text of a document on the left and how those procedures are extracted in the Excel file. This way, we could keep the structural form of the document, which opened the path to one of the tasks we created.

The rationale for including guidances, responsibilities, and procedures with policies clarifies their implementation. Responsibilities establish a clear chain of command for accountability during incidents. Procedures offer step-by-step directions to ensure policy objectives are met. Without such procedures, even detailed policies can be ineffective due to the lack of clear directives. Moreover, guidances, responsibilities, and procedures enhance a Large Language Model's understanding of a policy and its implications.

D. Legal NLP Tasks created

Using this dataset as a base, we created a set of Legal NLP tasks like Multiclass-Classification and Text-Co-Occurrence. Next, we list the set of tasks we proposed, but the dataset is broader than these.

Cluster/Subcluster Classification: Determine if a given policy, responsibility, or procedure belongs to a particular cluster or subcluster.

Type Classification: Determine if a text is a policy, responsibility, or procedure.

Purpose-Text Co-Occurrence: Determine if a given policy, responsibility, or procedure co-occurs a Purpose of a document.

Scope/Applicability-Text Co-Occurrence: Determine if a given policy, responsibility, or procedure co-occurs with the Scope/Applicability of a document.

Text-Text Co-Occurrence: Determine if a given subpart of a policy, responsibility, or procedure co-occurs with another subpart of a policy responsibility or procedure.

Co-occurrence tasks focus on context relevance, ensuring policies, responsibilities, and procedures align with a document's purpose and scope. Our Text-Text task aims to identify if a subitem semantically relates to its preceding context.

TABLE II: Documents distribution and extraction results by subcluster or cluster.

Subcluster/Cluster	Excels	Just PDF	Missing	Total
Lead and Govern	0	23	0	23
Design for the Fight	15	3	6	24
Develop the Workforce	8	2	4	14
Partner for Strength	4	3	3	10
Secure Data in Transit	17	2	3	22
Manage Access	13	3	8	24
Assure Information Sharing	6	0	0	6
Understand the Battlespace	3	5	0	8
Prevent and Delay Attackers and Prevent Attackers from Staying	10	9	7	26
Develop and Maintain Trust	4	0	2	6
Strengthen Cyber Readiness	9	7	0	16
Sustain Missions	10	2	8	20
Authorities	3	3	2	8
National/Federal	21	11	4	36
Operational/Subordinate	0	0	6	6
Total	123	73	53	249

Verification and comprehensiveness are also vital, involving automated checks for misplaced items within a document and verifying new items' compatibility with the existing context.

Improving policy design is crucial. Identifying patterns in document co-occurrences offers key insights, enabling designers to craft more effective policies with all necessary components. Given the growing complexity and volume of cybersecurity policies, there's a clear need for automated guidance. Large Language Models trained in co-occurrence tasks can guide professionals by suggesting aligned responsibilities or procedures for new or revised policies and providing insights into potential additions for current policies and procedures.

E. Composition of the Corpus

This section will describe the dataset's statistics and variants created depending on the Legal NLP task.

Table II shows how many documents by subcluster or cluster[4] from the chart we were able to extract labeled text, how many we only extracted the PDF, and how many we could not access. After the extraction process was finished, Table III shows the total of 7698 examples in the dataset and how they are distributed across the three types: Policy, Responsibility, and Procedures. Fig. 2 bar chart shows how many examples from the dataset are in each cluster. It is essential to highlight that every policy, responsibility, or procedure belongs to a cluster. However, the chart has clusters that do not have other subdivisions. Therefore, some of the 7698 examples are not assigned to a subcluster but just an outer cluster. Table V provides the distribution of the dataset concerning each subcluster. You might notice that the Lead and Govern subcluster has 0 text examples assigned. This is because all the documents in this subcluster were strategy documents, which is different from the focus of this dataset at the moment.

TABLE III: Dataset distribution by type

Type	Frequency
Policy	1531
Responsibility	4175
Procedures	1992
Total	7698

TABLE IV: Positives examples in each Text-Co-Occurrence task.

Task	Frequency
Purpose-Text	7197
Scope/App-Text	6617
Text-Text	8355

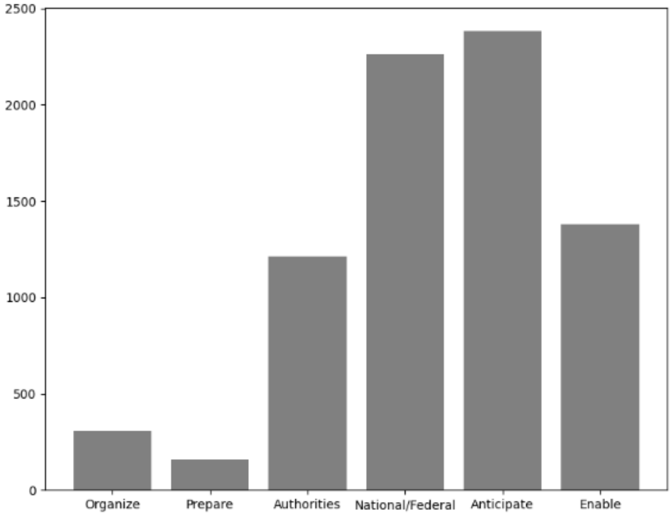


Fig. 2: Distribution of examples in the dataset by cluster.

TABLE V: Dataset distribution by subcluster

Subcluster	Frecuency
Lead and Govern	0
Design for the Fight	1002
Develop the Workforce	107
Partner for Strength	102
Secure Data in Transit	663
Manage Access	816
Assure Information Sharing	903
Understand the Battlespace	96
Prevent and Delay Attackers and Prevent Attackers from Staying	2165
Develop and Maintain Trust	43
Strengthen Cyber Readiness	151
Sustain Missions	112
Total	6160

In addition, Table IV shows for the Text-Co-Occurrence tasks how many positive examples were available for each task. It is important to notice that there are fewer examples in the Purpose-Text and Scope/App-Text tasks because some documents did not have a Purpose or Scope/App section. Furthermore, there are more Text-Text examples because we consider an example as the entire path from a root to a leaf in the Multiclass-Classification datasets. In the case of the Text-Text Co-Occurrence task, the positive examples are built by taking every tree built from a document policy, responsibility, or procedure, and every edge is a pair of Text-Text where the child co-occurs with the parent. For all the Text-Co-Occurrence tasks, we used a negative sampling strategy described as follows: (i) take the premise of a positive example at random; (ii) select the hypothesis of a different positive example that does not have the same premise; (iii) create a negative example using the randomly picked premise and hypothesis; (iv) continue this process until there are as many negative examples as positives.

[4] In case the cluster does not have any subcluster.

IV. Experiments and Results

In this section, we describe the evaluation of five transformer-based language models that have achieved state-of-the-art performance in most NLP tasks [20]. These models are trained in vast amounts of unlabeled text on Mask Language Modeling (MLM) and Next Sentence Prediction (NSP) tasks. We fine-tuned each of the following models in the particular version of the dataset corresponding to the task we were evaluating.

BERT: Most popular transformer language model proposed by [11]. It is trained in MLM and NSP tasks on the Wikipedia[5] and Bookcorpus [21] datasets.

RoBERTa: [12] implemented the RoBERTa model to improve BERT using a larger vocabulary and a dynamic masking technique to eliminate the NSP task. It was pre-trained on the same datasets as BERT.

Legal-BERT: Is another BERT-based model by [13] pre-trained from scratch on English legal data consisting of contracts, legislation, and court cases. The data sources are cited in the original paper and the Hugging Face Model Card.[6] The sub-word vocabulary of Legal-BERT is built from scratch with additions to legal terminology.

PrivBERT: A RoBERTa-based model [14]. It was pre-trained from scratch on one million privacy policy documents.[7]

We use the selected pre-trained models publicly available in Hugging Face.[8] Specifically, we used their base configurations with 12 Transformer blocks, 768 hidden units, and 12 attention heads. We train all the models using the Adam optimizer [22] and a learning rate of $5 * 10^{-5}$ during six epochs since, after several experiments, the models will reach their peak at the fourth or fifth epoch in every task. In the Multiclass-Classification tasks, we divided the dataset by randomly taking 60% of the examples in each class for training, 15% for validation, and 25% for testing; we selected this distribution to keep as much balance as possible and avoid overfitting to a class. In the Text-Co-Occurrence tasks, the datasets are balanced; we took 60% of the examples in each category for training, 15% for validation, and 25% for testing with the guarantees that all are balanced.

For the evaluation of the performance of all the models in all the tasks (Multiclass-Classification and Text-Co-Occurrence tasks), we used the *micro-F1* (μ-F1) and *macro-F1* (m-F1) metrics to take into account imbalance. Furthermore, for completeness and following the practice of several works [1], [10], [23], we report arithmetic, harmonic, and geometric means.

TABLE VI: Test results for all examined models across all Multiclass-Classification tasks.

Method	Type		Cluster		Subcluster	
	μ-F1	m-F1	μ-F1	m-F1	μ-F1	m-F1
BERT	0.963	0.95	0.96	0.921	0.911	0.734
RoBERTa	0.96	0.947	0.954	0.925	0.923	0.74
Legal-BERT	**0.969**	**0.96**	0.966	**0.954**	0.921	0.733
PrivBERT	**0.969**	0.959	**0.967**	0.938	**0.931**	**0.741**

TABLE VII: Test results for all examined models across all Text Co-Occurrence tasks.

Method	Purpose-Text		Scope/App-Text		Text-Text	
	μ-F1	m-F1	μ-F1	m-F1	μ-F1	m-F1
BERT	**0.749**	**0.735**	**0.581**	**0.503**	**0.715**	**0.699**
RoBERTa	0.539	0.427	0.494	0.397	0.607	0.556
Legal-BERT	0.738	0.721	0.574	0.495	0.688	0.666
PrivBERT	0.533	0.414	0.491	0.347	0.55	0.464

Table VI shows the models' results in all the dataset variants of the type Multiclass-Classifcation, and Table VII shows the models' results in all the dataset variants of the type Text Co-Occurrence. In addition, Table VIII presents the aggregated (averaged) results. Table VI shows how Legal-BERT and PrivBERT dominate on all tasks. This suggests that legal domain-specific models possess good transferability to the cybersecurity policies domain and outperform general-purpose models for some Multiclass-Classification tasks. However, in the case of the Text-Co-Occurrence tasks, BERT model dominated in all of them and all metrics. We believe that, in the case of RoBERTa and PrivBERT, the reason for such low performance is that they were not pre-trained in the NSP task. Still, BERT outperforms Legal-BERT, suggesting that domain transferability from models like Legal-BERT and PrivBERT to other domains is low for more complex tasks. Also, these results indicate the importance of including NSP as one of the pre-training tasks of a new transformer-based LM. Finally, based on the aggregated results in Table VIII, BERT performs overall better in all metrics. However, Legal-BERT is quite close, but it highlights how the difference in the performance of BERT and Legal-BERT in Text-Co-Occurrence tasks was more significant than in Multiclass-Classification tasks where Legal-BERT outperforms BERT.

V. Discussion and Future Work

To begin our discussion, it's vital to understand the implications of cybersecurity-related policies for organizations. Failures to grasp these policies can lead to data breaches [24], regulatory fines [25], loss of consumer trust [26], and reduced firm value [27]. Information security policies need to be clearer, more consistent, and intricate [28]. There are often misunderstandings between policy-makers and implementers [29], [30], and policy complexity can cause stakeholder stress and confusion, resulting in potential policy breaches [31], [32]. With a growing focus on cybersecurity, there's a shortage of clear policies from various regulators and organizations, making it challenging for executives to ensure compliance [30]. Many organizations remain unaware of relevant policies for cybersecurity incidents, risk mitigation strategies, or data breach remediation. This dataset aims to help both academia and industry in understanding and effectively applying policies, reducing misconceptions and negligence. The paper and data are pivotal, as security policies safeguard individuals, organizations, and society at large.

[5] https://dumps.wikimedia.org

[6] https://huggingface.co/nlpaueb/legal-bert-base-uncased

[7] https://privaseer.ist.psu.edu/data

[8] https://huggingface.co/models

TABLE VIII: Test results aggregated over all tasks: arithmetic (A), harmonic (H) and Geometric (G) mean.

	A-Mean		**H-Mean**		**G-Mean**	
Method	**μ-F1**	**m-F1**	**μ-F1**	**m-F1**	**μ-F1**	**m-F1**
BERT	**0.813**	**0.757**	**0.786**	**0.724**	**0.799**	**0.741**
RoBERTa	0.746	0.665	0.689	0.591	0.717	0.627
Legal-BERT	0.809	0.754	0.778	0.717	0.794	0.736
PrivBERT	0.74	0.643	0.676	0.549	0.707	0.595

As discussed in the previous section, we just explored a small set of the big family of transformer-based language models to be applied to this dataset. Ensemble combinations of these models and trying other models pre-trained in similar legal data are just a few of the future directions that can be followed to increase performance. Furthermore, it opens the door to investigating the correlation between cybersecurity and privacy policies. For example, we evaluated a model trained only on privacy policies like PrivBERT. Still, the entire text of our dataset could be used with more information on the internet to train a model expert in cybersecurity policies (CyberSecBERT) and evaluate it in privacy policies which provide insights in both directions of the transferability between both domains. Furthermore, this work strengthens the argument for more transferability among legal subdomains in some tasks. However, it showed that domain-specific legal LM and privacy policies in some Multiclass-Classification tasks perform better in cybersecurity policies. The low performance of all the models in the Text-Co-Occurrence tasks is because of their complexity. Some policies, responsibilities, or procedures are short and lack context to ensure they relate to a Purpose, Scope/Applicability, or if they are likely to come after a particular statement.

The current dataset is an initial version, with potential for expansion through the addition of new categories and further exploration of guidance documents accessible via the chart. It can be enhanced with tasks such as Named Entity Recognition (NER), Relation Extraction, Question Answering (QA), and Information Retrieval (IR). From the chart, a citation network can be constructed, referencing structures like those in [33]. The dataset presently contains 196 cybersecurity documents from the DoD in PDF format.

VI. Conclusions

In this paper, we discuss the importance of cybersecurity policies in modern digital life and highlight the need for datasets based on them, given the current progress in Legal NLP and exploration of subdomains within the legal field. Most of the research community is exploring privacy policies which are only one of the topics explored in cybersecurity policies. Our contribution to this issue was the creation of the CSIAC-DoDIN (V1.0) dataset that conformed to Cybersecurity-Related Policies and Issuances developed by the DoD Deputy CIO for Cybersecurity. In addition, we released baseline performances using classic and domain-specific transformer language models like BERT, RoBERTa, Legal-BERT, and PrivBERT. Our results showed good transferability from legal domain-specific LM to the cybersecurity policies in Multiclass-Classification tasks, but this is not the case on Text-Co-Occurrence tasks. Finally, we shared our code and dataset for future experimentation and reproducibility.

Limitations

Although our dataset is just based on DoD cybersecurity policies which is a small set of all the cybersecurity policies in the world, it will be a representative set of English cybersecurity policies datasets that will be created in the future and possibly a benchmark like [1]. In the current version of our dataset, only the English language can be evaluated, and we only provide Multiclass-Classification and Text-Co-Occurrence tasks. However, the dataset can be extended with more tasks. In addition, even when the dataset was built from a human knowledge base, it cannot still be considered a human expert performance comparison with any of the tasks provided in this dataset. We intend to collaborate with the DoD to assess human performance on these tasks in future works.

As internal threats, we have possible problems that might have arisen during the data extraction process. Extracting data is always a complex process that, in our case, could imply misclassified examples. We mitigate the miss classification since we are using the chart from the DoD as a knowledge base, and all the categories do not come from the annotators but are already present in the chart. We just adapted the chart to a processable format for NLP algorithms. However, the authors discussed with a legal expert to assess the correct classification in the few cases when it needed clarification.

The models we provided their performance in the dataset's tasks are only a small set of all the transformer language models currently out there. Also, we made the naive approach in all the tasks with a basic Hugging Face pipeline and used the [CLS] token encoding for classification to provide a baseline. Therefore, the results obtained in each model also carry the model's and our approach limitations. Furthermore, any conclusion from this paper cannot be generalized to any other NLP domain or out of this dataset's scope.

Ethics Statement

This dataset can be the base and inspiration to create new data related to cybersecurity policies. Furthermore, it can result in the automation of complex tasks like question answering and information retrieval. To put just an example, any new document created in the future could be automatically placed in the chart. All this data can be used to obtain high-performance transformer language models that can be used to develop applications to help users to understand cybersecurity-related policies, which tend to be sometimes long, complex, confusing, and inconsistent [28]. However, it is always important to consider the risks of transformer language models. Not only the misclassifications and errors, but they also discuss cases of bias, gender bias, toxicity language, and information leaking [34]–[36]. Especially the latter is a high risk when dealing with cybersecurity policies since this is sensitive data, and also some users might consider using these models in classified documents.

References

[1] A. Shankar, A. Waldis, C. Bless, M. Andueza Rodriguez, and L. Mazzola, "Privacyglue: A benchmark dataset for general language understanding in privacy policies," *Applied Sciences*, vol. 13, no. 6, p. 3701, 2023.

[2] W. U. Ahmad, J. Chi, Y. Tian, and K.-W. Chang, "Policyqa: A reading comprehension dataset for privacy policies," *arXiv preprint arXiv:2010.02557*, 2020.

[3] D. Song, A. Vold, K. Madan, and F. Schilder, "Multi-label legal document classification: A deep learning-based approach with label-attention and domain-specific pre-training," *Information Systems*, vol. 106, p. 101718, 2022.

[4] R. F. Lentz, "Deputy assistant secretary of defense for cyber, identity, and information assurance strategy. 1-32," 2009.

[5] E. Q. Caballero, P. Rivas, A. P. Arguelles, A. Rodriguez, J. Yero, D. Pienta, and T. Cerny, "Natural Language Understanding Dataset for DoD Cybersecurity Policies (CSIAC-DoDIN V1.0)," 11 2023. [Online]. Available: https://doi.org/10.6084/m9.figshare.22800185.v1

[6] S. Wilson, F. Schaub, A. A. Dara, F. Liu, S. Cherivirala, P. G. Leon, M. S. Andersen, S. Zimmeck, K. M. Sathyendra, N. C. Russell *et al.*, "The creation and analysis of a website privacy policy corpus," in *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2016, pp. 1330–1340.

[7] A. Kornilova and V. Eidelman, "Billsum: A corpus for automatic summarization of us legislation," *arXiv preprint arXiv:1910.00523*, 2019.

[8] A. Z. Wyner, B. J. Fawei, and J. Z. Pan, "Passing a usa national bar exam: a first corpus for experimentation," in *LREC 2016, Tenth International Conference on Language Resources and Evaluation*. LREC, 2016.

[9] R. Amos, G. Acar, E. Lucherini, M. Kshirsagar, A. Narayanan, and J. Mayer, "Privacy policies over time: Curation and analysis of a million-document dataset," in *Proceedings of the Web Conference 2021*, 2021, pp. 2165–2176.

[10] I. Chalkidis, A. Jana, D. Hartung, M. Bommarito, I. Androutsopoulos, D. M. Katz, and N. Aletras, "Lexglue: A benchmark dataset for legal language understanding in english," *arXiv preprint arXiv:2110.00976*, 2021.

[11] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint arXiv:1810.04805*, 2018.

[12] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov, "Roberta: A robustly optimized bert pretraining approach," *arXiv preprint arXiv:1907.11692*, 2019.

[13] I. Chalkidis, M. Fergadiotis, P. Malakasiotis, N. Aletras, and I. Androutsopoulos, "Legal-bert: The muppets straight out of law school," *arXiv preprint arXiv:2010.02559*, 2020.

[14] M. Srinath, S. Wilson, and C. L. Giles, "Privacy at scale: Introducing the privaseer corpus of web privacy policies," *arXiv preprint arXiv:2004.11131*, 2020.

[15] P. He, X. Liu, J. Gao, and W. Chen, "Deberta: Decoding-enhanced bert with disentangled attention," *arXiv preprint arXiv:2006.03654*, 2020.

[16] I. Beltagy, M. E. Peters, and A. Cohan, "Longformer: The long-document transformer," *arXiv preprint arXiv:2004.05150*, 2020.

[17] M. Zaheer, G. Guruganesh, K. A. Dubey, J. Ainslie, C. Alberti, S. Ontanon, P. Pham, A. Ravula, Q. Wang, L. Yang *et al.*, "Big bird: Transformers for longer sequences," *Advances in neural information processing systems*, vol. 33, pp. 17 283–17 297, 2020.

[18] L. Zheng, N. Guha, B. R. Anderson, P. Henderson, and D. E. Ho, "When does pretraining help? assessing self-supervised learning for law and the casehold dataset of 53,000+ legal holdings," in *Proceedings of the eighteenth international conference on artificial intelligence and law*, 2021, pp. 159–168.

[19] Z. Shaheen, G. Wohlgenannt, and D. Mouromtsev, "Zero-shot cross-lingual transfer in legal domain using transformer models," in *2021 International Conference on Computational Science and Computational Intelligence (CSCI)*. IEEE, 2021, pp. 450–456.

[20] R. Bommasani, D. A. Hudson, E. Adeli, R. Altman, S. Arora, S. von Arx, M. S. Bernstein, J. Bohg, A. Bosselut, E. Brunskill *et al.*, "On the opportunities and risks of foundation models," *arXiv preprint arXiv:2108.07258*, 2021.

[21] Y. Zhu, R. Kiros, R. Zemel, R. Salakhutdinov, R. Urtasun, A. Torralba, and S. Fidler, "Aligning books and movies: Towards story-like visual explanations by watching movies and reading books," in *The IEEE International Conference on Computer Vision (ICCV)*, December 2015.

[22] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[23] T. Shavrina and V. Malykh, "How not to lie with a benchmark: Rearranging nlp learderboards," 2021.

[24] F. Schlackl, N. Link, and H. Hoehle, "Antecedents and consequences of data breaches: A systematic review," *Information & Management*, p. 103638, 2022.

[25] J. Haislip, J.-H. Lim, and R. Pinsker, "The impact of executives' it expertise on reported data security breaches," *Information Systems Research*, vol. 32, no. 2, pp. 318–334, 2021.

[26] S. Goode, H. Hoehle, V. Venkatesh, and S. A. Brown, "User compensation as a data breach recovery action," *MIS Quarterly*, vol. 41, no. 3, pp. 703–A16, 2017.

[27] H. Cavusoglu, B. Mishra, and S. Raghunathan, "The effect of internet security breach announcements on market value: Capital market reactions for breached firms and internet security developers," *International Journal of Electronic Commerce*, vol. 9, no. 1, pp. 70–104, 2004.

[28] W. A. Cram, J. G. Proudfoot, and J. D'arcy, "Organizational information security policies: a review and research framework," *European Journal of Information Systems*, vol. 26, pp. 605–641, 2017.

[29] S. W. Schuetz, P. Benjamin Lowry, D. A. Pienta, and J. Bennett Thatcher, "The effectiveness of abstract versus concrete fear appeals in information security," *Journal of Management Information Systems*, vol. 37, no. 3, pp. 723–757, 2020.

[30] W. A. Cram, J. D'arcy, and J. G. Proudfoot, "Seeing the forest and the trees: a meta-analysis of the antecedents to information security policy compliance," *MIS quarterly*, vol. 43, no. 2, pp. 525–554, 2019.

[31] G. R. Milne and M. J. Culnan, "Strategies for reducing online privacy risks: Why consumers read (or don't read) online privacy notices," *Journal of interactive marketing*, vol. 18, no. 3, pp. 15–29, 2004.

[32] J. D'Arcy, T. Herath, and M. K. Shoss, "Understanding employee responses to stressful information security requirements: A coping perspective," *Journal of management information systems*, vol. 31, no. 2, pp. 285–318, 2014.

[33] S. Paul, P. Goyal, and S. Ghosh, "Lesicin: A heterogeneous graph-based approach for automatic legal statute identification from indian legal documents," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 36, no. 10, 2022, pp. 11 139–11 146.

[34] S. Gehman, S. Gururangan, M. Sap, Y. Choi, and N. A. Smith, "Realtoxicityprompts: Evaluating neural toxic degeneration in language models," *arXiv preprint arXiv:2009.11462*, 2020.

[35] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell *et al.*, "Language models are few-shot learners," *Advances in neural information processing systems*, vol. 33, pp. 1877–1901, 2020.

[36] N. Carlini, F. Tramer, E. Wallace, M. Jagielski, A. Herbert-Voss, K. Lee, A. Roberts, T. B. Brown, D. Song, U. Erlingsson *et al.*, "Extracting training data from large language models." in *USENIX Security Symposium*, vol. 6, 2021.