Towards Automated Regulation Analysis for Effective Privacy Compliance

Sunil Manandhar
IBM T.J. Watson Research Center
sunil@ibm.com

Kapil Singh
IBM T.J. Watson Research Center
kapil@us.ibm.com

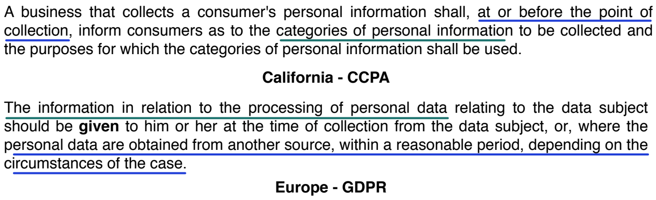
Adwait Nadkarni
William & Mary
nadkarni@cs.wm.edu

Abstract—**Privacy regulations are being introduced and amended around the globe to effectively regulate the processing of consumer data. These regulations are often analyzed to fulfill compliance mandates and to aid the design of practical systems that improve consumer privacy. However, at present, this is done manually, making the task error-prone, while also incurring significant time, effort, and cost for companies. This paper describes the design and implementation of ARC, a framework that transforms unstructured and complex regulatory text into a structured representation, the ARC tuple(s), which can be queried to assist in the analysis and understanding of regulations. We demonstrate ARC's effectiveness in extracting three forms of tuples with a high F-1 score (avg. 82.1% across all three) using four major privacy regulations: CCPA, GDPR, VCDPA, and PIPEDA. We then build ARCBert that identifies semantically similar phrases across regulations, enabling compliance analysts to identify common requirements. We run ARC on 16 additional privacy regulations and identify 1,556 ARC tuples and clusters of semantically similar phrases. Finally, we extend ARC to evaluate the compliance of privacy policies by comparing it against the disclosure requirements in the four regulations. Our empirical evaluation with the privacy policies of S&P 500 companies finds 476 missing disclosures, which when manually validated, result in 71.05% true positives, as well as the discovery of 288 additional missing disclosures from the partial matches identified by ARC.**

I. Introduction

Privacy and data protection regulations continue to be passed and amended around the globe, with 71% countries worldwide having their own privacy legislations at present [69]. Moreover, in the last few years alone, the United States has passed 5 new state-level data privacy legislations, and introduced 20 new privacy bills [37]. For these policy initiatives to be useful in advancing consumer privacy, the proliferation of complex regulations must be accompanied by systematic techniques and methods that enable researchers and practitioners to reason about them.

Particularly, there are two key stakeholders who need to analyze regulations: (1) *businesses/organizations*, who struggle to understand a diverse array of privacy regulations, employing legal experts who *manually ensure* that systems/processes continuously comply with the regulatory requirements [38], and (2) *security/privacy researchers*, who systematically evaluate the privacy posture of organizations against regulatory text, generally by *manually extracting privacy-requirements* from regulations to inform the design of privacy analyses [8], [33], [18], [17], [72], or evaluate their impact [45], [58], [7].


A business that collects a consumer's personal information shall, at or before the point of collection, inform consumers as to the categories of personal information to be collected and the purposes for which the categories of personal information shall be used.

California - CCPA

The information in relation to the processing of personal data relating to the data subject should be **given** to him or her at the time of collection from the data subject, or, where the personal data are obtained from another source, within a reasonable period, depending on the circumstances of the case.

Europe - GDPR

Fig. 1: A motivating example with two regulatory statements

For both researchers and practitioners, analyzing regulations *correctly* and *at scale* is extremely challenging, given the size, complexity, and sheer number of privacy regulations they must analyze. Numerous reports have shown that companies struggle to meet the requirements and are in danger of falling further behind [66], [31], [24], while at the same time spending billions to continuously comply with new/changed regulations (*e.g.*, GDPR is estimated to cost Fortune 500 companies $9 billion, with 40% of the costs spent on legal advice on regulations [52]). Similarly, most recent research that considers privacy regulations has been limited in scope and scale, either constrained to only a specific requirement (*e.g.*, evaluating "Do Not Sell" links based on the CCPA requirement [76], [49]), or the broad/imprecise impact of regulations on disclosure practices (e.g., analyzing changes in privacy policies [58], [48] or consent choices [70] after GDPR). These problems compound when we consider the frequent changes in regulations as new requirements are added, or existing ones are made more precise, requiring businesses to re-adapt their processes to comply, and researchers to re-examine the implications of their privacy analyses. A key hindrance in the practical analysis of regulations is the almost complete *lack of automation*: stakeholders (*i.e.*, practitioners/businesses and researchers) generally use *manual methods to extract pertinent information from complex privacy regulations, an approach that is labor-intensive, prone to manual error, difficult to scale, and exorbitantly expensive [51].*

We illustrate drawbacks of manually analyzing complex regulatory text with the example in Figure 1, which shows two statements from GDPR and CCPA, respectively. If we read the statements carefully, we see that both the statements effectively discuss the same idea: *that the business should*

Network and Distributed System Security (NDSS) Symposium 2024
26 February - 1 March 2024, San Diego, CA, USA
ISBN 1-891562-93-2
https://dx.doi.org/10.14722/ndss.2024.24650
www.ndss-symposium.org

inform consumers about what it collects. However, manually reasoning about regulations containing thousands of such statements is infeasible, for two key reasons. First, understanding and reasoning about these statements is a challenge in itself, as different regulations use regulation/jurisdiction-specific legal jargon, e.g., *'consumers'* in CCPA are *'data subjects'* for GDPR. Hence, the analyst must understand the vocabulary specific to each regulation to even begin to analyze it. Second, comparing two statements can be challenging because the effective similarity across sentences is not readily apparent. This is because regulation statements often include multiple contextual information that individually contribute to the overall meaning. For instance, both statements in Figure 1 discuss temporal condition (e.g., *"at the time of collection"*), which is an important context that helps with a precise understanding of the requirement. Hence, the analyst needs to identify and keep track of such contexts from multiple regulations for a proper comparison, which can be extremely labor-intensive. To summarize, the task of reasoning about thousands of such statements in each regulation, while also considering regulation-specific legal jargon, and contextual information that is critical for correctly extracting the semantics of the requirements, is infeasible to do manually. Thus, there is a fundamental gap in the area of privacy analysis and compliance: *the absence of a systematic and automated methods for understanding privacy regulations*.

The key argument in this paper is that there is a significant room for automation in tasks that form privacy regulation analysis, which can help both researchers and practitioners analyze regulations with correctness and scale, and help businesses adapt their processes to comply with evolving regulations. While the end goal is to reach full automation in this direction, this work builds the initial foundation that reduces the fully manual approach to address compliance requirements to a semi-automated one, only requiring *manual effort of value*. That is, we seek to build a semi-automated framework that automates aspects of regulation analysis that can be automated without significant legal expertise, thereby reducing manual effort and helping businesses, security and privacy researchers, and practitioners focus on their key end-goals of privacy analysis and/or compliance.

This paper proposes a framework that enables **Automated Representation and querying for privacy regulation Compliance (ARC)**. ARC transforms unstructured, complex, and contextually-rich regulatory text into a structured form that retains the context. ARC's design leverages Natural Language Processing (NLP) to: (1) identify key phrases that define the semantic meaning of a regulatory statement, (2) separate the core requirement in a statement from a plethora of complex clauses, and (3) express this digested form of regulatory text into a form of novel tuple-representation (known as the *ARC tuple*), which can be queried to accomplish several tasks that involve analyzing multiple privacy regulations. ARC's queries allow an analyst to easily compare several regulations, or use the regulations to evaluate privacy policies, thereby streamlining the evaluation of privacy compliance with various jurisdictions. We implement ARC and perform both *intrinsic* and *extrinsic* evaluation tasks, leading to *13 key results* ($\mathcal{R}_1 \rightarrow \mathcal{R}_{13}$) illustrating the effectiveness and usefulness of ARC in enabling privacy regulation and compliance analysis. The contributions of this paper are as follows:

- **The ARC framework:** We design and implement a novel framework, ARC, which creates structured representations of regulatory requirements, i.e., *ARC tuples*, from unstructured regulation text. We define three tuple representations: (i) the *data flow* tuple that extracts statements discussing the flow of information, (ii) the *definition* tuple that extracts the definitions of key terms, and (iii) the *rights* tuple that extracts information about the rights afforded to entities. These primitives lay the foundation for an automated extraction, representation, and analysis of text in privacy regulations.
- **Evaluation of Tuple Extraction (*intrinsic*):** We evaluate ARC's ability to consistently extract correct tuples from regulations with four major privacy regulations: CCPA, GDPR, Canada's Personal Information Protection and Electronic Documents Act (PIPEDA) [55], and Virginia's Consumer Data Protection Act (VCDPA) [71]. Our evaluation demonstrates an average F1-score of 83.4% for data flow tuples, 82% for definition tuples, and 81% for rights tuples.
- **Multi-regulation comparison using the *ARCBert* model (*extrinsic*):** ARC enables a novel approach for multi-regulation similarity analysis, *i.e.*, the ARCBert model, which identifies similar phrases across privacy regulations, and helps analysts identify similar statements as the previously discussed example in Figure 1. Our evaluation with the 4 regulations demonstrates how ARCBert outperforms baseline approaches by achieving a balance between two contrasting baselines: (a) keyword searches that identify few similar phrases (and miss many), and (b) an off-the-shelf model that identifies a large number of loosely similar phrases (that are significantly different). We further demonstrate how the similarity results from ARCBert can be used to understand phrases and statements across regulations.
- **Extracting ARC tuples and clusters of similar phrases from global privacy regulations (*intrinsic and extrinsic*):** We run ARC on 16 additional, diverse, global privacy regulations, and successfully extract 844 Data Flow Tuples, 536 Definition Tuples, and 176 Right Tuples. Further, we identify several clusters containing semantically similar phrases across 20 regulations using ARCBert. We leverage this evaluation, particularly the extraction of clusters from semantic role arguments, and discuss how it can be used to aid privacy regulation analysis.
- **Large-scale privacy compliance evaluation (*extrinsic*):** We build a module to analyze privacy policies for compliance using regulatory requirements extracted by ARC. We evaluate the privacy policies of S&P 500 companies with the 4 major privacy regulations, leading to the identification of 476 missing statements (*i.e.*, violations of disclosure requirements), including 111 CCPA violations by 38 companies, 173 VCDPA violations by 35 companies, and 192 GDPR violations by 49 companies. A manual validation of the results demonstrates that ARC was 90.1% correct in identifying *full matches* (*i.e.*, where the regulatory requirement from ARC's tuple is represented in the target privacy policy), which indicates a low false negative rate for the compliance analysis. Further, ARC was 71.05% correct in identifying *missing* statements, which indicates a reasonable false positive rate given the scale of the analysis (across regulations and policies) and the amount of manual effort reduced (i.e., relative to a manual comparison between every regulation and a policy). Given that S&P companies with presumably well-defined processes for regulatory compli-

ance failed to identify such violations/missing statements, our results highlight ARC's utility in providing automation for privacy compliance.

To enable future research, we have released the data and artifact [62]. Further, we have disclosed the privacy policy violations to the respective organizations. Listing 1 in the online Appendix [53] presents how we informed the organization.

II. Motivation

The goal of this paper is to automate key tasks related to the analysis of privacy regulations, such that it can help both researchers in designing or evaluating practical systems, and companies in complying with applicable privacy requirements.

Motivating Example 1 – Privacy Analysis for Researchers: Consider Alice, a researcher who is building a practical and effective system that can automatically analyze and reason about privacy practices of companies (e.g., analyzing privacy policies and/or code to identify issues [7], [8]). Alice prioritizes on important privacy issues, i.e., issues that if not fixed, can directly impact consumers. Since privacy regulations discuss mandatory requirements that businesses *must* comply with, Alice studies privacy regulation so that the goals of her system align with the regulatory requirements. After building the system, Alice also plans to report her results and findings to the companies, so that companies can take reasonable measures to fix the issues identified by Alice's system. To provide this context, Alice keeps track of relevant requirements from the regulation that the company may violate. However, with the introduction of newer jurisdiction-specific regulations and changes to the existing ones, Alice may need to: (1) reassess what the changes mean for her system, and (2) discuss her findings in the context of a given regulation. Alice spends an inordinate amount of time and effort in fully understanding as well as keeping up with current state of privacy regulations.

Motivating Example 2 – Privacy Compliance for companies: Consider Bob, a privacy compliance expert in a company that offers web services and operates in multiple geographical locations. Bob's task involves making sure that the services offered by the company are in compliance with the privacy regulations. For this, Bob needs to keep up with two moving pieces: **(1)** updates to company's processes as new services are deployed, and, **(2)** updates to privacy regulations in locations where services are provided. Hence, Bob converts the regulation requirements into compliance checklist, while also keeping track of company's processes. For example, if a new service starts collecting sensitive data from the user, Bob examines the manually curated checklist to make sure that it does not violate any requirements. However, every time there is change in the regulation, or when a new regulation becomes effective, Bob needs to repeat the process of examining the regulation and creating the manually-curated compliance checklist. While performing such an update, Bob needs to carefully interpret the information as the new regulation or amendment may not be straightforward to understand. In case of any error in his process, Bob risks legal penalties to the company as well as loss of company reputation.

The need for automation: Currently, both Alice and Bob rely on manual methods to understand existing regulations (*e.g.*, keyword searches), and can benefit from a framework that automatically extracts relevant regulatory requirements based on the provided query, all the while retaining and accounting for the contextual information present in the regulatory text. For instance, Bob would be able to automatically filter and extract information specifically related to the processing of sensitive information from different regulations. That is, such a query will return all the regulatory statements related to sensitive data, which can be used to reduce the cost of compliance, e.g., by prioritizing changes needed in business processes and evaluating legally-binding privacy policies for compliance. Similarly, Alice would be able to systematically and scalably evaluate her system against multiple regulations. This allows her to build a more efficient and practical system that can adjust with the changing regulatory requirements. That is, enabling the automated analysis of regulations would allow both Alice and Bob to understand, compare, and contrast regulations at scale, with significantly less time and effort.

III. Design Goals

Privacy regulations generally cover four areas associated with data privacy: **(1)** Scope and Definitions, **(2)** Rights and Obligations, **(3)** Privacy Principles for Data Processing, and **(4)** Enforcements Rules. Regulatory statements discuss these categories, describing mechanisms and requirements for compliance. Consider the CCPA statement presented in Figure 1:

Running Example: *"A business that collects a consumer's personal information shall, at or before the point of collection, inform consumers as to the categories of personal information to be collected and the purposes for which the categories of personal information shall be used."*

While compact, the sentence precisely describes the requirement *i.e.*, 'Business' needs to *inform* 'Consumers' *about* "categories of personal information and purpose for usage". Furthermore, the requirement includes additional context for each entity and object. For instance, phrases in the above running example are further clarified with the use of clauses, *e.g.*, "A business - that collects consumer's personal information", "personal information - to be collected", "purposes - for which categories of personal information shall be used".

Since our goal is to automate the process of understanding complex, diverse, and *unstructured*, regulatory statements, the first step towards automation would be to effectively represent these statements such that all of the important context is captured, while also making such a representation useful for privacy analysis. Motivated by this general requirement, we formulate the following key design goals that guide the design of our proposed framework for analyzing regulations:

G_1: Capturing the semantics of regulatory statements – It is important to identify semantically important phrases within a statement, which play a role in conveying its general meaning. Hence, the framework should individually identify key phrases that represent the core semantics of the statement.

G_2: Simplifying complex regulatory statements – Regulatory statements are often detail-oriented, and hence, in addition to stating the main requirement, may also include phrases that clarify the requirement or provide additional guidance on conditions that affect it. Therefore, for effective analysis, we

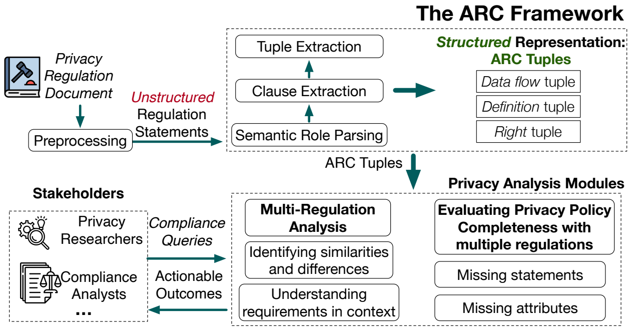


Fig. 2: Overview of ARC framework.

need to simplify such complex statements in a manner that separates the regulatory requirement from auxiliary *clauses* that clarify it, elaborate on it, or specify conditions to it.

G_3: Defining representations of the extracted semantics – Since our goal is to understand statements in privacy regulations, we need to formalize the structured representation such that it captures important details for privacy-focused analysis, and enables privacy compliance analyses.

G_4: Enabling the use of the previously defined representations to aid compliance tasks – We need to develop approaches that effectively leverages the representations defined in response to **G_3** for key tasks, such as the analysis of multiple privacy regulations for identifying similarities and contrasts (e.g., the tasks performed by Alice and Bob in Section II).

IV. The ARC Framework

Figure 2 shows ARC, a framework that systematically extracts structured information from complex privacy regulation documents, which can be used for privacy analysis. As shown in the figure, first, ARC pre-processes regulation documents to obtain unstructured regulation statements. ARC then parses these statements to extract semantic roles with regard to verbs that convey the semantics of the requirements. ARC further simplifies the statements by extracting phrases that provide additional context using constituency tree parsing [39]. Next, ARC maps each of the semantic roles and phrases to create a structured representation, which we call *ARC tuple*.

We build two modules to demonstrate the utility of ARC, the first of which enables the analysis of multiple regulations to identify similarities in requirements through the analysis of phrases. The second module allows comparison of regulation statements with policy statements (both represented as ARC tuples) to identify missing statements (*i.e.*, lack of compliance with disclosure requirements). These modules, which help with privacy regulation analysis can be used by stakeholders such as researchers and compliance analysts to significantly reduce the burden of manual effort required for compliance tasks described in Section II. We implement ARC by integrating multiple pluggable modules in the Spacy pipeline [34] (see Appendix A for implementation details). The rest of this section describes the design-level contributions of ARC.

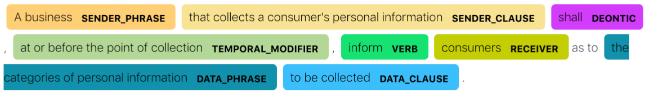

A business SENDER_PHRASE that collects a consumer's personal information SENDER_CLAUSE shall DEONTIC , at or before the point of collection TEMPORAL_MODIFIER , inform VERB consumers RECEIVER as to the categories of personal information DATA_PHRASE to be collected DATA_CLAUSE .

Fig. 3: Extraction of Phrases in the Data Flow tuple.

TABLE I: Verbs used to categorize statements.

Type	Category	Words
Deontic Modals & Verbs	obligation	must, shall, should bind.01, compel, obligate, oblige.02
	permission	can, could, may , would let.01, grant, leave, appropriate, permit, admit.01, tolerate, reserve, allow.01, allow.02
	prohibition	should not, must not, may not, could not, prohibit, forbid, nix, disallow, interdict, proscribe, veto
Data Flow Verbs	collect	collect, inform, check.01, know, obtain, access, receive, gather, solicit
	share	share.01, disclose, sell.01, provide, trade.01, return.02, transfer, give, rent, send.01, distribute, report, transmit
	retain	save.03, save.04, retain, store
	process	process, use.01, operate
	delete	delete, remove, rescind
Legal	definition	mean, define, refer.01, refer.03, include, exclude
	rights	entitle.01, have.03, invoke, include, exercise.01

A. *Semantic Parsing of Regulatory Text (**G_1**)*

Privacy regulations contain regulatory statements, which are significantly different from statements in documents such as privacy notices. This is mostly because regulations cover broader aspects related to consumer privacy and are written to precisely communicate the requirements in legal terms. Considering the complexity as described in Section I, a proper analysis and interpretation of regulatory text requires an understanding of individual context within each statement. To facilitate such understanding, Figure 3 presents the annotated version of the statement from the running example that discusses the data flow requirement. The extraction of *context* within the statement helps us fully understand the requirement. These contexts include *who* the sender and the receiver are, and the clauses tied to the requirement (e.g., temporal modifier).

We extract such context from unstructured regulatory text using Semantic Role Labeling (SRL), which is a technique for semantic parsing to identify the predicate-argument structure of sentences. For instance, for the verb "inform", SRL identifies the agent (i.e., "A business ..."), and theme ("the categories of personal information"), which are represented as arg0, and arg1 respectively. Hence, we retrain a BERT-based SRL model using a reimplementation of AllenNLP code [59] to include both argument identification (i.e., to identify semantic roles with respect to a verb), as well as verb sense disambiguation (i.e., to disambiguate the meaning of the verb). Our rationale behind including verb sense disambiguation is to facilitate easier mapping of verbs to respective arguments, without being limited to predefined set of verbs. That is, instead of directly creating mappings for verb-specific arguments (e.g., mapping 'arg0' for the verb 'share' to 'sender'), our framework considers verb senses (i.e., meaning of the verb in a given context) to help with a more precise identification of arguments. For example, Propbank lists three sense for the verb 'refer' [57], where the sense *refer.01* captures "thing being labeled", whereas *refer.02* represents "recommendation" (e.g., being referred to a lawyer). Since one of our goals is to identify definition statements (see Section IV-C), disambiguating verbs can help with the precise identification of definition tuples. We use the CONLL2012 dataset [56] for training, which is a standard dataset annotated with structural information (syntax and predicate argument structure), commonly used for evaluation [63]. Our model achieves F1 score of 86.0 for argument identification and 95.5 for predicate disambiguation.

B. Extracting Clauses from phrases ($\boldsymbol{G}_2$)

While SRL identifies arguments specific to a verb, the labeled phrases can still be complex. For example, SRL initially labels the phrase "A business that collects a consumer's personal information" as *arg0* for the verb *inform*, which is represented as a *sender* entity. However, we realize that a further simplification is needed to enable a better understanding of such phrases. That is, while one approach would be to simply extract entity from the phrase, we also want to keep track of the clause that may contain important details about the entity. Hence, we segment the phrase into <*Phrase, Clause*> pair (e.g., *Sender Phrase*, and *Sender Clause* in Figure 3) to identify and separately analyze the entity and its surrounding context. For this, we use benepar model [39] to obtain the constituency tree and perform a pre-order traversal on the tree to identify subordinate clause and its preceding phrase (see Appendix B for more details). We extract entity-specific subordinate clause to identify the phrase that provide additional context to entities. For example, we represent the subordinate clause "that collects a consumer's personal information" as a *sender_clause*, and the noun phrase "A business" as the *sender_phrase*.

C. Defining and Representing ARC Tuples ($\boldsymbol{G}_3$)

A major goal of our framework is to represent regulatory statements in a structured form so that they can be used for privacy analysis. SRL model enables such representation by extracting semantic roles specific to a verb. However, selecting verbs that are useful in the context of understanding privacy regulation can be challenging. For instance, our running example includes three verbs (i.e., collects, inform, collected) for which semantic roles can be individually obtained.

Hence, we scope the design of the framework by focusing on three major categories of statements that play important role in compliance tasks: (i) requirements that discuss the flow of data, (ii) definition statements that define terms specific to the regulation, and (iii) right statements that discuss the rights and obligations of entities. For each of these categories, we leverage the ontologies developed by prior work in legal [36] and privacy domain [7], [18], and create a list of verbs. Furthermore, we identify statements that discuss normative concepts (e.g., permission, obligation, and prohibition) using *deontic modals*, which are expressed with modal verbs (e.g., shall, must) [25]. We provide a full list of verbs in Table I. We also include verb senses for instances where multiple senses are available in the CONLL12 frame file and when disambiguation helps with the precise identification. Note that while ARC uses the verbs from existing literature [36], [18], [7], [42], our framework can be easily extended to handle new categories by creating a mapping for new verbs.

We formalize three types of ARC tuples by mapping the arguments identified by SRL model to the tuple attributes:

- *Data Flow Tuple:* We use the Contextual Integrity (CI) framework [47], which defines privacy as an appropriate flow of information. We adapt CI to represent data flow statements as:
  <*Sender, Deontic Modal, Data_Flow_Verb,*
  Receiver, Data Object, Transmission Principles>
  Thus, we represent the statement from the example in Figure 1 as: <*A business*, *shall*, *inform*, *consumers*, (*as to the categories of personal information, the purposes*), (A business *that collects consumer's personal information*, *at or before the point of collection*, as to the categories of personal information *to be collected*.>

TABLE II: Extraction of Data Flow Tuple

Regulation	Statements with Data Flow Verb	Precision	Recall	F1-Score
CCPA	410	0.768	0.75	0.759
GDPR	345	0.918	0.815	0.863
VCDPA	41	0.937	0.937	0.937
PIPEDA	196	0.943	0.857	0.898
Total/Average	992	0.866	0.804	0.834

- *Definition Tuple:* The *definition* tuple clarifies the meaning of the terminology used in regulations. We extract the description (i.e., Definiens) and the term being defined (i.e., Definiendum) using the verb (i.e., Definition Verb), which is represented as:
  <*Definiendum, Definition_Verb, Definiens*>
  We represent the statement "*personal information shall mean information about an identifiable individual.*" as: <*personal information, mean, information about an identifiable individual*>
- *Right Tuple:* The *right* tuple discusses the rights available to a specific entity. We extract the right tuple to provide an insight into how ARC can be adapted to compare rights across regulations. We represent the right tuple as:
  <*Entity, Deontic_Modal, Right_Verb, Right_Statement*>
  We represent the statement "*The data subject shall have the right to withdraw his or her consent at any time.*" as: <*The data subject, shall, have, right to withdraw his or her consent at any time*>

Note that ARC tuples are motivated by the need to enable: (i) a structured understanding of privacy requirements from regulations (i.e., Right Tuple help stakeholders understand consumer rights, whereas Definition Tuple provide context for the requirements), and (ii) privacy analysis using regulations (i.e., the Data Flow Tuple enables privacy analysis along the CI framework). Hence, ARC does not seek to comprehensively extract every category (e.g., power of member states), but only the aspects that help understand and evaluate privacy.

D. Mapping Tuple Arguments

We provide a complete list of mappings for verb-sense specific arguments in Table XI in the Appendix. We group general clauses identified using SRL model (i.e., argm_tmp, argm_prp, argm_pnc) and entity-specific clause obtained through constituency parsing (sender_clause, receiver_clause, data_clause) under the Transmission Principle attribute. Additionally, the purpose clause obtained for certain verbs (e.g., arg2 for the verb *use*) are also included under Transmission Principle.

V. Evaluating ARC Tuples Extraction

In this section, we evaluate ARC's performance in resolving regulatory statements into ARC tuples. We consider four major privacy regulations: California's California Consumer Privacy Act (CCPA) [19], Europe's General Data Protection Regulation (GDPR) [32], Canada's Personal Information Protection and Electronic Documents Act (PIPEDA) [55], and Virginia's Consumer Data Protection Act (VCDPA) [71]. For each regulation, we build a dataset, manually labeled by two authors both with more than 6 years of experience in security and privacy research. Note that for labeling in this case, we consider experience in privacy research to be more relevant than legal expertise, for two key reasons. First, the tuples are modeled after general-purpose structures (*e.g.*, definitions), or privacy-specific concepts, *e.g.*, we extracted Data Flow tuples that is based on the CI framework [47], which privacy researchers are familiar with. Second, we use a systematic methodology to identify tuples that do not rely on the expertise in the legal domain (e.g., we identify definition statements that requires the presence of both *term being defined* and *the explanation of the term*).

A. Data Flow Tuples

We first measure the effectiveness of ARC in capturing Data Flow Tuples. For this, we create a labeled dataset, where statements expressing data flow requirements (expressed as permission, obligation, and prohibition) are manually labeled. For instance, the statement "*The organization may collect personal information if..*" expresses a permission, where the organization is *allowed* to collect personal information. In contrast, the statement "*This part does not apply to organization in respect of personal information that it collects..*", expresses the scope (i.e., the section is not applicable for certain organization). Although both examples use the data flow verb (i.e., collect) and refer to the data (i.e., personal information), we only label the first statement as a data flow statement. Our goal behind this analysis is to assess the performance of ARC compared to manually labeled dataset.

Creating Labeled Dataset: We create a ground truth dataset by identifying a total of 992 statements that use at least one data flow verb listed in Table I. Two authors individually labeled each statement in the dataset with a calculated Cohen's Kappa score of 0.72, demonstrating substantial agreement. From this, we prepared the final dataset after resolving disagreements through discussion. Finally, we evaluated ARC's performance against the labeled dataset.

<u>Result 1</u>: ARC extracts Data Flow Tuples with an average F1-score of 83.4%. ($\mathcal{R}_1$) – As shown in Table II, ARC is effective in capturing Data Flow tuples expressed in regulations. In a few instances, the data flow verb is used as a noun phrase, resulting in our approach missing some of the regulatory statements; *e.g.*, in the statement: *"... transfer of personal data ... may take place ..."*, the deontic modal 'may' is associated with the verb 'take' and a noun phrase is used to describe data flow, which our approach does not currently consider. This can be improved by considering additional action verbs and noun phrases to identify data flow statements.

TABLE III: Extraction of Definition Tuples

Regulation	Statements	Metric	ARC	LexNLP
CCPA	214	Precision	0.88	0.98
		Recall	0.98	0.75
		F1-Score	0.93	0.85
GDPR	94	Precision	0.61	0.79
		Recall	0.80	0.68
		F1-Score	0.69	0.73
VCDPA	57	Precision	0.76	0.97
		Recall	1	0.81
		F1-Score	0.86	0.88
PIPEDA	73	Precision	0.77	0.97
		Recall	0.96	0.66
		F1-Score	0.85	0.78
Total/Average	**438**	**Precision**	**0.80**	**0.94**
		Recall	**0.95**	**0.73**
		F1-Score	**0.87**	**0.82**

TABLE IV: Extraction of Right Tuple

Regulation	Statements	Precision	Recall	F1-Score
CCPA	100	0.64	0.90	0.75
GDPR	237	0.77	0.92	0.84
VCDPA	27	0.66	0.66	0.66
PIPEDA	7	1	1	1
Total/Average	141	0.73	0.91	0.81

B. Definition Tuples

Next, we evaluate the Definition Tuples extracted by ARC. For this, we created a labeled dataset, where statements that define regulation-specific terminologies were labeled as definition statement. As described in Section IV, our criteria for labeling a statement as definition statement is the presence of both Definiendum (i.e., the term), and Definiens (i.e., description). Apart from evaluating ARC's perfomance against manual baseline, we also compare our approach against the extraction using LexNLP [42], which is a state of the art regex-based library for the evaluation of unstructured legal text.

Creating Labeled Dataset: We create a ground truth dataset by identifying statements from three sources: (i) "definition section" from each regulation, (ii) statements containing the definition verbs listed in Table I, and (iii) statements identified as definition by LexNLP. In total, we identified 438 statements. Two authors individually labeled the statements with a calculated Cohen's Kappa score of 0.83, demonstrating high inter-annotator agreement. We prepared a final dataset after resolving disagreements. We describe our results below:

<u>Result 2</u>: ARC extracts Definition Tuples with an F1-score of 87% on average, outperforming LexNLP ($\mathcal{R}_2$) – As shown in Table III, ARC identifies definition tuples with high precision of 80% and a recall of 95%. While LexNLP has considerably high precision, we find that ARC outperforms LexNLP in terms of recall. This is because LexNLP performs additional processing based on the regex expression (e.g., checking for quotes that highlight the defined term). Moreover, we find that most of the false positives for ARC are due to statements with the "include" verb, which can be fixed with a post-processing step that identifies the term. We also find that LexNLP simply fails in complex instances; *e.g.*, the sentence *"Biometric information includes, but is not limited to, imagery of the iris..."*, is captured by ARC but missed by LexNLP potentially because of phrase after the verb "includes".

C. Right Tuples

Finally, we evaluate the Right Tuples extracted by ARC. For this, we created a labeled dataset composed of statements that describe the right(s) given to an entity. As described in Section IV, our criteria for labeling a statement as a right statement is the presence of both the "entity", who is entitled to the right and "the right statement", which describes *what* the right is about. For example, in *"A consumer shall have the right to request that a business that collects.."*, "a consumer" is the entity, and "the right to request..." is the right statement.

Creating Labeled Dataset: We created a ground truth dataset by identifying a total of 141 statements that use the term *right* anywhere in the statement. We simply use the term *right* because we seek to identify statements that discuss rights granted to an entity in its strictest sense, also known as claim-rights [40]. Two authors individually labeled each statement with a calculated Cohen's Kappa score of 0.91, demonstrating high inter-annotator agreement. Next, we use the dataset (i.e., after resolving disagreements) to evaluate ARC's performance.

Result 3: ARC is able to extract Right Tuple with an average F1-score of 81%. ($\mathcal{R}_3$) – As shown in Table IV, ARC identifies rights tuples with high precision of 73% and a recall of 91%. Similar to complexities identified in extracting Data Flow tuples, we find that ARC misses out on identifying Rights tuple in some complex statements. For example, in the phrase "any rights the consumer may have to appeal the decision to the business", ARC fails to identify right_statement (i.e., arg1) for the verb 'have'. However, our results show that ARC extracts tuples reasonably well even with an unoptimized approach. Given ARC's modularity, each component can be separately extended to focus on a specific use case.

VI. Multi-Regulation Analysis using ARC ($\mathbf{G}_4$)

We develop a phrase representation model to enable the comparison of ARC tuples. As we discussed in Section II, analysts currently rely on manual methods to identify similarities and differences between regulations. However, even for an expert analyst, manual methods (e.g., keyword search based approach) can be unreliable because *the searches* are solely dependent on lexical similarity (i.e., the use of similar words). Relying on lexical similarity can be particularly ineffective when we seek to compare regulations written by legal experts from around the world. An alternative would be to use a model that directly compares statements (e.g., using a sentence classifier). While a direct comparison can be useful to some extent, this approach can only provide a coarse reasoning for its results. ARC addresses both of these issues by building a model, ARCBert, which enables querying and comparison of statements at a phrase level granularity.

To understand the utility of identifying phrase-level similarity, consider the following phrases from two different regulations (that our model identified as similar):

$\mathbf{P}_1$ – *in a form that is reasonably accessible to consumers*
$\mathbf{P}_2$ – *in a commonly used electronic form*

It is evident that the nature of the phrases is similar as both describe the manner in which information is to be delivered, and an analyst can use this result to quickly identify and examine requirements of this (similar) nature.

To enable the comparison of statements at phrase-level granularity, we train ***ARCBert***, a `BERT` based model, which creates phrase embeddings for phrase representation. For this, we obtain paraphrased phrases used by Phrase-BERT [73], which helps the model to learn general semantic relatedness between phrases. We then add legal and privacy context by including 100K most frequent contextual phrases extracted from (22k US privacy bill [28], 37K EU legal documents [29], and 56 global privacy regulations [68]). We then fine-tune BERT on this dataset to build ARCBert. We create ARCBert by adopting the same process as Phrase-BERT for fine-tuning step, where the goal is to bring together semantically similar contexts. Both the process for curating the phrase-in-context dataset and the fine-tuning process are further described in Appendix C. Finally, we use ARCBert and report similarity scores between phrases based on cosine similarity, which is a metric commonly used to measure text similarity [73], [46].

A. Baseline Comparison of Phrase Similarity

We evaluate ARCBert's ability to identify similar statements in comparison with key baseline approaches: (1) a naive method that uses *lemmatized keyword search*, which represents the best keyword search approach that analysts may currently perform, and (2) a *GloVe vectors* based similarity search.

To provide an understanding of how ARCBert can help with the identification and prioritization of similarity results, we evaluate the similarity scores for each approach. That is, a model returning very few similar results may limit analysis of important contexts by only delivering phrases that are exactly the same (akin to a keyword search), whereas one that returns all phrases that are even slightly/broadly similar, thereby returning an extremely high number of similar results, can make it hard to identify and prioritize truly semantically-similar phrases. Thus, we seek a middle ground, an approach that will output a *reasonable* similarity score that expresses the semantic-relatedness between phrases, without being overly restrictive or permissive in its notion of similarity.

To understand our evaluation of similarity results, let us consider the example of two similar phrases (from Canada and GDPR regulations) with similarity score of 0.782:

$\mathbf{P}_1$ – *without delay on that matter*
$\mathbf{P}_2$ – *no later than thirty days after the date of the request*

ARC determined P_2 to be one of the most similar phrases to P_1. We observe that while the keyword search completely fails to find similarity, ARC provides high similarity score, which can be attributed to the similarity gleaned from semantics of the phrase (i.e., both phrases talk about temporal requirement). Note, however, that estimating *similarity* manually is not always straightforward, and requires a good understanding of *user-expectations* and *expertise* before it can be used for automated reasoning. Hence, we develop a similarity analysis approach relying on cosine similarity that would effectively investigate the expected properties of similarity results. We now evaluate ARC's similarity results against baseline approaches.

Methodology: We compare the similarity scores generated by ARCBert against two additional methods. We describe how we obtain similarity scores for each approach below:

1. Naive Approach: We consider a naive approach, where

TABLE V: Examples of similarity results across regulations. Phrase pairs #1 to #4 were manually validated as similar, #5 to #8 as not similar.

No.	Phrase 1	Phrase 2	Naive	ARCBert	GloVe
1.	in an intelligible and easily accessible form	in a form that is generally understandable	0.29	0.88	0.86
2.	for purposes other than those for which it was collected	for any purpose other than those expressly listed in this section	0.48	0.77	0.92
3.	in writing and without delay	in electronic form	0.22	0.73	0.61
4.	unless otherwise required by law	Except as otherwise provided in this chapter	0.17	0.88	0.79
5.	access to personal information	personal data	0.33	0.69	0.82
6.	to third parties having access to the information in question	the other information sought by the consumer	0.24	0.62	0.88
7.	of the request made by the individual	a record of the request	0.5	0.65	0.9
8.	that the organization notified an institution or part under paragraph (2.2)(a)	Personal data processed by a controller pursuant to this section	0	0.47	0.78

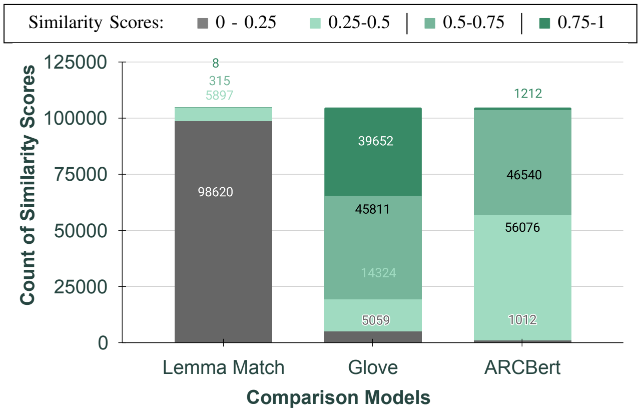


Fig. 4: Similarity Score Comparison for PIPEDA vs GDPR

for a given pair of phrases p1 and p2, each containing unique set of word lemmas, we obtain similarity score as:

$$similarity\_score = \frac{intersection(p1, p2)}{average(len(p1), len(p2))} \quad (1)$$

2. GloVe: We evaluate similarity scores based on GloVe vectors, which builds vector representation for words. To obtain the similarity score between phrases, we average the pre-trained token embeddings as done by prior work [73].

3. ARCBert: We calculate cosine similarity based on phrase embeddings produced by ARCBert model.

To scope our analysis, we obtain a list of unique phrases (sliced to 15 word tokens per phrase for a fair comparison) from Data Flow tuples identified by ARC. In total, we obtain 1,134 phrases (i.e., 679 Entity Phrases, and 455 Transmission Principle phrases) from four regulations. We identified 218 phrases for PIPEDA, 49 for VCDPA, 481 for GDPR, and 386 for CCPA. Note that all the baseline approaches are already benefitting from ARC's extraction of clauses from statements. We then compare each phrase with the list of phrases in a separate regulation and report the similarity scores (between 0 - 1), which we divide into four ranges for brevity. Figure 4 plots the number of similar phrases found with similarity scores in each given range, for the two baselines and ARCBert.

Result 4: ARC improves over naive method of extracting similar phrases in all four regulations. ($\mathcal{R}_4$) – As illustrated in Figure 4, we see that naive method is not able to extract similarity results as only 8 pairs are identified to have > .75 similarity score. In contrast, ARCBert identifies 1212 phrase-pairs with a high degree (0.75-1) of similarity. We observe a similar trend in the remaining set of regulations (see Appendix E) *i.e.*, the naive method generally returns similarity score below 25%, which speaks to the difficulty in identifying similar phrases through a keyword search based methods.

Result 5: ARC shows improvement over off-the-shelf models (GloVe) by restrictively identifying similar phrases. ($\mathcal{R}_5$) – In contrast to the naive approach, GloVe is too permissive and essentially marks *everything* as broadly similar, with a large number of phrase pairs $(39,652)$ falling in the 0.75-1 similarity score range, and a majority $(45,811)$ in the 0.5-0.75 range. In contrast, we find that ARCBert is restrictive as it limits to $1,212$ highly similar phrases. As presented in Appendix E, we see a similar trend across all regulation comparisons *i.e.*, ARCBert consistently identifies lower number of phrase pairs with (>.75) similarity score, and higher number of pairs with (<.25) similarity scores compared to the GloVe vector based approach. Hence, our results show that ARCBert is more restrictive in identifying similar phrases, leading to better prioritization of phrases that are truly similar (instead of marking everything as very similar, as GloVe does).

To further illustrate how ARCBert strikes a balance between the overly restrictive naive (keyword-search) approach, and the extremely permissive off the shelf Glove model, we present example phrase-pairs with similarity scores. As shown in Table V, the naive method often results in lower similarity score and is unable to reason about the semantics of the phrases that are semantically similar in nature. In contrast, GloVe consistently returns higher similarity score relative to ARCBert, for all but 3 of the results, and in all but one instance, returns a similarity score of >0.75. We find that ARCBert accurately assigns higher similarity scores for phrases that are used in a similar context. For example, in example 3, both of the phrases (i.e., phrase 1: 'in writing..', and phrase 2: '..electronic form') discuss similar context, which represents *how* the document should be presented. ARCBert detects high similarity score of 0.73 compared to 0.22 by naive method and 0.61 by GloVe model. In contrast, for example 6, where the phrases are not similar, (i.e., *'to third parties having access...'* vs *'information sought by the consumer'*, ARCBert appropriately assigns a lower similarity score (0.62), whereas GloVe permissively assigns much higher (0.88), presumably since both phrases talk about information, which is too broad a similarity to be useful. Another interesting instance is the dissimilar phrases in example 7, where ARCBert correctly assigns lower similarity score despite having majority of the words in common with phrase 1, in contrast to GloVe. This shows that ARCBert does not solely rely on lexical similarity, which is one of the

TABLE VI: Count of Top-1 Similar Definition with score > 0.75

	CCPA	GDPR	VCDPA	PIPEDA
CCPA	-	41	82	55
GDPR	26	-	13	27
VCDPA	22	20	-	14
PIPEDA	13	10	11	-

TABLE VII: Top-3 Similar Definitions across Regulations

Regulation	Top-3 Similar Definiendum	Score
CCPA vs GDPR	(processing, processing) (biometric information, genetic data) (business, controller)	0.93 0.87 0.81
CCPA vs PIPEDA	(financial incentive, commercial activity) (signed, electronic signature) (biometric information, personal health	0.87 0.70 0.70
CCPA vs VCDPA	(control or controlled, control or controlled) (service provider, processor) (processing, process or processing)	0.99 0.83 0.96
GDPR vs PIPEDA	(data concerning health, personal health information) (processing, electronic document) (personal data breach, breach of security safeguards)	0.87 0.85 0.88
GDPR vs VCDPA	(processing, process or “processing” (controller, controller) (profiling, profiling)	0.97 0.97 0.95
VCDPA vs PIPEDA	(personal data, personal information) (State agency, Responsible Authority) (Sensitive data, personal information)	0.87 0.63 0.74

primary reason for fine-tuning with paraphrased phrases using Phrase-BERT [73]. This evaluation shows that ARC effectively extracts similar phrases relative to the naive approach and GloVe. Next, we perform a manual validation of the similarity results produced by ARCBert.

B. Extrinsic Evaluation of Phrase Similarity

We perform an evaluation of similar phrases identified by ARCBert, where two authors independently validated the semantic similarity results. We describe our evaluation below.

Methodology: We build an evaluation dataset of 100 phrases (<15 words), which contains an equal distribution of random phrases belonging to different SRL arguments (i.e., for all entity and transmission principle arguments). For each phrase, we extract top 3 similar phrases with a similarity score >0.75 by comparing it against phrases in corresponding regulations. Two authors independently labeled a total of 237 phrases identified by ARCBert with a binary label (similar vs. not_similar). We consider two phrases to be similar if they essentially discuss similar concepts. For example, two phrases describing *“by 25 May 2018”* and *“prior to 24 Jan 2016”* will be considered similar because both phrases discuss a specific date/deadline.

Result 6: ARCBert is able to identify similar phrases across regulations ($\mathcal{R}_6$) – We find 186/226 (82.30%) instances, where both evaluators consider the similar phrases identified by ARCBert to be similar. Moreover, evaluators marked at least 1 phrase to be similar among top-3 predictions in 87/100 (87%) cases. In contrast, 40/226 individual instances were given conflicting labels by the evaluators. We observe that the disagreement stemmed from the varying notion of similarity and the level of abstraction at which similarity was being considered. While judging the similarity may not always be straightforward, we notice that ARCBert can be highly useful when phrases are used in context of the entire statement. For example, consider the phrase-pair that ARCBert extracted as similar (and evaluators found similar as well): *“on behalf of the business that provided the personal information”* and *“on behalf of a controller”*. We were able to quickly use this similarity extracted by ARCBert to identify the following two statements that discuss a processor-specific requirement, from two different regulations:

CCPA – *“A service provider shall not retain, use, or disclose personal information .. except to process or maintain personal information on behalf of the business..”*
GDPR – *“Each processor .. shall maintain a record of .. processing activities carried out on behalf of a controller”*

C. Statement Analysis using Definition Tuples

Since ARC enables automated identification of similar phrases, one direct evaluation of utility can be to analyze the similar definitions across regulations by comparing *definiens*. We now evaluate how well ARC identifies similar terms defined across regulations.

Methodology: Recall that a definition tuple is represented as <*Definiendum, Definition_Verb, Definiens*>. To understand similarity results, we extract all manually validated definition tuples for each regulation. We obtain 180 definitions for CCPA, 41 for GDPR, 37 for VCDPA, and 25 for PIPEDA. We then compare definiens between regulations and sort it based on similarity score. We study the most similar definitions with the score above 0.75 for each definition tuple. We now discuss the insights based on our analysis.

Result 7: ARC can be used to understand similarities across regulations ($\mathcal{R}_7$) – As shown in Table VI, ARCBert identifies similar definitions across regulations. Among all the comparisons, we find that CCPA shares the highest number of similar definitions with VCDPA. Intuitively, this result is reasonable given that both CCPA and VCDPA are U.S. based privacy regulations. In contrast, we see that PIPEDA shares least number of definition similarities with all other regulations. This can be attributed to lower number of similarities in definitions as well as lower number of defined terms in PIPEDA (i.e., 25 definition tuples).

Result 8: ARC identifies similar definition terms, which may be missed by manual comparison methods. ($\mathcal{R}_8$) – As shown in Table VII, we find that ARC is able to identify semantically similar definition terms with high accuracy. The table shows top-3 similar terms that ARC determined to be the most similar among the entire list of definitions statements across regulation. While most of the results in Table VII are self-evident because of the use of same definiendum, we also find multiple instances where the similarity results required further investigation. We elaborate on our analysis next.

1. ARC helps to improve understanding of terminologies used in privacy regulations: We find instances where multiple regulations use similar terminologies (e.g., *profiling* in GDPR is also defined as *profiling* in VCDPA). However, we also find cases where a completely different definiendum is used to describe similar concept. For example, when comparing GDPR and CCPA, we find that "controller" has the highest similarity to the definition of the term "business" (score of 0.84). We present the definitions below:

GDPR – ***'Controller'*** *means the Union institution or body or the directorate-general or any other organizational entity which, alone or jointly with others, determines the purposes and means of the processing of personal data; where the purposes and means of such processing are determined by a specific Union act, the controller or the specific criteria.*

CCPA – ***"Business"*** *means a sole proprietorship, partnership, limited liability company, corporation, association, or other legal entity that is organized or operated for the profit or financial benefit of its shareholders or other owners that collects consumers' personal information or on the behalf of which that information is collected and that alone, or jointly with others, determines the purposes and means of the processing of consumers' personal information, that does business in the State of California.*

After reading more carefully, we notice that both of the sentences discuss similar concept, describing an entity that "determines the purpose of processing...", which is termed as "Controller" in GDPR, and "Business" in CCPA. This example demonstrates ARC's usefulness in automatically identifying similar definitions across regulations.

2. ARC helps in the identification of contrasting difference across privacy regulations: We also found instances where top similar definitions had a relatively lower similarity score. For instance, while comparing CCPA with Canada's regulation, we noticed that the closest term to "Biometric information" was found to be "personal health information", with a similarity score of 0.70. We present the definitions below:

CCPA – ***'Biometric information'*** *includes, but is not limited to, imagery of the iris, retina, fingerprint, face, hand, palm, vein patterns, and voice recordings, from which an identifier template, such as a faceprint, a minutiae template, or a voiceprint, can be extracted, and keystroke patterns or rhythms, gait patterns or rhythms, and sleep, health, or exercise data that contain identifying information.*

PIPEDA – ***'personal health information'***, *with respect to an individual, whether living or deceased, means: information concerning the physical or mental health of the individual*;"

Since "*personal health information*" was identified to be most similar to "*biometric information*", we investigated further to confirm that Canada's PIPEDA regulation does not define "biometric information" separately. We found that PIPEDA does not use the word "biometric" anywhere in the document. Similarly, we find that VCDPA does not define "biometric information" either. Analysts can use ARC for a similar analysis by quickly evaluating similar statements, instead of manually searching for terms across regulations.

VII. Applying ARC to New Privacy Regulations

To further explore the effectiveness of our approach, we run ARC on 16 additional privacy regulations to analyze its general applicability. This extrinsic evaluation demonstrates ARC's performance on a diverse set of regulations, and shows how the results can be used to aid privacy regulation analysis.

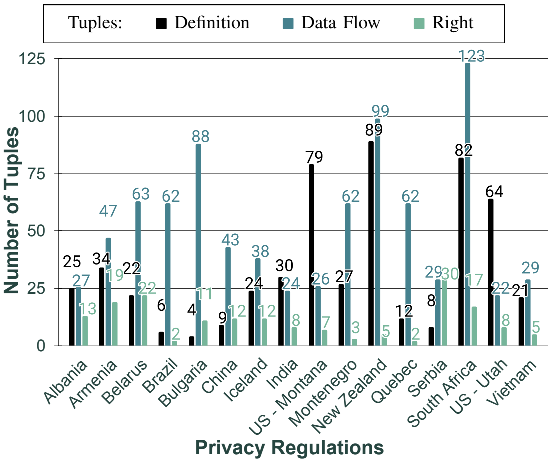


Fig. 5: Extraction of Tuples from recent regulations using ARC

Methodology: We explore two aspects relevant to privacy regulation analysis: (i) the generalizability of ARC's tuple extraction on tuples from diverse regulations, and (ii) the utility of ARCBert model in identifying similar phrases, i.e., the clustering of similar phrases.

1. Extracting ARC Tuples: We obtain 16 additional privacy regulations from respective government websites [69], which includes unofficial translations by legal experts of 2/16 (China [26], and Brazil [44]). We use the methodology described in Section IV to extract Definition, Rights, and Data Flow Tuples from each regulation separately.

2. Clustering Semantically Similar Phrases: We identify clusters of similar phrases across all 20 regulations (i.e., including the prior four) by applying k-means [61]. For this, first we obtain a list of phrases belonging to each SRL argument, where the longer phrases (>15 words) containing multiple verbs are broken down into smaller constituents. Second, we compute k-means clustering using the phrase embeddings obtained from ARCBert, where the number of clusters is estimated using Gap Statistic algorithm [67] by setting the upper bound of 100 clusters. Finally, instead of manually assigning a topic to each cluster, we use BerTopic [12] to create interpretable topics. Note that we perform analysis on semantic roles (e.g., arg0) instead of limiting to ARC tuple attributes (e.g., sender attribute) because we seek to capture diverse phrases used in different context throughout the regulation. Since our models are more generalized, they can be used to predict cluster categories for different ARC tuple attributes.

Results: We extract ARC tuples based on 16 new privacy regulations demonstrating ARC's effectiveness in identifying important representations from regulations. Moreover, we also build clusters of semantically similar phrases based on 31,822 unique phrases extracted from 20 regulations. Our extraction of semantically similar phrases can be used to understand complex requirements at a finer granularity (e.g., identifying different categories of purpose/temporal requirements), and to build automated systems for privacy analysis (e.g., prior

TABLE VIII: Phrase Clusters and Representative Examples

SRL Args	Cluster Count	Example Cluster Topic	Representative Examples
arg0	81	data_controller-personal_data-processor	the processor of personal data the controller and processor the personal data processor
arg1	94	personal-information	information on his or her personal data the personal information collected personal information concerning him
arg2	90	person_individual-body	the person or body a california resident a resident of the state
argm_prp	79	order_data_protect	in order to ensure adequate data security in order to protect public order and security in order to protect national security
argm_loc	78	member_state-third_country	in third countries on the territory of a member state outside of the eu
argm_tmp	98	least_months_12	for at least 12 months before next in the preceding 12 months for at least one year

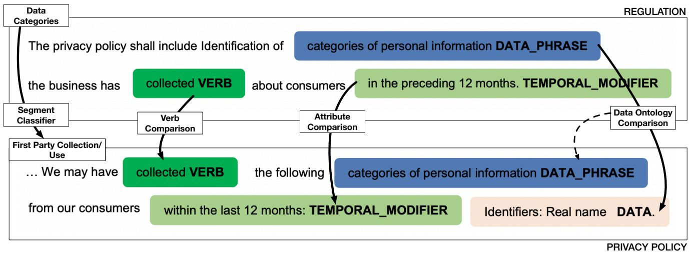

Data Categories
REGULATION
The privacy policy shall include Identification of
categories of personal information DATA_PHRASE
the business has
collected VERB
about consumers
in the preceding 12 months. TEMPORAL_MODIFIER
Segment Classifier
Verb Comparison
Attribute Comparison
Data Ontology Comparison
First Party Collection/ Use
… We may have
collected VERB
the following
categories of personal information DATA_PHRASE
from our consumers
within the last 12 months: TEMPORAL_MODIFIER
Identifiers: Real name DATA.
PRIVACY POLICY

Fig. 6: Comparing Privacy Policy with Regulation.

work built a taxonomy of purpose phrases from privacy policies [18]). Our results show that ARC can be generally applied to diverse privacy regulations, as described below:

<u>Result 9</u>: ARC consistently extracts ARC tuples from diverse regulations ($\mathcal{R}_9$) – ARC extracts 536 Definition Tuples, 844 Data Flow Tuples, and 176 Right Tuples from 16 regulations. As shown in Figure 5, ARC consistently identifies high number of Data Flow and Definition tuples, which are most relevant for privacy researchers. We also observe that Right tuples are not as common in regulations, which is expected since our tuple only extracts entity-specific rights. For example, Quebec's [41] privacy regulation mentions the keyword "right" only 16 times throughout the document.

<u>Result 10</u>: ARC can be used to identify meaningful cluster of phrases ($\mathcal{R}_{10}$) – We present our results for identifying phrase clusters for different semantic roles. Table VIII highlights the results for 6 major semantic roles, describing the number of clusters, examples of cluster topic, and three representative examples for each cluster. For example, we identified 98 clusters for argm_tmp (temporal modifiers). The topic "least_months_12" gives an initial idea of specific time constraints clustered under this category. The representative examples provides three common phrases included in the cluster. We include the full list of semantic phrases for this category in Figure 8 in the Appendix. Note that we exclude examples of clusters for other semantic roles (e.g., argm_mnr) for brevity. We release the dataset of clustered phrases for each semantic role [62].

VIII. Privacy Policy Compliance Analysis

In this section, we demonstrate how ARC can be used to validate the privacy policy against privacy regulations. Figure 6 presents a representative example of our policy analysis showing how we perform validation using context derived as ARC's tuple attributes. For the regulatory requirement that discusses a specific requirement-type *(e.g., Data Categories)*, we first identify the respective policy statement belonging to same category, using a policy segment classifier (described below). We then compare individual attributes (e.g., *Data, Verb, and Temporal*) between two statements to verify the compliance of privacy policies.

We now describe the methodology that enables such comparison followed by the evaluation results.

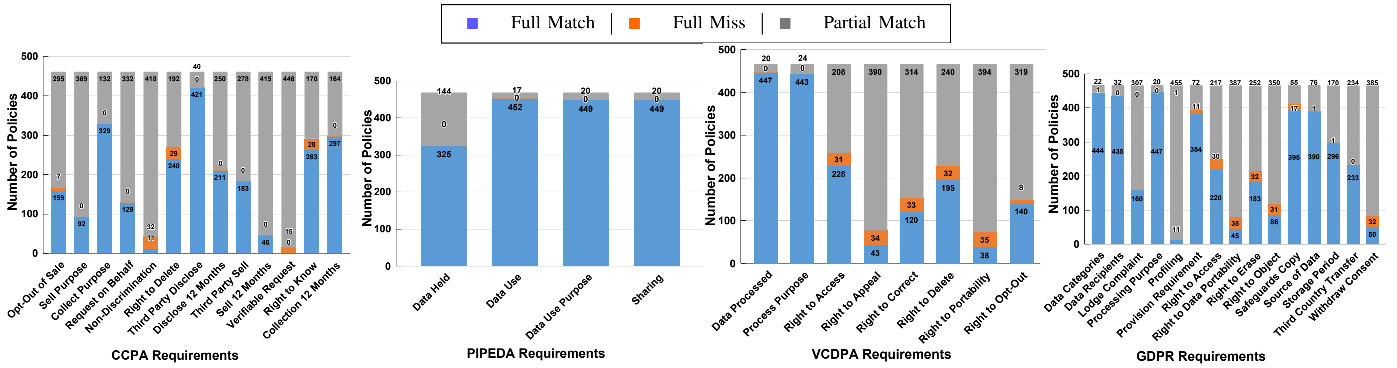
Methodology: To enable the analysis of privacy policy against privacy regulations, we devise a methodology that performs a series of tuple attribute comparisons, described as follows:

1. Collecting Privacy Policies: We use semi-automated method to obtain the privacy policy of S&P 500 companies from three different jurisdictions (i.e., United States, Canada, and Europe). We crawled the privacy policy link using VPN and manually resolved challenges when multiple policies were available (e.g., AllState [5]), or additional links were embedded within privacy policy page (e.g., Stryker [65]). We also handled instances where a separate website was maintained for a jurisdiction (e.g., Amazon Canada [6]), or a subsidiary hosts the policy (e.g., Alphabet Inc does not provide privacy policy, hence we downloaded Google's). In one instance, *i.e.*, Coterra [22], we could not find any privacy policy. Furthermore, we also encountered instances where privacy policy text could not be obtained because of crawling/parsing issues (e.g., Nucor [2] uses PDF as supplemental links). This resulted in $1,864$ regulation-specific privacy policies.

2. Building Multi-label Policy Segment Classifier: We built a segment classifier to identify the category of policy segments, which we map with the requirement tuples. For this, we train the category level multi-label classifier based on BERT, using the OPP-115 dataset in a manner similar to prior work [33]. We used 65 policies for training and kept 50 policies as the testing set, and calculated macro-average of the precision in predicting the presence of each label. Table X in Appendix F presents the results for category-level classification, which are comparable to the results in prior work [33], with an average F1-score of 0.86.

3. Extracting Privacy Policy Tuples: We use ARC to extract tuples from privacy policies, just as we do for privacy regulations. To adapt ARC for privacy policy documents, we simply extract tuples using the *main verb* for each statement instead of relying on deontic modal (which we use for regulations).

4. Building Requirement Tuples: We obtain requirement tuples from each of the four privacy regulations by identifying requirements that apply to disclosure practices in the privacy policy. After manual evaluation, we curated 40 regulatory requirements that apply to privacy policy (i.e., 13 requirements from CCPA, 4 from PIPEDA, 8 from VCDPA, and 15 from GDPR). We also preprocessed the tuple attributes extracted



(a) CCPA (b) PIPEDA (c) VCDPA (d) GDPR

Fig. 7: Compliance Analysis Results of S&P 500 companies across four regulations

by ARC to help with the direct comparison with privacy policies. This step involved incorporating additional keywords and removing extra phrases. For example, to enable verb class comparison, we included keywords identified in Table I. Similarly, we processed *data phrase* to only include data object for an easier comparison. Lastly, we prepared keyword lists to enable term comparison by filtering out irrelevant terms (e.g., stopwords). Note that out of 40 requirements, 5 requirements did not belong to Data Flow or Right Tuples. For example, CCPA requires privacy policy to provide instruction on how an authorized agent can make request on a consumer's behalf. Hence, we created an additional tuple that simply performs a term match. Note that since ARC extracts the tuples for each requirement, there was significantly less manual effort in adapting the tuples to aid our analysis. Furthermore, we created mappings for each requirement to the most appropriate category labels. Our intuition is to reduce manual effort by providing analysts with the most relevant statements instead of entirely discarding new requirements (see Section X). Figure XII in the Appendix presents our mapping of requirements with the categories in the privacy policy.

5. Tuple Analysis: We compare attributes from policies against the regulation requirements, as shown in Figure 6. The compliance module performs three main comparisons:

- **Data Ontology Comparison –** We use the ontology released by PolicyLint [7] and extend it with data objects from privacy regulations to compare data objects. We check for a subsumptive relation for a match, *i.e.*, we consider it to be a match if the data object in the policy tuple is subsumed under the data object defined in the regulation tuple.
- **Attribute Value Comparison –** We perform a keyword search of terms within the extracted tuple attribute for verb, right, and term. We consider it to be a match if the given keyword is identified within the tuple attribute object, e.g., collect verb as defined in Table I being used by both tuples.
- **Attribute Presence Comparison –** In contrast to performing ontology match or keyword search, we simply check for the presence of attributes to compare purpose, temporal, sender, and receiver attributes. This choice is based on our observation that regulations often have a general requirement related to these attributes. While our analysis can be further enhanced by enabling comparison at a finer granularity, building advanced taxonomy of phrases for privacy policies is beyond the scope of this work (see Section X). Our results in Table IX show reasonable precision in identifying missing and matching statements using our approach.

TABLE IX: Privacy Policy Compliance Validation Results

Regulation	Missing Statements	Full Match Statements	Matches in Partial Match
CCPA	76/111 (68.46%)	193/262 (73.66%)	268/350 (76.57%)
GDPR	128/192 (66.67%)	340/370 (91.89%)	190/305 (62.29%)
PIPEDA	0	173/175 (98.85%)	36/64 (56.25%)
VCDPA	135/173 (78.03%)	244/253 (96.44%)	163/226 (72.12%)
Average	**71.05%**	**90.21%**	**66.80%**

6. Privacy Compliance Analysis: We compare requirements in the regulation with the privacy policy statements. We create three major categories to aid compliance analysis: (1) *Full Match*, where a match is identified by the segment classifier, and passes the tuple analysis, (2) *Partial Match*, where segment classifier finds a match but it fails the tuple analysis, and (3) *Full Miss*, where even the segment classifier finds nothing. Hence, for each requirement tuple, ARC maps all privacy policy tuples to the relevant categories, making it easier for experts to review the compliance. Note that during manual validation, we limit our analysis to check the consistency of policy statement only against the requirement tuple. For example, for a given requirement, we consider the privacy policy to be a Full Match if any one of the policy tuple satisfies the requirement, i.e., we do not check inconsistency between privacy policy statements.

Results: We performed analysis for $1,864$ policies for a total of 40 regulatory requirements. In total we analyzed, $66,186$ policy segments for CCPA-specific requirements, $71,134$ for GDPR, $35,817$ for PIPEDA, and $42,670$ for VCDPA. The results of the compliance analysis are illustrated in Figure 7.

Result 11: ARC identifies 476 **instances of missing statements across S&P 500 companies.** ($\mathcal{R}_{11}$) – For CCPA, we find 111 instances of missing statements across 38 company policies with missing requirements for 5 major requirements (e.g., Right to Non-Discrimination, Opt-out of Sale). For VCDPA, we find 111 instances of missing statements across 35 policies with missing statements for 6 requirements. Similarly, for GDPR, we find 192 instances of missing statements across 49 policies with missing statements for 9 requirements.

Further, we find that majority of the policies contain statements that partially match the requirements. This is because of two main reasons: (i) we focus on precision in identifying *Full Match* and *Full Miss* statements, and (ii) our framework operates at statement level and does not retain the context of statements across paragraphs (which may lead to full matches at other locations in the regulation). Hence we leave remaining statements that the framework cannot confidently reason for further analysis. Even for the post-analysis of these partial matches, ARC significantly reduces manual effort by categorizing statements represented as tuples under each requirement that the analyst can evaluate to validate compliance (see $\mathcal{R}_{13}$).

Manual Validation: We performed a manual validation for a sample of 50 policies across each category (i.e., full match, full miss and partial match), validating statements from a total of 121 companies with regard to CCPA requirements, 243 companies with regard to GDPR, 94 companies for PIPEDA, and 120 companies for VCDPA. ARC provides the result in a JSON form (see Appendix 9 for example), where for a given statement, all the matched statements together with the attribute specific comparison is provided. We use the JSON file to aid our compliance analysis. That is, for full match statements, we simply study the statements listed under *Full Match* category to verify that the policy does indeed satisfy the requirement. For partial match, we study statements and the missing attributes shortlisted under *partial_match* category to check for scenarios, where the requirement can be satisfied after considering multiple statements (i.e., attribute matches are dispersed across multiple statements). Finally, in case of full miss statements, and scenarios where none of the partial matched statements meet the requirement, we read through the privacy policy to confirm that none of the statements meet the requirement. The *Full Miss* cases were independently evaluated by two authors with >5 years experience in privacy research.

Result 12: ARC reasons about policy statements with high accuracy ($\mathcal{R}_{12}$) – We found that ARC is able to reason about policy statements across regulations with high accuracy based on our manual validation. As shown in Table IX, we find that we get an average of 72.12% accuracy in identifying missing statements. Similarly, for full match statements, ARC obtains the accuracy of 90.13%. Lastly, we find that the 66.80% of the statements in partial matches are indeed full match after manual analysis. Moreover, we were able to quickly identify 288 instances of missing statements from this list, since ARC provided context into what was missing in the policy statements, which allowed for easier manual querying for a specific section in the privacy policy.

Result 13: ARC reduces effort required for manual validation ($\mathcal{R}_{13}$) – Since ARC provides context for every result by enlisting the requirement-specific statements and the attributes, the validation required anywhere between 2-5 minutes. Among these, the *Full Match* cases took the shortest amount of time as the evaluator could quickly go through the matching statements, whereas *Full Miss* cases took the longest as it required going through privacy policies. For *Partial Match*, ARC made the analysis a lot easier by providing context into *why* policies are flagged as partial matches. Note that this kind of analysis would not have been possible with a simple classifier that directly reasons on a statement. Figure 10 in the Appendix shows the list of attributes that ARC found missing across four regulations (e.g., ARC identifies a segment match but partially misses the sender attribute). This significantly reduces the overhead in analyzing policies, which is also the reason why we were able to identify 33.2% missing statements while analyzing 945 instances of partial match statements identified by ARC.

IX. Related Work

We now describe the closest prior work relevant to ARC, both in terms of the technique (analyzing legal text using NLP), and the application domain (privacy analysis).

Legal Text Processing using NLP: Processing information from unstructured legal text has remained an exciting area for both researchers and practitioners for a long time. For instance, LexPredict [43] enables segmentation, fact extraction, and classification from contracts, plans, policies, and procedures, whereas, BlackStone [1] provides models trained on long-form text containing common law and entity concepts. Similarly, there is an ongoing effort to improve Legal Question Answering. For instance LegalAI [75] provides legal judgement prediction, similar case matching, and legal question answering. Fawei et. al. [30] developed semi-automated legal ontology generation tool for legal question answering. Similarly, Legal-BERT [20] trains BERT model to evaluate the performance in the task of multi-label classification. However, in all of these cases, the frameworks focus on analyzing legal text for a generic task or solving challenges in criminal law. In contrast, our approach adapts existing NLP tools and techniques to enable an automated analysis of privacy regulations.

Privacy Analysis: In privacy research, the closest work that aids privacy compliance is PrivGuard [3]. PrivGuard performs static analysis of programs and compares it with *base policies* to verify compliance, where the base policies are encoded manually by experts. Our work complements frameworks like PrivGuard by reducing the manual effort to convert regulation text into formalized policies. Similarly, to understand the privacy promises, prior works have mainly focused on analyzing the content of privacy policies to evaluate inconsistencies [8], [18], vagueness [13], [17], consent and opt-out choices [60], [50], contradictions [7], [23], [74], and regulatory compliance [14], [15], [58]. Moreover, prior works have also focused on enabling automated understanding of privacy policies [27], [4], [33]. Additionally, prior work has used the theory of Contextual Integrity [47] to study the alignment with U.S. Children's Online Privacy Protection Act (COPPA) [9], and to evaluate its viability in privacy policies [64]. However, none of these works focus on analyzing privacy regulations, which is the focus of our work.

X. Discussion and Limitations

This paper demonstrates ARC's effectiveness in representing privacy regulation statements and enabling regulation and compliance analysis. We now identify the limitations of our work and discuss areas of improvement.

1. Optimizing the NLP pipeline: The goal of this paper is to lay the ground work towards automated and scalable analysis of privacy regulations. Therefore, while ARC leverages several state of the art NLP techniques, there is certainly room for improving and optimizing the NLP pipeline. For instance, to improve precision of tuple attributes, Coreference Resolution techniques can be applied to resolve references to entities (e.g., use of pronouns). Similarly, adapted NER models can be used to identify regulation specific entities and reason over them. Our evaluation showed that these factors do not hurt ARC's overall accuracy, however, improvement along this area, easily integrated using ARC's modular architecture, can significantly help make ARC more versatile. Note that while we identified scenarios where ARC needed improvement (as discussed in Section V) we avoided additional improvements to avoid biasing the results towards our evaluation set.

2. Using OPP-115 in the privacy policy compliance analysis: We use the existing OPP-115 dataset and map the labels to the most applicable requirement tuples. While we acknowledge that OPP-115 dataset (which was created in 2016) may not adequately capture new requirements, our current approach creates the mapping based on the observation of the classifier performance on new privacy policy text. For example, "Right to Object" is a new requirement in GDPR, which is mapped to "User Access, Edit, and Deletion". This is based on our observation that *segments* that describe how users may access, edit or delete data often also include *statements* that discuss user-specific rights. Hence, our approach includes these statements as candidate statements under this category because they have higher chances of fitting the requirement. Our results from *Full Match* and *Partial Match* category in Table IX demonstrates the efficacy of using this approach.

3. Simple tuple comparison heuristic in the privacy compliance analysis: To analyze privacy policy compliance, we compare tuple attributes using a simple heuristic. That is, for some tuple attributes, we simply check for the existence of attributes in tuples short-listed by our segment classifier. While our results show that even simple comparisons can help identify missing and matching cases with reasonable precision, future work can extend the approach by building systems that can analyze complex phrases.

4. Large Language Models (LLMS): While models such as ChatGPT [54] have demonstrated exceptional performance in a variety of NLP tasks, researchers are still grappling to understand the risks and reliability [11] of using these models in important use-cases, particularly as they may suffer from hallucination [10]. As privacy analysis is mission-critical, these models need to be properly evaluated before being used in an end-to-end system such as ARC. However, there are opportunities for underlying LLMs to be fine-tuned to focus on specific downstream tasks (e.g., use of few-shot learning [16] for text classification), which can be incorporated in our NLP pipeline. The fine-tuning on LLMs can leverage lessons from our existing contextualization of NLP techniques for regulations, whereas our labeled dataset can be used to evaluate the performance of these models.

XI. Conclusion

This paper presented the design and implementation of ARC, a framework that lays the foundation for systematic extraction, representation, and querying of privacy regulations. ARC benefits both security and privacy researchers by systematically extracting regulatory text and enabling the design and evaluation of practical privacy analyses or systems. Similarly, as ARC's tuples are machine consumable, it lays a foundation for organizations to automate their privacy compliance by connecting appropriate business operations with the tuple representation of rules. We evaluated ARC to demonstrate that it not only extracts tuples with considerable accuracy, but can also be effectively used for identifying similarities across regulations, or analyzing the compliance of privacy policies to multiple regulations.

Acknowledgment

The authors have been supported in part by the NSF-CNS-2237012. Any opinions, findings, and conclusions expressed herein are the authors' and do not reflect those of the sponsors.

References

[1] "How does nlp benefit legal system: A summary of legal artificial intelligence," https://github.com/ICLRandD/Blackstone.

[2] "Nucor privacy policy."

[3] "PrivGuard: Privacy regulation compliance made easier," in *31st USENIX Security Symposium (USENIX Security 22)*. Boston, MA: USENIX Association, 2022. [Online]. Available: https://www.usenix.org/conference/usenixsecurity22/presentation/wang-lun

[4] J. M. D. Alamo, D. S. Guaman, B. García, and A. Diez, "A systematic mapping study on automated analysis of privacy policies," 2021.

[5] AllState, "AllState Privacy Policy," https://www.allstate.com/privacy-center.

[6] Amazon, "Canada Amazon," https://amazon.ca/.

[7] B. Andow, S. Y. Mahmud, W. Wang, J. Whitaker, W. Enck, B. Reaves, K. Singh, and T. Xie, "PolicyLint: Investigating Internal Privacy Policy Contradictions on Google Play," in *Proceedings of the USENIX Security Symposium*, 2019.

[8] B. Andow, S. Y. Mahmud, J. Whitaker, W. Enck, B. Reaves, K. Singh, and S. Egelman, "Actions Speak Louder than Words: Entity-Sensitive Privacy Policy and Data Flow Analysis with PoliCheck," in *Proceedings of the USENIX Security Symposium*, 2020.

[9] N. Apthorpe, S. Varghese, and N. Feamster, "Evaluating the contextual integrity of privacy regulation: Parents' iot toy privacy norms versus coppa," 2019.

[10] Y. Bang, S. Cahyawijaya, N. Lee, W. Dai, D. Su, B. Wilie, H. Lovenia, Z. Ji, T. Yu, W. Chung, Q. V. Do, Y. Xu, and P. Fung, "A multitask, multilingual, multimodal evaluation of chatgpt on reasoning, hallucination, and interactivity," 2023.

[11] E. M. Bender, T. Gebru, A. McMillan-Major, and S. Shmitchell, "On the dangers of stochastic parrots: Can language models be too big?" 2021.

[12] BerTopic, "BerTopic Topic Modeling," https://maartengr.github.io/BERTopic/index.html, Accessed March 2023.

[13] J. Bhatia, T. D. Breaux, J. R. Reidenberg, and T. B. Norton, "A Theory of Vagueness and Privacy Risk Perception," in *Proceedings of the IEEE International Requirements Engineering Conference (RE)*, 2016.

[14] J. Bowers, B. Reaves, I. N. Sherman, P. Traynor, and K. Butler, "Regulators, Mount Up! Analysis of Privacy Policies for Mobile Money Services," in *Proceedings of the Symposium on Usable Privacy and Security (SOUPS)*, 2017.

[15] J. Bowers, I. N. Sherman, K. Butler, and P. Traynor, "Characterizing Security and Privacy Practices in Emerging Digital Credit Applications," in *Proceedings of the ACM Conference on Security and Privacy in Wireless and Mobile Networks (WiSec)*, 2019.

[16] T. B. Brown, B. Mann, N. Ryder, M. Subbiah, J. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, S. Agarwal, A. Herbert-Voss, G. Krueger, T. Henighan, R. Child, A. Ramesh, D. M. Ziegler, J. Wu, C. Winter, C. Hesse, M. Chen, E. Sigler, M. Litwin, S. Gray, B. Chess, J. Clark, C. Berner, S. McCandlish, A. Radford, I. Sutskever, and D. Amodei, "Language models are few-shot learners."

[17] D. Bui, K. G. Shin, J.-M. Choi, and J. Shin, "Automated extraction and presentation of data practices in privacy policies." 2021.

[18] D. Bui, Y. Yao, K. G. Shin, J.-M. Choi, and J. Shin, "Consistency analysis of data-usage purposes in mobile apps," in *Proceedings of the 2021 ACM SIGSAC Conference on Computer and Communications Security*, 2021.

[19] CCPA, "California Consumer Privacy Act," https://leginfo.legislature.ca.gov/faces/codes_displayText.xhtml?division=3.&part=4.&lawCode=CIV&title=1.81.5.

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

[21] Colorado, "Propbank Guideline," http://clear.colorado.edu/compsem/documents/propbank_guidelines.pdf.

[22] Coterra, "Coterra Website," https://coterra.com/who-we-are/.

[23] L. F. Cranor, P. G. Leon, and B. Ur, "A Large-Scale Evaluation of US Financial Institutions' Standardized Privacy Notices," *ACM Transactions on the Web*, 2016.

[24] DataGrail, "Gartner Data Subject Request," https://www.datagrail.io/blog/product/gartner-subject-rights-requests-2021/#:~:text=As%20highlighted%20in%20the%20market,%241%2C524%20in%20the%202021%20report, Accessed March 2023.

[25] Deontic Logic, "Deontic Logic," https://plato.stanford.edu/entries/logic-deontic/, Accessed March 2023.

[26] DigiChina, "Translated Privacy Regulation for China," https://digichina.stanford.edu/work/translation-personal-information-protection-law-of-the-peoples-republic-of-china-effective-nov-1-2021/, Accessed March 2023.

[27] M. d'Aquin, S. Kirrane, S. Villata, A. Oltramari, D. Piraviperumal, F. Schaub, S. Wilson, S. Cherivirala, T. Norton, N. Russell, P. Story, J. Reidenberg, N. Sadeh, M. d'Aquin, S. Kirrane, and S. Villata, "Privonto: A semantic framework for the analysis of privacy policies," 2018.

[28] V. Eidelman, "Billsum: A corpus for automatic summarization of us legislation," *Proceedings of the 2nd Workshop on New Frontiers in Summarization*, 2019. [Online]. Available: http://dx.doi.org/10.18653/v1/D19-5406

[29] EurLex Dataset, "EUR-LEX," http://www.ke.tu-darmstadt.de/resources/eurlex.

[30] B. Fawei, J. Z. Pan, M. Kollingbaum, and A. Z. Wyner, "A semi-automated ontology construction for legal question answering."

[31] Forbes, "Data Privacy is an Organization Wide Issue," https://www.forbes.com/sites/forbesbusinesscouncil/2021/02/24/data-privacy-isnt-just-an-it-issue-its-an-organization-wide-issue/?sh=b6847c927614, Accessed March 2023.

[32] GDPR, "General Data Protection Regulation," https://gdpr-info.eu/.

[33] H. Harkous, K. Fawaz, R. Lebret, F. Schaub, K. G. Shin, and K. Aberer, "Polisis: Automated Analysis and Presentation of Privacy Policies Using Deep Learning," in *Proceedings of the USENIX Security Symposium*, 2018.

[34] M. Honnibal and I. Montani, "spaCy 2: Natural Language Understanding with Bloom Embeddings, Convolutional Neural Networks, and Incremental Parsing," *To appear*, 2017. [Online]. Available: https://www.spacy.io

[35] HTMLToPlainText, "Convert HTML to Plaintext," https://github.com/benandow/HtmlToPlaintext.

[36] Humphreys, Llio, and Boella, Guido, and van der Torre, Leendert, and Robaldo, Livio, and Di Caro, Luigi, and Ghanavati, Sepideh, and Muthuri, Robert , "Populating legal ontologies using semantic role labeling."

[37] IApp, "US State Privacy Legislation Tracker," https://iapp.org/resources/article/us-state-privacy-legislation-tracker/, Accessed March 2023.

[38] ITIF, "The Looming Cost of a Patchwork of State Privacy Laws," https://itif.org/publications/2022/01/24/looming-cost-patchwork-state-privacy-laws/, Accessed March 2023.

[39] N. Kitaev and D. Klein, "Constituency parsing with a self-attentive encoder," 2018.

[40] Legal Rights, "Hohfeldian Analysis of Rights," https://plato.stanford.edu/entries/legal-rights/, Accessed March 2023.

[41] Legis Quebec, "Quebec Privacy Regulation," https://www.legisquebec.gouv.qc.ca/en/document/cs/p-39.1, Accessed March 2023.

[42] Lex Predict, "LexNLP," https://github.com/LexPredict/lexpredict-lexnlp, Accessed May 2022.

[43] LexPredict, "LexPredict," https://lexpredict.com.

[44] LGPD Brazil, "Translated Privacy Regulation for Brazil," https://www.lgpdbrasil.com.br/wp-content/uploads/2019/06/LGPD-english-version.pdf, Accessed March 2023.

[45] S. Manandhar, K. Kafle, B. Andow, K. Singh, and A. Nadkarni, "Smart home privacy policies demystified: A study of availability, content, and coverage," 2022.

[46] Nils Reimers and Iryna Gurevych, "Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks," *CoRR*, vol. abs/1908.10084.

[47] H. Nissenbaum, "Privacy as contextual integrity."

[48] M. V. Nortwick and C. Wilson, "Setting the bar low: Are websites complying with the minimum requirements of the ccpa?" 2022.

[49] S. O'Connor, R. Nurwono, A. Siebel, and E. Birrell, "(un)clear and (in)conspicuous: The right to opt-out of sale under ccpa," 2021.

[50] E. Okoyomon, N. Samarin, P. Wijesekera, A. E. B. On, N. Vallina-Rodriguez, I. Reyes, Álvaro Feal, and S. Egelman, "On the Ridiculousness of Notice and Consent: Contradictions in App Privacy Policies," in *Workshop on Technology and Consumer Protection (ConPro)*, 2019.

[51] Oliver Smitch, "GDPR Racket," https://www.forbes.com/sites/oliversmith/2018/05/02/the-gdpr-racket-whos-making-money-from-this-9bn-business-shakedown/?sh=34f82d2034a2, Accessed March 2023.

[52] Oliver Smith, "GDPR Compliance Cost," https://www.forbes.com/sites/oliversmith/2018/05/02/the-gdpr-racket-whos-making-money-from-this-9bn-business-shakedown/?sh=1b1d3dcb34a2, Accessed May 2022.

[53] Online Appendix, "Online Appendix," https://github.com/Secure-Platforms-Lab-W-M/ARC/blob/main/appendix/appendix-arc.pdf.

[54] OpenAI, "Chat GPT," https://chat.openai.com/, Accessed March 2023.

[55] PIPEDA, "Personal Information Protection and Electronic Documents Act," https://leginfo.legislature.ca.gov/faces/codes_displayText.xhtml?division=3.&part=4.&lawCode=CIV&title=1.81.5.

[56] S. Pradhan, A. Moschitti, N. Xue, O. Uryupina, and Y. Zhang, "CoNLL-2012 shared task: Modeling multilingual unrestricted coreference in OntoNotes."

[57] Probank, "Propbank Predicates," https://verbs.colorado.edu/propbank/framesets-english-aliases/refer, Accessed March 2023.

[58] T. A. Rahat, M. Long, and Y. Tian, "Is your policy compliant? a deep learning-based empirical study of privacy policies' compliance with gdpr," in *Proceedings of the 21st Workshop on Privacy in the Electronic Society*, 2022.

[59] Reimplementation of AllenNLP for Semantic Role Labeling, "Semantic Role Labeling," https://github.com/Riccorl/transformer-srl, Accessed May 2022.

[60] K. M. Sathyendra, S. Wilson, F. Schaub, S. Zimmeck, and N. Sadeh, "Identifying the Provision of Choices in Privacy Policy Text," in *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*, 2017.

[61] Scikit-Learn, "K-Means Clustering," https://scikit-learn.org/stable/modules/generated/sklearn.cluster.KMeans.html, Accessed March 2023.

[62] Secure Platforms Lab - William & Mary, “Data and Artifact for ARC,” https://github.com/Secure-Platforms-Lab-W-M/ARC, Accessed November 2023.

[63] P. Shi and J. Lin, “Simple bert models for relation extraction and semantic role labeling,” 2019.

[64] Y. Shvartzshnaider, N. Apthorpe, N. Feamster, and H. Nissenbaum, “Going against the (appropriate) flow: A contextual integrity approach to privacy policy analysis,” *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, vol. 7, no. 1, pp. 162–170, Oct. 2019. [Online]. Available: https://ojs.aaai.org/index.php/HCOMP/article/view/5266

[65] Stryker, “Stryker Privacy Policy,” https://www.stryker.com/us/en/legal/privacy.html.

[66] Thomson Reuters, “GDPR Report Compliance,” https://legalsolutions.thomsonreuters.co.uk/blog/wp-content/uploads/sites/14/2019/12/Thomson-Reuters-GDPR-Report.pdf, Accessed March 2023.

[67] R. Tibshirani, G. Walther, and T. Hastie, “Estimating the number of clusters in a data set via the gap statistic.”

[68] UNCTAD, “Privacy Regulation Dataset,” https://unctad.org/page/data-protection-and-privacy-legislation-worldwide.

[69] UNCTAD Website, “Privacy Regulation Worldwide,” https://unctad.org/page/data-protection-and-privacy-legislation-worldwide.

[70] C. Utz, M. Degeling, S. Fahl, F. Schaub, and T. Holz, “(un)informed consent: Studying gdpr consent notices in the field,” 2019.

[71] VCDPA, “Virginia Consumer Data Protection Act,” https://lis.virginia.gov/cgi-bin/legp604.exe?211+ful+SB1392.

[72] L. Wang, U. Khan, J. Near, Q. Pang, J. Subramanian, N. Somani, P. Gao, A. Low, and D. Song, “PrivGuard: Privacy regulation compliance made easier,” 2022.

[73] Wang, Shufan and Thompson, Laure and Iyyer, Mohit.

[74] L. Yu, X. Luo, X. Liu, and T. Zhang, “Can We Trust the Privacy Policies of Android Apps?” in *Proceedings of the IEEE/IFIP International Conference on Dependable Systems and Networks (DSN)*, 2016.

[75] H. Zhong, C. Xiao, C. Tu, T. Zhang, Z. Liu, and M. Sun, “How does nlp benefit legal system: A summary of legal artificial intelligence,” 2020.

[76] S. Zimmeck and K. Alicki, “Standardizing and implementing do not sell,” 2020.

APPENDIX

A. *Implementation Details of ARC*

We implement ARC by integrating multiple pluggable modules in the Spacy pipeline [34], after runnning HtmlToPlainText [35] on the HTML regulation document to obtain the plain text. HtmlToPlainText normalizes bullet points and unicode characters, resolving list objects to full sentences to retain the semantic context. To integrate the SRL model, we encode every verb token with semantic roles out of any of the 68 available semantic roles [21]. Moreover, we map arguments for the list of verbs based on their verb senses to extract the formalized tuple as shown in Table XI. We also integrate “benepar model” [39] in the pipeline for constituency parsing to identify subordinate clauses. Similarly, we also integrated adapted Named-Entity Recognition model to identify data objects [7].

For extraction, we filter requirement-specific statements by performing a lemma comparison with a set of verbs and deontic modals (as discussed in Figure I). Hence, we obtain Data Flow tuple by extracting statements where the *Data Flow verb* is associated with one of the deontic modals presented in Table I (e.g., in the phrase “a business shall collect”, the verb “collect” is associated with the modal “shall”). Similarly, we obtain the Definition tuple by identifying statements that contain *Definition_Verbs*, *Definiendum*, and *Definiens*. That is, we make sure that the definition tuple contain both *term* being defined and its *description*. For example, the statement “*personal data shall mean information about identifiable individual*” contains all three parameters and is represented as a Definition Tuple. Finally, we obtain the Right Tuple by identifying statements that contain *Right_Verb*. After the identification of the right verb, ARC searches for the word *‘right’* in the arg1 attribute of the SRL object. For example, in the phrase “*A consumer shall have the right to object*”, ARC first identifies the Right_Verb i.e., ‘have’, followed by identification of word ‘right’ in the in the arg1 attribute, which is then represented as a Right_Tuple. We save JSON object containing the semantic role attributes and clause information, along with ARC tuple for each regulation statement to enable querying and further analysis. Figure 13 in the online appendix [53] presents Spacy’s pipeline modified for ARC.

B. *Constituency Tree Parsing*

Figure 12 in the online appendix [53] represents simplified version of statement extracted from CCPA regulation. We obtain two separate child node subordinate clause (represented as “SBAR” in the diagram), instead of parent SBAR constituent. That is, we include “that collects information”, and “that sells that personal information” as clauses, instead of considering their parent clause, *i.e.*,, “A business that collects information and that sells that personal information”.

C. *Fine-tuning Phrase-BERT*

We build ARCBert using the same procedure as specified in the Phrase-BERT [73]. However, we adapt it in our context by using phrases extracted by semantic roles instead of constituency chunks. We extract top 60K phrases for legal context, and remaining 40K phrases from privacy regulation text. For each phrase, we curate the positive examples by identifying a statement that contains the phrase, and negative example by selecting a random statement that does not contain the phrase. ARCBert uses contrastive objective similar to Phrase-BERT [73] to fine-tune BERT. That is, for a given anchor phrase, the task is to bring semantically similar context i.e., positive context together, whereas the negative context is pushed apart.

D. *Multi Regulation Analysis - Similarity Scores*

Figure 11 demonstrates the similarity score across all four regulations as discussed in Section VI.

E. *Mapping Regulation Requirements to Privacy Taxonomy*

Table XII shows our mapping of requirements from four different regulations to the privacy taxonomy based on OPP-115 labels.

F. *Privacy Policy Segment Classifier*

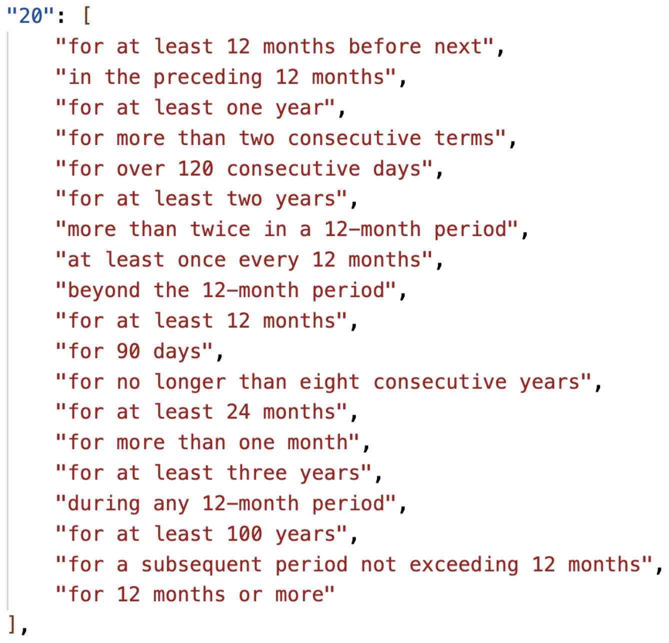
Table X shows the segment classification results for different labels based on OPP-115 dataset.

G. *Privacy Compliance Analysis*

Figure 9 shows the result from ARC. For a given privacy policy, ARC provides analysis based on individual regulation,

TABLE X: Classification Results for the Segment Classifier

Label	Precision	Recall	F1	Support
1st Party Collection	0.87	0.87	0.87	681
3rd Party Sharing	0.89	0.87	0.88	499
User Choice/Control	0.82	0.82	0.82	281
Data Security	0.90	0.82	0.85	183
Specific Audiences	0.96	0.89	0.92	169
User Access, Edit and Deletion	0.90	0.83	0.86	99
Policy Change	0.89	0.87	0.88	85
Data Retention	0.80	0.77	0.79	67
Do Not Track	1	0.97	0.98	15
Other	0.80	0.77	0.77	751
Average	0.88	0.84	0.86	



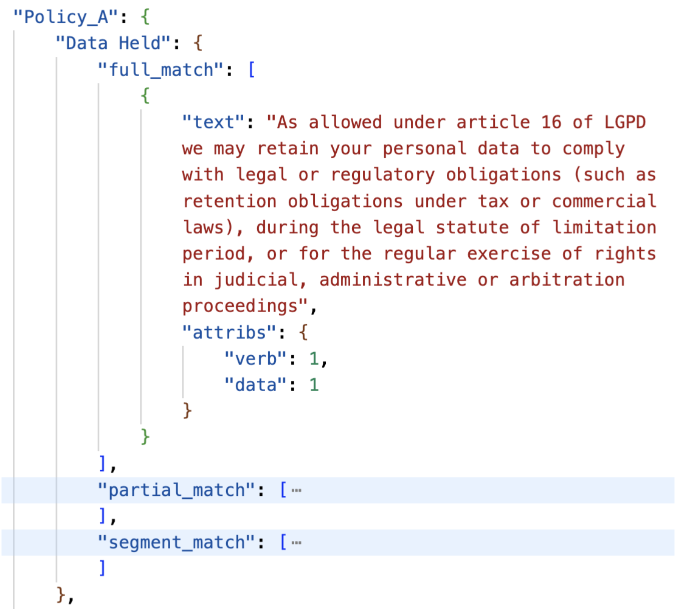
```
"20": [
    "for at least 12 months before next",
    "in the preceding 12 months",
    "for at least one year",
    "for more than two consecutive terms",
    "for over 120 consecutive days",
    "for at least two years",
    "more than twice in a 12–month period",
    "at least once every 12 months",
    "beyond the 12–month period",
    "for at least 12 months",
    "for 90 days",
    "for no longer than eight consecutive years",
    "for at least 24 months",
    "for more than one month",
    "for at least three years",
    "during any 12–month period",
    "for at least 100 years",
    "for a subsequent period not exceeding 12 months",
    "for 12 months or more"
],
```

Fig. 8: Example of Clustered Phrases for argm_tmp

short-listing all the policy statements within the categories such as full_match, and partial_match, while also providing the results for tuple attributes.

H. Missing Attributes in Partial Match Results

Figure 10 illustrates the number of policy statements and the missing attributes that were identified as a partial match.



```
"Policy_A": {
    "Data Held": {
        "full_match": [
            {
                "text": "As allowed under article 16 of LGPD
                we may retain your personal data to comply
                with legal or regulatory obligations (such as
                retention obligations under tax or commercial
                laws), during the legal statute of limitation
                period, or for the regular exercise of rights
                in judicial, administrative or arbitration
                proceedings",
                "attribs": {
                    "verb": 1,
                    "data": 1
                }
            }
        ],
        "partial_match": [⋯
        ],
        "segment_match": [⋯
        ]
    },
```

Fig. 9: Example of Full Match statement identified by ARC

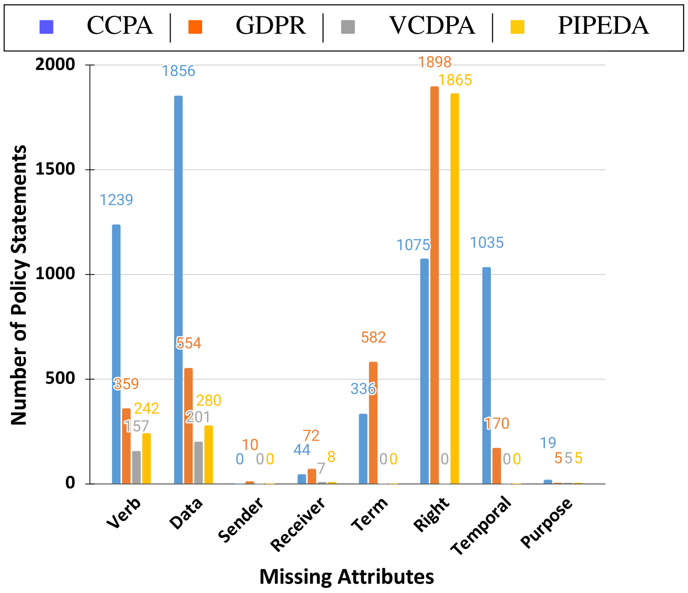


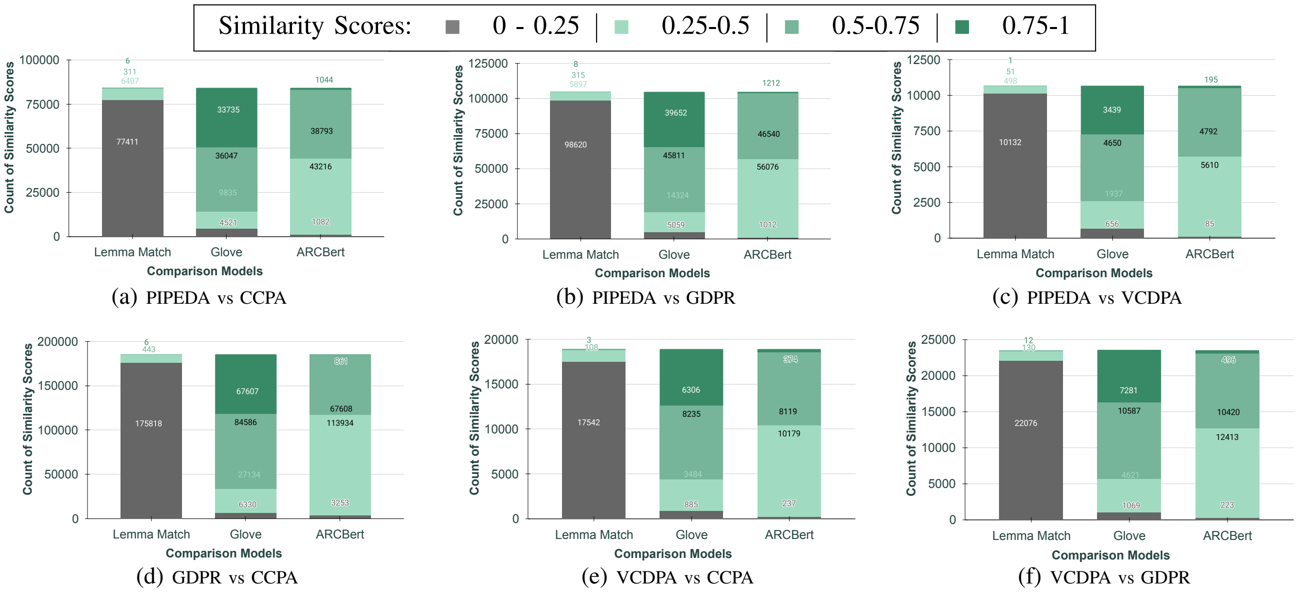
Fig. 10: Partial Match Attributes

TABLE XI: Mapping Arguments to Tuple Attributes

Verb	Sender	Receiver	Data	Purpose	Deontic	Temporal	Definiendum	Definiens	Right_Entity	Right_Phrase	Others
Collect	Arg2	Arg0	Arg1				-	-	-	-	Argm_loc or
Share	Arg0	Arg2	Arg1				-	-	-	-	Argm_ext or
Use	-	Arg0	Arg1	Argm_Prp			-	-	-	-	Argm_mnr or
Retain	-	Arg0	Arg1	or	Argm_Mod	Argm_Tmp	-	-	-	-	Argm_prd or
Process	-	Arg0	Arg1	Argm_Pnc			-	-	-	-	Argm_dir or
Delete	-	Arg0	Arg1				-	-	-	-	Argm_cau or
Include	-	-	-				Arg2	Arg1	Arg2	Arg1	sender_clause or
Mean	-	-	-				Arg0	Arg1	-	-	data_clause or
have	-	-	-				Arg0	Arg1	Arg0	Arg1	receiver_clause

TABLE XII: Regulation Requirements and Mapping to Privacy Taxonomy

CCPA	GDPR	VCDPA	CANADA	Privacy Taxonomy
-	Data Categories	Data Processed	Data Use	1st Party Collection/Use
Purpose of Collection	-	-	-	1st Party Collection/Use
-	Purpose of Processing	Purpose of Processing	Data Use Purpose	[1st Party Collection/Use, 3rd Party Sharing/Collection]
-	Source of Data	-	-	1st Party Collection/Use 3rd Party Sharing/Collection]
-	Profiling	-		1st Party Collection/Use
Data Collection 12 Months	-	-	-	1st Party Collection/Use
Third Party Disclose	Data Recipients	Third Party Share	Data Disclose	3rd Party Sharing/Collection
Third Party Disclose 12 Months	-	-	-	3rd Party Sharing/Collection
Sell Purpose	-	-	-	3rd Party Sharing/Collection
Third Party Sell	-	Third Party Sell	-	3rd Party Sharing/Collection
Third Party Sell 12 Months	-	-	-	1st Party Collection
-	Provision Requirement	-	-	User Choice/Control
-	Third Country Transfer	-	-	International and Specific Audiences
-	Safeguards Copy	-	-	Data Security
-	-	-	-	Policy Change
Data Retention	Storage Period	-	Data Held	Data Retention
Right to Know	Right to Access	Right to Access	-	User Access, Edit and Deletion
Right to Request Deletion	Right to Erase	Right to Delete	-	User Access, Edit and Deletion
Right to Opt-Out	Right to Object	-	-	User Access, Edit and Deletion
Right to Non-Discrimination	Right to Lodge Complaint	Right to Appeal	-	User Access, Edit and Deletion
-	Right to Data Portability	Right to Portability	-	User Access, Edit and Deletion
-	Right to Withdraw Consent	Right to Correct	-	User Access, Edit and Deletion
Opt-Out of Sale		Right to Opt-Out	-	[User Access, Edit and Deletion User Choice/Control]
Verifiable Consumer Request	-	-	-	Other
Request on Behalf	-	-	-	Other



(a) PIPEDA vs CCPA

(b) PIPEDA vs GDPR

(c) PIPEDA vs VCDPA

(d) GDPR vs CCPA

(e) VCDPA vs CCPA

(f) VCDPA vs GDPR

Fig. 11: Comparison of Similarity Scores for Phrases