Creation and Analysis of an International Corpus of Privacy Laws

Geetika Gopi
Carnegie Mellon University

Harish Balaji
Carnegie Mellon University

Sonu Gupta
Penn State University

Nora O'Toole
Penn State University

Siddhant Arora
Carnegie Mellon University

Thomas Norton
Fordham Law School

Norman Sadeh
Fordham Law School

Shomir Wilson
Penn State University

Abstract

The landscape of privacy laws and regulations around the world is complex and ever-changing. National and supernational laws, agreements, decrees, and other government-issued rules form a patchwork that companies must follow to operate internationally. To examine the status and evolution of this patchwork, we introduce the Government Privacy Instructions Corpus, or GPI Corpus, of 1,040 privacy laws, regulations, and guidelines, covering 183 jurisdictions. This corpus enables a large-scale quantitative and qualitative examination of legal foci on privacy. We examine the temporal distribution of when GPIs were created and illustrate the dramatic increase in privacy legislation over the past 50 years, although a finer-grained examination reveals that the rate of increase varies depending on the personal data types that GPIs address. Our exploration also demonstrates that most privacy laws respectively address relatively few personal data types, showing that comprehensive privacy legislation remains rare. Additionally, topic modeling results show the prevalence of common themes in GPIs, such as finance, healthcare, and telecommunications. Finally, we release the corpus to the research community to promote further study [1].

1 Introduction

Privacy is a growing topic of attention for legislative and regulatory bodies around the world, and a growing number of documents produced by governments provide instructions for this topic. These government-issued instructions include legally binding documents such as laws and regulations, and also non-legally binding documents such as guidelines for following a law. Legal jurisdictions around the world have their own sets of government privacy instructions (GPIs, government privacy instructions), shaping the legal framework surrounding privacy within their particular jurisdictions

At the same time, text analysis techniques have made it possible to study legal texts on a large scale. Prior efforts have studied legal text about privacy in the form of privacy policies, yielding insights for legal scholars and language models for the creation of privacy-enhancing technologies [29, 43, 52]. Other efforts have applied NLP to legal text more generally [37, 45]. However, despite growing interest in privacy, privacy law and NLP researchers have lacked a large-scale collection of texts of privacy documents from around the world. This stems from the non-trivial nature of this process. Often there are several official and unofficial versions of a document on the web. Among the official ones, governments also publish instructions to "simply" the adaptation of these laws [2] [3] which makes it challenging to distinguish them and legally enforceable documents. The task is further exacerbated by the absence of official translations of these laws.

We address these challenges and present the Government Privacy Instructions Corpus (GPI Corpus). To the best of our knowledge, the GPI Corpus is the most comprehensive corpus of government privacy instructions to date, with natural language text in original languages and English. [4] The texts are paired with extensive metadata on the documents' electronic sources (i.e., URLs), relevant jurisdictions, dates of enactment, relation to international agreements, and other significant information. We coin the term government privacy instructions, or GPIs to characterize these documents, as the corpus encompasses laws, regulations, and government-issues guidelines and recommendations intended to instruct citizens, organizations, law enforcement, or lawmakers on actions to protect digital privacy. We include legally binding documents such as laws and government-produced non-legally binding documents such as guidelines in the corpus. Together, they provide a comprehensive view of the privacy instruction infor-

[1] https://usableprivacy.org/data

[2] https://www.oaic.gov.au/__data/assets/pdf_file/0012/8013/privacy safeguardcombinedchapters.pdf

[3] https://www.priv.gc.ca/en/privacytopics/technology/online-privacy-trackingcookies/trackingandads/gl_ba_1112/

[4] In this study we focus on GPIs originally in English or translated to English, to match the authors' expertise. However, we acknowledge the importance of multilingual analysis, which motivated our inclusion of original non-English documents in the corpus for future use by ourselves or others.

mation provided by governments. In order to trace the history of such documents and the ways in which they may inherit vocabulary, concepts, and precedents from one another, the included documents comprise ones that are binding or relevant today, along with ones that have been in effect in the past. We also present the first large-scale study of GPIs using natural language processing (NLP) tools. We examine temporal and topical trends in GPIs, showing a dramatic increase in attention to privacy over the past 50 years, a varied and nuanced distribution of mentions to personal information types, and a set of common themes that GPIs address.

We structure the rest of the paper as follows. In Related Work, we describe some prior efforts toward privacy text corpora and law corpora. In Corpus Creation, we describe the types of documents that comprise the GPI Corpus and the criteria, and how we gathered them from the web. In Distribution of GPIs, we show the differences in document availability and quantities of documents across geographic regions and across time. In addition, we use the metadata collected about the documents to make observations about the prevalence of GPIs over time and their availability in English. In Text Analysis, we study the distribution of mentions to personal data types across the corpus. We also apply LDA-based topic modeling techniques to extract the privacy topics discussed in the corpus. In the Discussion, we share the challenges of text analysis, and the limitations of this work. We conclude with Future Directions and Conclusions.

2 Related Work

We describe prior efforts toward language resource creation and NLP applications on four related domains of text: laws in general, legal documents, privacy policies, and privacy laws.

Law Corpora: Prior work has created international law corpora with varying foci. Elliott [18] curated a master list of 779 international human rights instruments from 1863 to 2003 to highlight significant violations of those rights. Adams et al. [12] created a dataset of 63 labor laws to generate the Centre for Business Research - Labor Regulation Index (CBR-LRI) dataset. Deakin and Sarkar used this data to estimate the impact of labor regulation on unemployment. The authors further expanded this dataset to include 117 countries [11]. Similar efforts have created datasets of non-English policies. González Ferrer and Mezger [19] developed ImPol, a database, to estimate immigration policies in three European countries (France, Italy, and Spain) from 1960 to 2008. With the recent progress in NLP, language models have been created to apply law corpora to practical problems. Researchers used statutes of the US Internal Revenue Code to extract a set of rules along with a collection of natural language questions that can be answered correctly only by consulting these rules [28]. The authors also developed a StAtutory Reasoning Assessment dataset (SARA) for question answering and statutory reasoning in tax Law entailment. Lame [33] proposed an NLP-based technique to extract concepts and relations from 57 French codes gathered from government websites that constitute 59,000 articles.

Legal Document Corpora: The analysis and interpretation of text dominates the field of law. Lawyers, judges, and regulators continuously compose legal documents such as memos, contracts, patents, and judicial decisions. Accordingly, there is a body of research about creating corpora of such legal documents. These corpora facilitate building Natural Language Understanding (NLU) technologies to assist legal practitioners. Malik et al. [34] introduce a corpus of Indian legal documents toward building an automated system for predicting the outcome of a legal trial as well as explaining the outcome. These automated systems can assist judges and help expedite the judicial process. Another similar study [32] annotates Indian legal documents for rhetorical roles, which has applications for both legal judgment prediction and legal summarization. Chalkidis et al. [17] create a benchmark dataset for various legal NLU tasks and evaluate different pretrained Language models on these tasks. There have also been similar efforts to develop legal documents corpora for nonEnglish languages, such as Mauri et al. [35], who created a corpus of Canadian legal documents with legally equivalent texts in English and French, respectively. These corpora have enabled the creation of automated methods to interpret these legal documents. Josi et al. [31] aims at automatic extracting text from signed PDF legal documents. Similarly, there has also been an interest in performing named entity recognition in legal domains [40]. Another work [36] aims at automating the extraction of information from legal judgments to assist lawyers on the case at hand. There has also been work in summarising legal text like court judgment documents to help legal professionals and ordinary citizens to get relevant information with little effort [30, 39]. Similar efforts [14, 43, 53, 54] have also been made to interpret legal documents in non-English languages.

Privacy Policy Corpora: Over the last decade, there has been significant growth in research about online privacy policies. The existence of data and high-quality annotations are essential for the application of both natural language processing and crowd-sourcing techniques to address the challenges posed by online privacy policies. This requirement has generated two threads in online privacy policy research: (i) annotation of privacy policy documents to facilitate future analysis and (ii) large-scale collection and analysis of privacy policies. The initial annotation attempts involved manual annotation of privacy policies by legal experts and crowd workers. Two such corpora are OPP-115 [51] and APP-350 [55]. OPP-115 consists of 115 web privacy policies (267K words) with 23K finegrained data practices annotations. Although these corpora are relatively small, their annotations enable several researchers use them to train machine learning models to extract salient details from privacy policies [46, 47]. In an attempt to create a larger corpus, Harkous et al. [27] collected 130K mobile applications' privacy policies from the Google Play Store. Authors used the corpus to train a privacy policy-centric language model and built a set of neural network-based classifiers for both high-level and fine-grained aspects of privacy practices. In a similar effort, Srinath et al. [48] collected 1.4M privacy policies and developed a privacy policy search engine, PrivaSeer, which enables text query-based search across the collection. In follow-up work, authors [49] trained a transformer-based language model using this corpus, resulting in the state of the art performance on classification and question answering tasks [13, 44].

Privacy Law Corpora: In 2011, Graham Greenleaf performed the first global survey of data privacy laws and identified 76 countries that meet minimum international data privacy standards of international data protection and privacy agreements [20]. After a decade, the seventh edition of this work [25] expanded the global table to 145 and 23 countries with Data Privacy Laws and bills, respectively. This corpus has been used to analyze the momentum toward global ubiquity of data privacy laws [24], the networks of data privacy authorities [23], and progress for international data privacy standards [5]. World Legal Information Institute (WorldLII) developed a privacy research library that consists of links to case laws, commentaries, legislation, and more that several Legal Information Institutes originally maintain (LIIs) [8]. DLA Piper, a global law firm, presents an overview of data protection laws for 89 jurisdictions [41]. Along with global corpora, there are studies of laws of specific regions. In [1], researchers analyzed the rising data protection systems in Africa concerning cultural differences across countries in addition to their socio-economic and political landscape. Authors compared 32 African data privacy laws at a fine-grained level against 30 features of data privacy law such as data quality, access, and collection [26]. Further, researchers [16] highlighted the similarities and differences between the South African Protection of Personal Information Act (PoPI) and the international data protection laws. Similarly, in [21], the author discussed and analyzed Asian data privacy laws in-depth. Our work closely aligns with the previous work by Greenleaf. We take a broader perspective of the data protection laws and broaden the inclusion criteria to extend our corpus by including more jurisdictions and documents (e.g., guidelines). In addition, all the above efforts present only qualitative analysis. In contrast, we employ both quantitative and qualitative methodologies. We also leverage NLP tools and machine learning algorithms to study this large-scale corpus. Lastly, unlike previous work that shared the list of the names of these documents, we share the original text of all the documents. We also consider the multi-lingual dimension and share both the non-English and English translations.

3 CORPUS CREATION

Corpus creation required a series of overarching tasks: searching by jurisdiction for document that ought to be included in the corpus, determining precise jurisdiction and document inclusion criteria, manually collecting GPIs for the selected jurisdictions from the internet, and categorizing these documents into three subdivisions. We summarize the entire pipeline of the corpus creation tasks in **Figure 1**.

3.1 Jurisdiction

Intending to achieve extensive coverage of nation-level jurisdictions worldwide, we curate a list of candidate jurisdictions prior to collecting their GPI documents. First, to build this list, we defer to the existing work by Greenleaf [22], and leading legal experts that provide such information with their online legal resources such as Data Guidance [2] and DLA Piper [4]. The jurisdictions mentioned in these sources serve as the starting point for our work, and for the duration of the corpus creation process, we often refer to them. For the sake of simplicity, we call them reasoning documents. Next, we instate a series of inclusion criteria to scope our list of jurisdictions.

The initial list of jurisdictions derived from Greenleaf's table is limited. We believe privacy researchers can benefit from a more comprehensive list with better coverage of documents from around the world at a national level. This requires the development of a set of rules to filter out the jurisdictions that are outside the scope of our work. These criteria facilitate our manual search for the jurisdictions across the web to expand our initial list, weighing each jurisdiction against these criteria to decide whether it ought to be included to make the corpus representative of a consistent group of jurisdictions that fulfill certain requirements.

To develop a criterion to represent all available documents from various nations, as well as all categories of non-nation locales represented in the Greenleaf [22], we decided to focus on country-level jurisdictions together with a few special categories. We include a jurisdiction if it satisfies one of the two requirements: (i) it is a country recognized as either a member or observer state of the United Nations by at least one other member state as of 2020, and (ii) a jurisdiction falls into the following special categories: (a) self-governing British Overseas Territories (Bermuda, Gibraltar, Cayman Islands), (b) crown dependencies (Guernsey, Jersy, Isle of Man), (c) Chinese Special Economic Regions (Macau and Hong Kong), (d) Qatar economic free zones (Qatar Financial Centre), (e) United Arab Emirates economic free zones (Abu Dhabi Global Market, Dubai International Financial

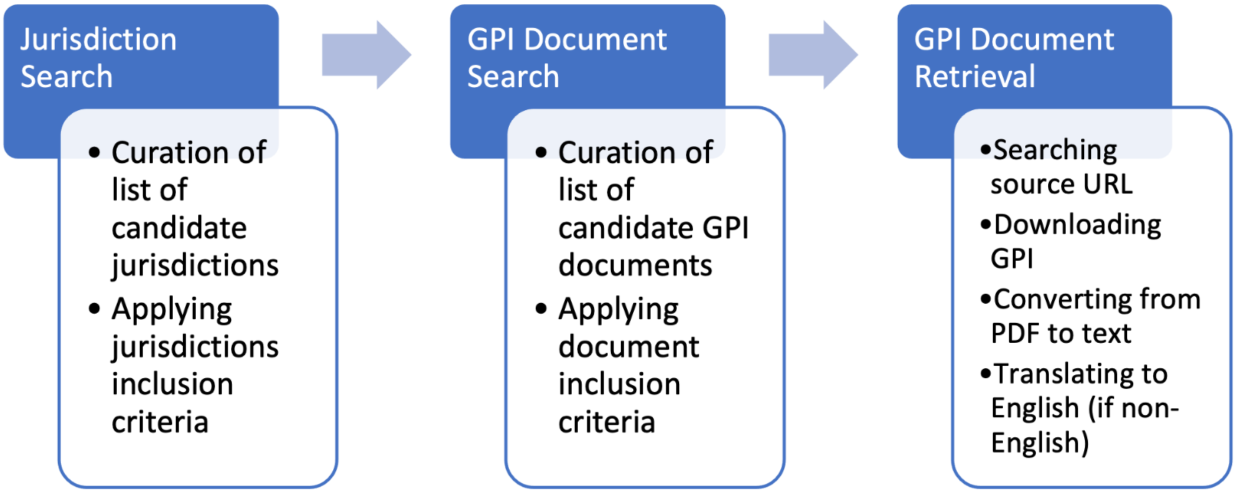
Jurisdiction Search
• Curation of list of candidate jurisdictions
• Applying jurisdictions inclusion criteria
GPI Document Search
• Curation of list of candidate GPI documents
• Applying document inclusion criteria
GPI Document Retrieval
•Searching source URL
•Downloading GPI
•Converting from PDF to text
•Translating to English (if non-English)

Figure 1: The end-to-end pipeline of the creation of GPI corpus.

Centre, Dubai Healthcare City), and (f) the states which are not recognized as UN members or observers (The Republic of China (Taiwan) and Kosovo). We also include one US state (California), with a rationale explained below.

A subset of the jurisdictions from the special jurisdictional categories is reverse-engineered from the list of jurisdictions originally mentioned in the Data Privacy Tables developed by Greenleaf [22, 25] . To ensure consistent presentation of all the jurisdictions from each of the described jurisdiction types, we add several individual jurisdictions that are not part of Greenleaf's table. We further expand our list of relevant documents with the help of information present in the documents that satisfy our inclusion criteria. With an exception, we include one US state, California, in our corpus, due to its significance and weight in defining privacy legislation that impacts the entire US economic system [42].

In summary, this process resulted in a list of 183 jurisdictions, with 161 at the country level (86% of 193 United Nations member states [7]). For the remaining 27 countries in United Nations member states, either no GPI exists or it was irretrievable on the internet.

3.2 Government Privacy Instruction Documents

To create an initial list of candidate documents, we refer to the list of 132 privacy laws collected by Greenleaf [22] and, by default, include documents present in it. Several other documents are included because of their inclusion in online resources compiled by legal experts for public viewing and use, pertaining to the applicable privacy legislation and guidelines within each of several jurisdictions. Iteratively, we extend this list if an existing document of this list points to other candidate documents. Upon discovery of additional documents, we apply our GPI document inclusion criteria to determine whether they should be included in the corpus.

Our goal is to curate an initial exhaustive list of all candidate GPI documents. However, we need to filter out the documents that fall beyond our scope. We develop two sets of rules based on document type and source to address this. All the documents to be included must satisfy at least one criteria from both sets of the rules.

Criteria based on document type — Each document must meet at least one of the following inclusion criteria:

1. The document is legally enforceable (or once-enforceable and now defunct, or assumed to be enforceable upon some future date of effect), which is promulgated in a complete state to the general public for the purposes of awareness of the law and enforcement if it is in force, which may include laws and regulations

2. The document contains rules, clarification, or similar resources directed towards lawmakers or law enforcement for the purposes of enforcing the aforementioned document.

3. The document contains a non-enforceable list of guidelines, which serve as official guidance directed towards the general public, or specific sectors of the public, for the purposes of advising them on how to comply with a document of another type.

A detailed description of excluded document types — These document types exclude case law, which establishes legal precedents through individual court decisions. Although such cases are valuable pieces of information and form precedents for decisions regarding compliance with laws related to privacy, they neither form an explicit legal directive or instruction nor a document explicitly instructing the reader about how to enforce or comply with such instructions. Additionally, due to the overwhelming scope and limited resources for acquiring case law notices or summaries globally, case law is categorically excluded from this work.

This corpus also excludes discussions of legal rationale unaccompanied by content that matches the aforementioned document types. Much like case law, discussions and arguments explaining the rationale behind a legal directive are a malleable resource that can be used to understand the application of the law. However, we exclude them because such documents also do not provide any direct instruction or guidance to the reader and, instead summarize lawmakers' theoretical decisions.

The final notable type of document excluded from this work is national constitutions, which provide established principles with significance both in their own right as legal documents and as a potent precedent for other laws developed in the country. We categorically exclude national constitutions because, in the overwhelming majority of cases, allusions to a right to privacy in a constitutional document were found to lack actionable details regarding expectations, instructions, or enforcement. Thus, although such mentions within national constitutions may act as a guiding principle in the development of subsequent legal documents regarding privacy, we find that these constitutional documents do not provide enough instruction to lawmakers, enforcers, or citizens regarding privacy to be a meaningful and effective part for our corpus.

Criteria based on source type — The documents must meet any one of the following inclusion criteria for document source:

1. The document contains more content than a notification containing some update regarding the legal status of another document. A decree that says only that a different law is now in effect, providing no further guidance or substance, is excluded.

2. The document is released by a government entity, such as (but not limited to) an executive order released by a president, a law passed by a congress or parliament, or a set of rules released by a government agency. Documents released by non-government entities, such as rules released by corporations and non-profit organizations for the internal governance of data privacy, guidelines released for the general public, and others, are excluded.

3. The document is released to the general public with the intent of circulating the document in its current, complete state for the purposes of understanding or enforcement of the document. Such circulation resources may include government websites and legal journals. This implies that the following types of documents are excluded.

   (a) Private documents are not meant for such release to the public.

   (b) Rules that describe internal procedures not directly relevant to the privacy laws and concepts in question, such as documents that merely describe which agencies or positions are charged with particular enforcement duties, are not included in this corpus. This is because these documents do not provide meaningful context into how the meaning of law itself is interpreted and enforced.

   (c) Activity reports of government agencies, meant primarily for internal review and as a resource regarding the state of enforcement. Because of their conceptual removal from the types of documents of interest to the researchers.

   (d) Strategy and action plans designed for internal use by enforcement agencies.

   (e) Enforcement decisions and records of fines. This is because they are notices aimed towards the specific audience of a given punished entity, without a desired audience of the general civic public.

4. While future versions of the document may be released with changes, the document is released within its given form with the understanding that this form is immutable and is to be understood as-is until further documents are released to update it. This implies that the following types of documents are excluded

   (a) Bills and similarly unfinished documents released in various drafts for the purposes of transient public forum discussion.

   (b) Forms, software tools, and other tools that require active constituent participation for effective use. As the form of these artifacts extends beyond the static, immutable document states that we wish to analyze here.

5. The documents are promulgated in their included region by or before December 31, 2020. We set that date significantly in the past to promote higher recall in the final years of the corpus, recognizing that documents from some jurisdictions are not immediately available online.

3.3 Document Collection

If a document is deemed fit to be included within the corpus, we look for the source of the downloadable version on the web.

We begin the search from our set of reasoning documents. We are able to find a few direct sources to downloadable documents mentioned in the reasoning documents. However, in most cases, they do not provide direct links to the source documents we sought to include; for several documents, only the associated legal document titles are mentioned without a link to the law or other legal document. This is especially true for documents that are not originally composed in the English language, in which case we seek to collect both an original language version and, if available, a human-translated English version of the document. Since few reasoning documents are linked to only one language version of a document or to no version, we leverage the mentioned legal document titles to search for the sources of other legal documents that are not present in the reasoning documents.

We collect all the documents within this corpus manually from the internet using the document inclusion criteria described above. The collection activities include locating, downloading, and uploading documents to the repository and recording the metadata. It took two researchers from our team approximately 60 labor hours combined to complete these tasks. Since this process of manual collection is conducted within the broad scope of any candidate documents we might find on the internet, rather than a closed list of automatically retrieved results, we take every caution to apply the inclusion criteria described above comprehensively to all the documents we find in the process of collection.

We download all the relevant documents from the web in PDF, to preserve visual formatting. This includes both the document in the original language and their English translation wherever available. Since we cannot directly extract the text content from PDF documents, we convert them into a text file format (.txt) using Apache Tika [50]. We attempt to use OCR [38] to convert scanned documents, although the software was not always successful. In instances when we were only able to collect a non-English version of a document, we translated its text file to English using online tools. We elaborate on the issues and challenges of the translation process in section 4.1.

3.4 Subdivision of the Corpus

We divide the GPI Corpus into three sets: **the primary set, the untranslatable set, and the irretrievable set**. This categorization is necessary to segregate the documents based on their availability for content analysis. The primary set comprises documents within the canonical body of the corpus. It includes documents that exist as an available English language text within the corpus, whether this text is the original document, a human translation, or a machine translation. Next, as the name suggests, the untranslatable set contains documents for which we present original, non-English-language text, and we are unable to translate into a usable English-language version. Lastly, the irretrievable set is a list of documents we sought to include in the corpus but are ultimately unable to retrieve from the internet in a usable state either in English or in some original non-English language. The failure modes for the untranslatable set and the irretrievable set are detailed in Section 4.1 and Section 4.3. In addition, if available, we retrieve the metadata (e.g., the year of enactment, if in effect or repealed) for all the documents present in these three sets and use it for the temporal analysis of the corpus

Set Name	No. of Docs	Percentage
Primary Set	1043	87.21%
Untranslatable Set	14	1.17%
Irretrievable Set	139	11.62%

Table 1: Summary of the subdivision of the documents.

4 DISTRIBUTION OF GPIs

The process described in the above section results in 1,040 documents for analysis. Based on the coverage, we classify the jurisdictions into three categories, (1) National, (2) International, and (3) State/Province. As shown in **Table 3**, the majority of the documents cover national level jurisdictions and contribute to 95.38% of the documents in the corpus with 183 unique jurisdictions, whereas 161 distinct countries make up 91.15% of the documents. Countries that participated in various international agreements also have their own unique sets of documents within the jurisdiction of their own country.

The collected documents are laws and regulations, rules, guidelines, and other government-released documents, communiques, notices, circulars, orders, decrees, and decisions. Each document is in its current state or the latest state of revision if any. The latest revision date is recorded for each document from the law category. We add the promulgation date as the last revision date if no revisions have occurred.

We show the number of enforceable privacy laws in our corpus per country as a map in **Figure 2**. We observe that Turkey has the most GPIs, followed by Japan, Uzbekistan, and France. We explore the reason for Turkey's exceptional number in Section 2.

4.1 Translations

The corpus comprises documents in 54 languages, with 37.11% documents in English, making it the most common language. In addition, 85 documents are written in both English and the native language of the region in a single document. Chad is the only exception where the document is written in two languages (French/Arabic), and neither of them is English. There are six languages that only appear in combination with the English language. For instance, all the 11 documents from Malta are written in Maltese/English

We attempt to create English translations for all the non-English documents in the corpus, to establish a uniform natural language for text analysis. Based on the translation, we divide the corpus into four classes: (1) Originally in English, (2) Official translation, (3) Unofficial translation, and as the name suggests, and (4) Non-Government Machine. If a document is in a language other than English, we seek an official English translation provided by the source of the official non-English document. Sometimes, official sources provide a translated version but call it an unofficial document for legal purposes.In the absence of the availability of such translations, we turn to international privacy expert sites, with exact sources noted in the corpus metadata. However, if the translation is still unavailable, we use translation tools like Google Translate [10]. Machine translated documents are referred to as 'non-government machine'. As we present in **Table 2**, 53.65% documents within the corpus do not contain a translation. It emphasize the difficulty in searching and accessing the non-English GPIs due to language barrier.

Out of 654 non-English documents, We utilize their English titles to locate the source of the English version of the document on the web. In the absence of non-English titles, we turn to Google translate. However, it fails to provide a usable translation for a few titles in the Russian language. Therefore, we turn to Yandex Translate [3]. Yandex is a Russian technology company that provides internetrelated products and services [9]. We also utilize Yandex Translate to scan the contents of a few non-English websites and look for the required document.

Translation Type	No. of Docs	(%) of Total
Non-Government Machine	558	53.65
Originally English	386	37.12
Official Translation	66	6.35
Unofficial Translation	19	1.83
Originally in other languages	11	1.06

Table 2: Distribution of the sources of English translations.

4.2 Temporal Distribution

We examine the distribution over time of the creation of GPIs, as the corpus contains documents dated as early as 1803 and as recently as December 2020. We illustrate the pace of GPIs enacted over this date range in **[Figure 3]**. It is a dual-vertical axis graph where the left and right vertical axes show the cumulative and total number of GPIs enacted over the years, respectively. We perform a chi-square test for the goodness of fit and detect a trend, starting in 1966, that the pace of GPIs grows exponentially with rate parameter (λ) equal to 0.054. We also find a sharper exponential growth in the 21st century with (λ) equal to 0.036. It should be noted that in a few documents first written in the late 80s and early 90s, the data privacy statements were included only after revisions. For example, the criminal code of Finland was enacted in 1889, but a data privacy section wasn't added to it until 2015.

We notice two discernible peaks in 2016 and 2018. In 2016, 86 GPIs were issued, out of which 32 were published in Turkey only. After the failed July 15, 2016 coup attempt [10] in Turkey, several emergency decrees were published [15]. We speculate that the coup attempt was the cause of the sudden increase in GPIs in Turkey. We also observe that the largest number of GPIs were issued in 2018, with a total of 131 GPIs in 67 distinct jurisdictions. We speculate that the enactment of GPDR in early 2018 may have caused the increase in the new documents as 16.79% of documents explicitly mention GDPR in their title. Additionally, GDPR may have encouraged the presence of documents to be in digital format and available over the web. We note a continuous increase every decade, with the most number of jurisdictions receiving their first GPI between 2010 and 2019. There are 62 such jurisdictions, including countries like Costa Rica, Barbados.

4.3 Excluded Documents

There were 153 documents that we attempted to include in the primary set but were ultimately unable to add due to several failure modes. This led to the categorization of documents into three sets as discussed in Section 3.4. We present this categorical distribution in **Table 1**. The primary set consists of 1,040 documents that are either originally in English or translated into English. For 14 documents that could not be fully translated into English and the percentage of English words within the document was less than 95%, we added them to the untranslatable set. For three documents, Google Translate could not interpret the entire document due to a document size constraint [6]. We split these documents into smaller chunks in such scenarios and then perform the translation. Further, Google Translate does not offer services for all languages in the corpus. One such language not offered by Google Translate is Dzongkha, a language used in one of Bhutan's GPIs in the corpus. Lastly, we mark a document irretrievable if we are unable to locate the original English or NonEnglish document on the web in a machine-readable format. This resulted in the removal of 139 documents from the corpus. Below are other failure modes that contributed to the irretrievable set:

1. Errors with Optical Character Recognition: Running a scanned document through Optical Character Recognition (OCR) [66] results in a machine-readable text data. We utilize OCR for documents originally present on the

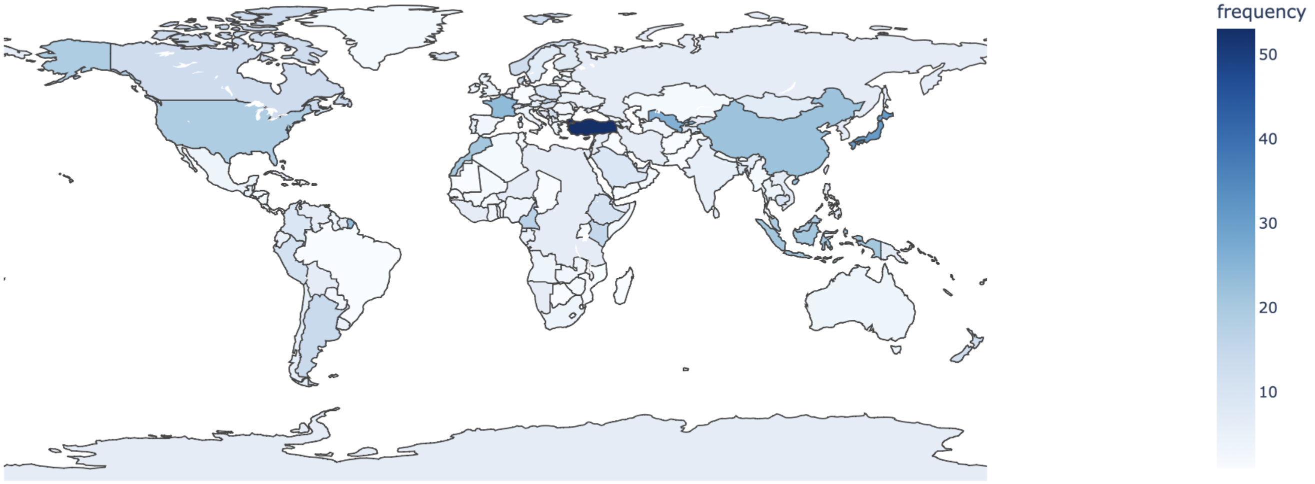

Figure 2: GPIs enacted over time.

Jurisdiction Type	Coverage Type	# Unique Jurisdictions	# Documents	Examples
Countries	N	161	948	Albania
British Overseas Territories	N	3	24	Cayman Islands
Crown Dependencies	N	3	20	Isle of Man
Special Administrative Regions	S/P	3	18	Macau
International Organizations	I	4	14	United Nations
Special Economic Zones	S/P	4	8	Qatar Financial Centre
Intergovernmental Organizations	I	4	5	US + 23 Countries
State	S/P	1	3	California(USA)

Table 3: Summary of corpus composition. In the Coverage Type column, N, I, and S/P represent National, International, and State/Province jurisdictions, respectively.

web in a scanned version. However, 58 documents could not be accurately or completely converted.

2. Document Not Found: This includes 50 documents mentioned in several reasoning documents or other GPIs, but we could not find them on the web.

3. Not Found: This includes 20 documents for which we are able to source a URL, but the available URL resulted in a 404 HTTP error

4. Suspicious URLs: If the web browser blocks access to a URL, we exclude it. We have two such links associated with the jurisdiction of Montenegro.

5. Cannot Access: Nine documents are not available on the web directly and require a paid subscription of a service like Guidance Notes [2].

5 TEXT ANALYSIS

We use text analysis methods to examine the composition of the GPI Corpus and trends in GPIs. In the rest of this section, we use corpus and primary set interchangeably. **[Table 4]** shows some summary statistics for the corpus. The documents vary widely in length, from **[[40 words to 509,094]]**.

To explore mentions to technologies in GPIs, we hand-curate a list of 32 keywords used to describe technologies relevant to consumer privacy. We started with a small list of words referring to common technologies (e.g., computer, website) and after multiple iterations of discussion with privacy and legal experts, we expanded it to 32 keywords. For text analysis, we convert all the keywords into lower case and their singular forms (e.g., “Emails” is changed to email). We perform the same preprocessing steps on the corpus documents and then calculate the statistics. We consider the text of primary set documents and find that out of 1,043 documents,

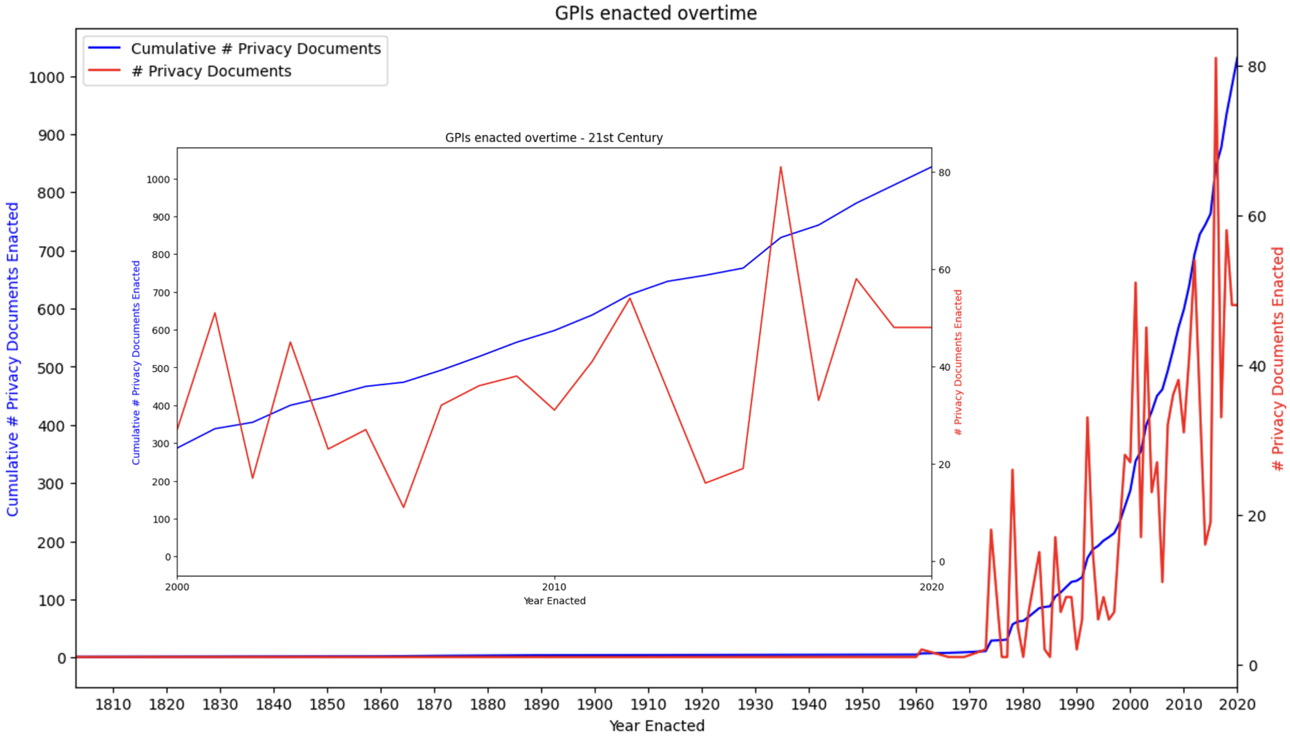

Figure 3: GPIs enacted over time.

73 documents do not contain any of the keywords. We present top ten most frequent keywords in **[Table 5]**. We observe that email and phone, two of the oldest methods of electronic communication in our list, lead the frequency ranking with a large gap before the fourth (computer).

5.1 Personally Identifiable Information

Personal identifiable information (PII) includes any information associated with an identified or identifiable living person in particular that can be connected to an identifier such as a name, national identification number, email address, and more [2]. GPIs often explicitly include a descriptive definition of PII at the beginning of the document. For instance, "...personal information refers to personal information: (1) About an individual's race, ethnic origin, marital status, age, color, and religious, philosophical..." - Data Privacy Act (2012), Philippines [6].

We examine the distribution of mentions of PII types [6] in GPIs and their trends over time to identify differences in attention and temporal trends. To do this, we create a list of ***154*** PII keywords with the help of following official sources:

- The U.S. National Archives and Records Administration [18]
- The U.S. National Institute of Standards and Technology (NIST) [14]
- The U.S. Federal Trade Commission (FTC) [9]
- The European Commission [8]

We expand the list by including several country-specific alternates for each PII keyword. For example, national ID schemes are known by many names, including Social Security Number in the US[7] , Documento Nacional de Identidad in Argentina [8] , and Aadhaar in India [9] . We also include variations for non-country specific terms. For example, a name could be mentioned as a middle name, first name, last name, mother's maiden name, surname, and more such variations

We place our PII keywords into 15 unique categories, as shown in **[Table 6]**, and create a miscellaneous category for four keywords that do not fit any of these themes. The complete list will be provided upon acceptance. To perform analysis, we convert all the keywords to lower case and their singular forms (e.g., "Ages" is changed to age). Similarly, we consider abbreviations of the keywords as well. For instance, we look for "mobile number" as well as "mobile no." and count them as the same entity. We perform the same preprocessing steps on the corpus documents and then calculate the statistics

[Figure 4] shows that for every category besides biographical, the percentage of documents that contain keywords from the category is less than half.We also observe that the tracking IDs, a category that contains relatively technical terms compared to other categories, do not often appear in the GPIs. This suggests that laws refrain from committing to regulating specific representation of the information. Additionally, **[Figure 5]** represents the number of PII types present per document in the corpus, shows that privacy legislation that addresses a large number of different PII types is rare. About 60.79% GPIs represent three or fewer PII types. There are only 0.19% GPIs that are comprehensive enough to cover the 14 PII types listed in the **[Table 6]**. According to this measure, the two most comprehensive privacy laws are California Consumer Privacy Act (CCPA) and California Privacy Rights Act (CPRA).

We further compute the pair-wise correlations between the occurrence of PII types in each document to investigate the linear relationship between PII types. We show the results in **[Figure 6]**. We note that the correlations between all the PII types are always positive but differ for all the PII pairs. Although the correlations are positive, they are below 0.6, showing that no one pairing is dominant. The highest correlation is between Race/Ethnicity and Beliefs. GPIs that fall under this scenario include documents from 95 distinct countries and cover 16.29% of documents in the corpus. We also observed a high correlation between Genetic and Biometric PII types, which share a biological focus.

In **[Figure 7]** we show that all the PII types exhibit increase (i.e., the second derivative is positive) in frequency over time, but the rate of increase varies across the categories. For instance, we observe that for the "Biographical/Demographic" and "Contact Number" increase has been rapid; however, for "Tracking IDs" and "Photo" the rate of increase is more sedate.

5.2 Topic Modeling

To explore the range of the topics covered in the GPIs, we turn to algorithmic methods. In machine learning, topic modeling is an unsupervised learning technique that identifies major themes or topics in a collection of documents. We leverage Latent Dirichlet Allocation (LDA), a probabilistic model, to extract latent semantic topics in the GPIs [26]. LDA model assumes that each document consists of several topics and that each topic is a distribution of words. Although every document in the GPI Corpus concerns privacy, there are several dimensions to this topic. Therefore, we partition GPIs into paragraphs to explore at a finer-grained level themes they contain.

Our GPIs are stored in text files, but there are no discernible patterns to extract the paragraph structure precisely for a text document. Therefore, we use two newline characters () as a proxy indicator of a new paragraph unit to extract the paragraphs from a document. It results in paragraphs with a vast range (1- 27,890 words) of length. To balance this range, we take a step further to divide the larger paragraphs into smaller paragraph units. We take a threshold of ten sentences and divide all the paragraphs we extracted in the previous step into the chunks of at most ten sentences. To filter out extremely small paragraphs, we remove all the paragraphs with less than nine words. With this technique, we are able to reduce the range to 9-1,133 words per paragraph. Each of these chunks forms a single input document unit for the LDA model.

We apply the following steps to preprocess the input text segments:

1. We tokenize all the segments into uni-grams,
2. We curate a custom list of stopwords. Stopwords are words that carry very little information. For our context, words like "article", "chapter", "number" provide insufficient information and, we include typical stop words such as "the", "is" , "an" from gensim [15] We then remove all the stopwords from the text segments,
3. We lemmatize all the tokens using WordNetLemmatizer [16],
4. We remove all the tokens with less than three characters, and
5. We filter out the tokens that occur less than 15 times and the ones present in more than 50

We generate a dictionary with the remaining tokens. We next compute the vector representation of each token using TF-IDF [63] and give it to the LDA model. One hyperparameter of the LDA is the number of topics (k) to be considered. We experiment with six values for k, equal to 5, 6, 10, 12, 15, and 20, and by manual analysis, we find that the cohesiveness of the resulting clusters decreases with an increase in the k. We also experiment with a combination of uni-gram and bi-gram inputs and find that uni-gram results in a higher coherence score.

We manually interpret each output topic cluster by inspecting each topic's top ten relevant terms and the relevant documents. We get the best results for k equals six and show our results in Table 7. We observe discernible meanings in eight clusters. Out of these six clusters, four clusters show notable strong connections to the significant privacy concerns. These clusters cover Telecommunications, Bank and Finance, Healthcare, and Consumer Payments, which are intuitively common industries for privacy concern. We also observe subtle similarities between the Prosecution Process and Penalty clusters as the first describes the various aspects of prosecution, latter talks about the outcome of the prosecution. It is worth noting that these two topics suggest criminal laws. Given the broad inclusion criteria, we also include the criminal laws published by various jurisdictions that have sections devoted to privacy and data protection concerns.

For the last two clusters, the top relevant terms point to a combination of topics instead of a single topic. The topic nine contains terms like "processor", "subject", "person", "right", and "regulation" which seems to be about People and Regulations. Similarly, with terms like "fine", "federal", "national", and "sanction", topic ten appears to be about National Level Authorities and Sanctions.

6 DISCUSSION

We discuss the issues in text analysis and legal enforceability of the GPI documents, and conclude with a discussion on the limitations of this work.

Legal enforceability: In the corpus, we mark each English translated document with whether the translation is completed by a human government translator, a 3rd-party government translator or by our own machine translation. However, levels of legal enforceability of each type of document vary widely among countries and individual instances. For example, there are English translations of laws released by both third-party groups and government resources that proclaim that they are for informational use only and that the law is only legally enforceable in the original non-English language. In contrast, some jurisdictions appear to provide their laws in multiple languages but fail to specify which version of the document is legally enforceable.

6.1 Balancing column at last page

For balancing the both column length at last page use :

```
\vadjust{\vfill\pagebreak}
```

at appropriate place in your TEX file or in bibliography file.

References

[1] African Data Privacy Laws | SpringerLink.

[2] Dataguidance. https://www.dataguidance.com. Accessed: May 10, 2023.

[3] Dictionary and online translation between English and over 90 other languages - Yandex Translate.

[4] DLA Piper. https://www.dlapiper.com/en-us/. Accessed: May 10, 2023.

[5] Global Data Privacy Laws 2021: Uncertain Paths for International Standards by Graham Greenleaf :: SSRN.

[6] Translate documents & websites - Computer - Google Translate Help.

[7] United Nations. https://www.un.org/en/about-us/member-states. Accessed: May 10, 2023.

[8] WorldLII - Help: 404 File not found.

[9] Yandex.

[10] 2016 Turkish coup d'état attempt, May 2023. Page Version ID: 1154360104.

[11] Adams, Z., Bishop, L., and Deakin, S. Cbr labour regulation index (dataset of 117 countries). *Cambridge: Centre for Business Research* (2016).

[12] Adams, Z., Bishop, L., Deakin, S., Fenwick, C., Martinsson, S., and Rusconi, G. Labour regulation over time: new leximetric evidence. In *4th Conference of the Regulating for Decent Work Network, Developing and Implementing Policies for a Better Future for Work. ILO, Geneva* (2015).

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

[14] Avram, A.-M., Pais, V., and Tufis, D. Pyeurovoc: A tool for multilingual legal document classification with eurovoc descriptors. *arXiv preprint arXiv:2108.01139* (2021).

[15] Bazy Malaurie, C., Cleveland, S., Kiener, R., Suchocka, H., Tuori, K., and Velaers, J. Opinion on emergency decree laws nos 667-676 adopted following the failed coup of 15 july 2016 in turkey. Adopted by the Venice Commission at its 109th Plenary Session (9-10 December 2016), 2016.

[16] Botha, J., Grobler, M., Hahn, J., and Eloff, M. A high-level comparison between the south african protection of personal information act and international data protection laws. In *ICMLG2017 5th International Conference on Management Leadership and Governance* (2017), p. 57.

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

[18] Elliott, M. A. The institutional expansion of human rights, 1863–2003: A comprehensive dataset of international instruments. *Journal of Peace Research 48*, 4 (2011), 537–546.

[19] González Ferrer, A., and Mezger, C. The impol database: A new tool to measure immigration policies in france, italy and spain since the 1960s.

[20] Greenleaf, G. Global data privacy laws: Forty years of acceleration. *Privacy Laws and Business International Report*, 112 (2011), 11–17.

[21] Greenleaf, G. *Asian data privacy laws: trade & human rights perspectives*. OUP Oxford, 2014.

[22] Greenleaf, G. *Global data privacy laws 2019: 132 national laws & many bills*. The Privacy Laws & Business International Report, 2019.

[23] Greenleaf, G. Global data privacy 2021: Dpas joining networks are the rule.

[24] Greenleaf, G. Global data privacy laws 2021: Despite covid delays, 145 laws show gdpr dominance.

[25] Greenleaf, G. Global tables of data privacy laws and bills (january 2021).

[26] Greenleaf, G., and Cottier, B. Comparing African Data Privacy Laws: International, African and Regional Commitments, Apr. 2020.

[27] Harkous, H., Fawaz, K., Lebret, R., Schaub, F., Shin, K. G., and Aberer, K. Polisis: Automated analysis and presentation of privacy policies using deep learning. In *27th {USENIX} security symposium ({USENIX} security 18)* (2018), pp. 531–548.

[28] Holzenberger, N., Blair-Stanek, A., and Van Durme, B. A dataset for statutory reasoning in tax law entailment and question answering. *arXiv preprint arXiv:2005.05257* (2020).

[29] Hosseini, M. B., Breaux, T. D., Slavin, R., Niu, J., and Wang, X. Analyzing privacy policies through syntax-driven semantic analysis of information types. *Information and Software Technology 138* (2021), 106608.

[30] Jain, D., Borah, M. D., and Biswas, A. Summarization of legal documents: Where are we now and the way forward. *Computer Science Review 40* (2021), 100388.

[31] Josi, F., Wartena, C., and Heid, U. Preparing legal documents for nlp analysis: Improving the classification of text elements by using page features. In *Computer Science & Information Technology (CS & IT)* (2022), AIRCC Publishing Corporation, pp. 17–29.

[32] Kalamkar, P., Tiwari, A., Agarwal, A., Karn, S., Gupta, S., Raghavan, V., and Modi, A. Corpus for automatic structuring of legal documents. *arXiv preprint arXiv:2201.13125* (2022).

[33] Lame, G. Using nlp techniques to identify legal ontology components: concepts and relations. *Law and the Semantic Web: Legal Ontologies, Methodologies, Legal Information Retrieval, and Applications* (2005), 169–184.

[34] Malik, V., Sanjay, R., Nigam, S. K., Ghosh, K., Guha, S. K., Bhattacharya, A., and Modi, A. Ildc for cjpe: Indian legal documents corpus for court judgment prediction and explanation. *arXiv preprint arXiv:2105.13562* (2021).

[35] Mauri, F. S., Gijón, P. S., and González, A. O. Cadlaws–an english–french parallel corpus of legally equivalent documents. *Mutatis Mutandis: Revista Latinoamericana de Traducción 14*, 2 (2021), 494–508.

[36] Mistica, M., Zhang, G. Z., Chia, H., Shrestha, K. M., Gupta, R. K., Khandelwal, S., Paterson, J., Baldwin, T., and Beck, D. Information extraction from legal documents: A study in the context of common law court judgements. In *Proceedings of the The 18th Annual Workshop of the Australasian Language Technology Association* (2020), pp. 98–103.

[37] Moreno-Schneider, J., Rehm, G., Montiel-Ponsoda, E., Rodriguez-Doncel, V., Revenko, A., Karampatakis, S., Khvalchik, M., Sageder, C., Gracia, J., and Maganza, F. Orchestrating nlp services for the legal domain. *arXiv preprint arXiv:2003.12900* (2020).

[38] Mori, S., Nishida, H., and Yamada, H. *Optical character recognition*. John Wiley & Sons, Inc., 1999.

[39] Nguyen, D.-H., Nguyen, B.-S., Nghiem, N. V. D., Le, D. T., Khatun, M. A., Nguyen, M.-T., and Le, H. Robust deep reinforcement learning for extractive legal summarization. In *Neural Information Processing: 28th International Conference, ICONIP 2021, Sanur, Bali, Indonesia, December 8–12, 2021, Proceedings, Part VI 28* (2021), Springer, pp. 597–604.

[40] Păiș, V., Mitrofan, M., Gasan, C. L., Coneschi, V., and Ianov, A. Named entity recognition in the romanian legal domain. In *Proceedings of the Natural Legal Language Processing Workshop 2021* (2021), pp. 9–18.

[41] Piper, D. Data protection laws of the world: full handbook. *DLA Piper 1* (2017), 1–50.

[42] Pratum. The national impact of ccpa. https://pratum.com/blog/425-the-national-impact-of-ccpa, 2020. Accessed: May 10, 2023.

[43] Ravichander, A., Black, A. W., Norton, T., Wilson, S., and Sadeh, N. Breaking down walls of text: How can nlp benefit consumer privacy? In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing* (2021), vol. 1.

[44] Ravichander, A., Black, A. W., Wilson, S., Norton, T., and Sadeh, N. Question answering for privacy policies: Combining computational and legal perspectives. *arXiv preprint arXiv:1911.00841* (2019).

[45] Robaldo, L., Villata, S., Wyner, A., and Grabmair, M. Introduction for artificial intelligence and law: special issue "natural language processing for legal texts", 2019.

[46] Sathyendra, K. M., Wilson, S., Schaub, F., Zimmeck, S., and Sadeh, N. Identifying the provision of choices in privacy policy text. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing* (2017), pp. 2774–2779.

[47] Shvartzshanider, Y., Balashankar, A., Wies, T., and Subramanian, L. Recipe: Applying open domain question answering to privacy policies. In *Proceedings of the Workshop on Machine Reading for Question Answering* (2018), pp. 71–77.

[48] Srinath, M., Sundareswara, S. N., Giles, C. L., and Wilson, S. Privaseer: A privacy policy search engine. In *Web Engineering: 21st International Conference, ICWE 2021, Biarritz, France, May 18–21, 2021, Proceedings* (2021), Springer, pp. 286–301.

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

[50] Team, T. A. T. D. Apache tika. https://tika.apache.org, 2021. Accessed: May 11, 2023.

[51] Wilson, S., Schaub, F., Dara, A. A., Liu, F., Cherivirala, S., Leon, P. G., Andersen, M. S., Zimmeck, S., Sathyendra, K. M., Russell, N. C., 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.

[52] Wilson, S., Schaub, F., Liu, F., Sathyendra, K. M., Smullen, D., Zimmeck, S., Ramanath, R., Story, P., Liu, F., Sadeh, N., et al. Analyzing privacy policies at scale: From crowdsourcing to automated annotations. *ACM Transactions on the Web (TWEB) 13*, 1 (2018), 1–29.

[53] Xiao, C., Hu, X., Liu, Z., Tu, C., and Sun, M. Lawformer: A pre-trained language model for chinese legal long documents. *AI Open 2* (2021), 79–84.

[54] Yamada, H., Teufel, S., and Tokunaga, T. Building a corpus of legal argumentation in japanese judgement documents: towards structure-based summarisation. *Artificial Intelligence and Law 27* (2019), 141–170.

[55] Zimmeck, S., Story, P., Smullen, D., Ravichander, A., Wang, Z., Reidenberg, J. R., Russell, N. C., and Sadeh, N. Maps: Scaling privacy compliance analysis to a million apps. *Proc. Priv. Enhancing Tech. 2019* (2019), 66.