PolicyPulse: Precision Semantic Role Extraction for Enhanced Privacy Policy Comprehension

Andrick Adhikari
University of Denver
andrick.adhikari@du.edu

Sanchari Das
University of Denver
sanchari.das@du.edu

Rinku Dewri
University of Denver
rinku.dewri@du.edu

***Abstract*—The effectiveness of natural language privacy policies continues to be clouded by concerns surrounding their readability, ambiguity, and accessibility. Despite multiple design alternatives proposed over the years, natural language policies are still the primary format for organizations to communicate privacy practices to users. Current NLP techniques are often drawn towards generating high-level overviews, or specialized towards a single aspect of consumer privacy communication; the flexibility to apply them for multiple tasks is missing. To this aid, we present PolicyPulse, an information extraction pipeline designed to process privacy policies into usable formats. PolicyPulse employs a specialized XLNet classifier, and leverages a BERT-based model for semantic role labeling to extract phrases from policy sentences, while maintaining the semantic relations between predicates and their arguments. Our classification model was trained on $13,946$ manually annotated semantic frames, and achieves a F1-score of 0.97 on identifying privacy practices communicated using clauses within a sentence. We emphasize the versatility of PolicyPulse through prototype applications to support requirement-driven policy presentations, question-answering systems, and privacy preference checking.**

I. Introduction

Privacy policies detail how user information is collected, used, secured, and shared by organizations [48], [74]. As such, they are vital documents that aid users in understanding data access and manage their privacy. However, the current state of privacy policies, characterized by their lengthy, complicated, and ambiguous nature, continues to pose challenges for users [39], [49], [29], [28]. Researchers have put forth alternative designs and recommendations aimed at making privacy policies more accessible. These include the use of short notices [31], [63], multi-layer policies [20], and graphical representations [32], [40]. However, these approaches often face challenges due to ambiguity and limited adoption [44], prompting researchers to explore Natural Language Processing (NLP) solutions that can effectively work with natural language policies [59].

NLP plays a crucial role in the privacy domain, with applications including paragraph classification [34] and clustering [53], [46], and keyword extraction [58]. However, tasks such as control choice detection and compliance checks often require labor-intensive efforts in creating task-specific annotated data corpora [55], [73], [14]. Despite the considerable attention policy text classification has received [59], [7], [34], [51], [50], particularly using the OPP-115 corpus [67], it is rarely combined with information extraction tasks. We identify this gap and aim to provide a platform that can enable granular information extraction from privacy policies that can help in achieving usable and comprehensible privacy policy design.

To achieve this goal, we introduce PolicyPulse, an approach that combines automated classification with transformer-based semantic information extraction. Unlike traditional methods that operate at the sentence level, PolicyPulse dissects natural language policies into granular components and labels privacy-relevant components based on their semantic role. At the same time, this approach preserves the semantic relationships between predicates and their respective arguments within a sentence. The knowledge base generated using PolicyPulse can be leveraged to automatically generate alternative policy designs, and create customized & usable summarization, among others, without necessitating additional involvement from policy authors. PolicyPulse applies a BERT (Bidirectional Encoder Representations from Transformers)-based model for semantic role labeling and utilizes two XLNet classifiers arranged serially to identify relevant sentence clauses. It then proceeds to encode generic language semantic roles into privacy-specific roles related to information collection, usage, sharing, retention, and user control & access. The granularity enables extraction of relationships between data types, collectors, and purposes, supplemented with rich annotations such as user-triggered actions, opt-in/opt-out, user control methods, location specifics, sharing terms & consequences, and data retention periods.

Through this work, we make the following significant contributions:

1. We present PolicyPulse which contextualizes English language privacy policies using semantic frames. It has achieved a F1-score of 0.97 for information categorization.

2. PolicyPulse has multifaceted usable applicability such as to develop applications for policy completeness checks, designing alternative presentations (short notice, nutrition labels), automated query answering, and user preference checking.

3. We manually annotate a corpus of $13,946$ semantic frames using which we provide frame-level classification of $129,856$

Network and Distributed System Security (NDSS) Symposium 2025
24-28 February 2025, San Diego, CA, USA
ISBN 979-8-9894372-8-3
https://dx.doi.org/10.14722/ndss.2025.240916
www.ndss-symposium.org

policies to analyze complexity in sentence composition.

4. We also provide a mapping from generic natural language predicate arguments to privacy-specific roles for 146 commonly used verbs in privacy policies, and five data practice categories. This mapping allows for a granular capture of relationships between actors, actions, purposes, triggers and consequences, often embedded in policies through the use of nuanced language semantics.

In the remainder of the paper, we present related work and background in Section II, highlighting gaps in the current state and positioning PolicyPulse as a solution aimed at addressing them. Through Section III, we present the methodology and evaluation process utilized to develop PolicyPulse. Section IV presents observed tendencies in privacy policies analyzed using PolicyPulse, followed by a discussion on potential applications in Section V. Finally, we present limitations with references to future work in Section VI and conclude in Section VII.

II. Background & Situating PolicyPulse

Privacy policies are legal documents primarily designed to convey user-centric information about an organization's usage and access to their data [74], and ensure compliance with relevant regulations [25]. Yet, privacy policies face several challenges that hinder the general public from effectively utilizing them. Primary among these obstacles are concerns regarding readability [66], [39], [49], [29], [28], ambiguity [57], [56], [45], and accessibility [39], [37], [33] of information. To address this, there has been prior applications of natural language processing (NLP) aimed at mitigating the challenges posed by privacy policies. NLP has found diverse and valuable applications in the privacy policy domain including information extraction [36], [9], content summarization [71], automated query answering [60], text classification [68], [47], [67], and text alignment [53], [46].

Intersection of NLP and privacy policies have seen substantial advancement in automated policy text classification [59], [7], [67]. These encompass domain-specific embeddings, neural networks, deep learning models, and transformer-based models [51], [50]. Automated classification is applied to obtain a high-level overview of a policy, as exemplified by Harkous et al. [34], who used classifiers to identify data types and purposes, and created graphical visualizations of such information. Despite the strengths, it primarily uses paragraph-level classification, which limits its ability to achieve the granularity needed for extracting phrases related to specific privacy policy roles and precise semantic matching.

In the effort to extract vital policy-specific information, Bhatia et al. developed a lexicon of personal information types by identifying noun phrase chunking patterns from 15 human-annotated privacy policies [15]. Similar approaches that focus on extracting data types, entities, and purposes, have resulted in tools such as PolicyLint [9], PoliCheck [10], and OVRseen [65]. These tools find utility in linking data types and entities, identifying contradictions within privacy policies, and aligning privacy policy statements with observed data collection practices. Similarly, Cui et al. utilized Named Entity Recognition (NER) to identify collected data, entities collecting information, served purposes, and subsumption relations, thus aggregating the information dispersed across a policy [26]. These extracted entities provide concise information, but needs detailed and accompanying textual descriptions. Hence, to enhance information extraction, we use Semantic Role Labeling (SRL) for parsing long and complex phrases. This approach includes details about user-triggered actions, user-controllable practices, data retention processes, and user data access levels.

For semantic analysis, Bhatia et al. manually coded semantic frames in 202 statements from five privacy policies, resulting in 17 semantic roles and 281 instances of data actions [16]. Their work was eventually extended with 15 manually annotated privacy policies [17]. While specialized approaches exist that use syntax-driven semantic analysis methods to construct partial ontologies, and context-free grammar for inferring semantic relations [35], deep learning and NLP can facilitate automated methods for improved and scalable extraction of semantic frame representations of policies, and enable large-scale analysis. Shvartzshnaider et al. proposed information extraction through semantic role labeling using domain-specific rule-based heuristics to include information for a predefined list of verb predicates [62]. In the NLP domain, Zhang et al. has demonstrated that utilizing keyword-based or syntactic patterns on SRL for querying can be used for argument extraction [72]. Among previous applications of SRL in the privacy domain, PurPliance [19] has utilized it to handle lengthy and complex phrases within purpose clauses. These predicates are mostly limited to first-party collection and use of data and rarely include information beyond the highly coupled application-specific requirements. By incorporating semantic analysis with frame classification and mapping predicate-specific arguments to broader domain-specific roles, we aim to further generalize policy processing with NLP and realize a platform that can support a wider array of tasks. This will make the platform more versatile and effective for various applications without requiring additional efforts in specialized corpus creation.

Bannihatti et al. developed a corpus of 236 website privacy policies to automate the extraction of opt-out statements [14]. Granular extraction of privacy-specific artifacts, however, can also be achieved without the need for specialized corpus creation. Such scalable corpus can improve other areas including information alignment [53], [46] and potentially enhancing the specificity of responses in query-answering systems [60]. Automated generation of usable policy formats using such a corpus will help improve wider adoption, which other machine-readable formats such as P3P [24] and similar initiatives [11], [13], [18], [38], [30] have struggled with due to complexities and scalability issues [23]. In such, alternative designs for privacy policy representation [31], [63], [20] can also benefit much from a real-time generation using such generic knowledge bases. Through PolicyPulse we offer a more granular characterization of policy texts with privacy

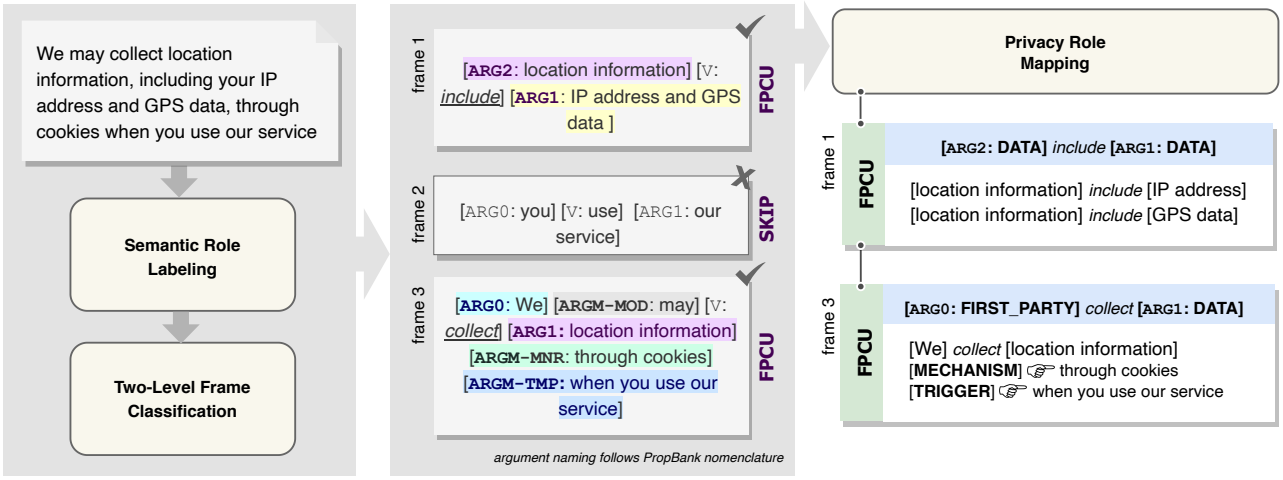

Fig. 1. Overview of privacy-specific role extraction in PolicyPulse

specific roles showing its effectiveness with scalable corpus. Alongside collector, data type, and purpose information, we include and associate details such as user-triggered actions, practices controllable through user controls and choices, the data retention process, and the access levels granted to users for their data.

III. Privacy Role Extraction

In PolicyPulse[1], we apply a BERT-based model to facilitate role extraction from *semantic frames*—cognitive structures designed to encapsulate distinct scenarios, situations, or concepts. We categorize the semantic frames using a two-level classification architecture. In the first level we determine the frame's privacy-specific relevance. If the frame is deemed relevant, it is then further categorized into a privacy practice category. Depending on this category and the action verb in the frame, semantic roles are extracted, which are then translated into *privacy roles*—privacy-specific data entities. An overview of this process is illustrated in Figure 1.

A. Semantic Role Labeling

SRL is a NLP task that involves identifying the roles that different words play in a sentence, or a clause, with respect to a specific predicate [43]. For example, in the clause *"We collect location information,"* SRL would identify 'We' as the agent (doer) of the 'collect' verb (action), and 'location information' as the object of the action. In this context, the clause is also referred to as a *frame*, and the identified components (doer and object in the example) are the *roles* (arguments) played by words/phrases in the frame with respect to the predicate's meaning. Role identification in SRL is centered around a predicate (mostly verb), which could be the main predicate of a sentence, or a secondary predicate appearing in a clause.

SRL Bert is a state-of-the-art semantic role labeling model [61] that leverages the BERT architecture, and capitalizes on the contextualized embeddings provided by BERT [45]. We utilized the trained model provided in the AllenNLP platform [1], without applying any modifications. SRL model training in AllenNLP is based on SRL annotations available in the English PropBank corpus [52]. The corpus offers semantic role annotation samples for roughly $50,000$ predicates ($1,118$ verbs), as well as role definitions for each of the predicates, denoted by numbered arguments, with each number corresponding to a distinct role based on predicate usage. While there may be nuances in PropBank roles depending on the verb and sentence context, the core definitions for each numbered argument can be found in Appendix B. SRL is pivotal to enhance the granularity of information extracted from a privacy policy. Using SRL Bert, we are able to extract the semantic frames, and the roles within, in a given privacy policy document. However, to limit the scope of analysis to privacy-specific categories, we prune the set of extracted frames before analyzing the roles in a frame. The pruning is performed using a XLNet classifier, as discussed next.

B. Frame Classification

We process each extracted frame from a policy document to identify if a frame relates to a distinct privacy practice category. For privacy practice categories, we use OPP-115, a corpus of 115 website policies annotated with 12 high-level data practice categories [67]. Please refer to Appendix A for brief descriptions of all the categories in the OPP-115 corpus. Among the 12 high-level categories in the OPP-115 corpus, we decided to categorize frames with 5 specific categories—First-Party Collection/Use (FPCU), Third-Party Sharing/Collection (TPSC), User Choice/Control (UCC), User Access, Edit, and Deletion (UAED), and Data Retention (DR). We collectively refer to these categories as the KEEP category. Our decision to prioritize the aforementioned 5 categories is rooted in the complexity of their language, often entailing multiple concepts within a sentence. This complexity necessitates a breakdown of information and categorization of extracted granular details

[1]The data from this project is made available at https://github.com/crisp-du/ppevo

to ensure direct alignment with the fundamental aspects of privacy practices.

We excluded the Do Not Track (DNT), Data Security (DS), Policy Change (PC), Privacy Contact Information (PCI), and International Specific/Audience (ISA) categories. In our initial analysis, we found that these categories represent a small portion of a policy, with concise sentences providing complete information. On average, PC and PCI account for less than 3%, DNT less than 0.3%, and DS and ISA less than 5% of sentences. Given their small presence, using our information extraction pipeline, which aims to extract concise relevant information from large volumes of text, would be redundant. Omitted categories such as DNT and DS often convey information in brief sentences with simpler language. PolicyPulse's in-depth extraction may be perceived as unnecessary for such cases, as the information is readily apparent. Additionally, Introductory/Generic (IG) categories are excluded due to their broad nature and lack of clear relevance to privacy concepts. These categories are more grounded in their inclination towards more organization-centric facets and encompass topics, which, while significant, do not depict the core essence of user-centric privacy concerns [67]. The chosen five categories comprehensively encompass aspects of data collection, usage, sharing, and control. The language used in policy paragraphs often blurs the boundaries between these practices, making the semantic frame-level analysis crucial. PolicyPulse excels in this regard by providing granularity, facilitating the disambiguation of distinctions that might be challenging for other models like Polisis.

1) Frame Category Annotation: In order to effectively train and evaluate a frame classification model, it is imperative that we have access to sample frame annotations. We utilized SRL Bert [61] on each of the $10,717$ policy sentences extracted from the OPP-115 corpus to generate $48,783$ semantic frames. As roles within each frame are governed by their association to a specific verb, we organized the frames for annotation by grouping them according to the associated verb. We identified a total of 929 unique verbs within the frames from the corpus. Following a manual verification process, we determined that 146 of these verbs are associated with the 5 privacy practice categories. Through elimination of unrelated verb frames, we arrived at a final count of $13,946$ frames for annotation. We excluded verbs like "can," "be," etc., which do not convey actions describing specific privacy concepts. The rationale for focusing on privacy-specific actions was to map the generic PropBank argument definitions to privacy-specific roles, as discussed in Section III-C. In addition to the selected five categories, we also introduced a SKIP category to eliminate noise at a granular semantic level encompassing:

1. Frames belonging to excluded OPP-115 categories, namely DNT, DS, PC, PCI, ISA, and IG. As detailed in Section III-B, thorough extraction from these categories is deemed unnecessary for our use case. Among $13,946$ frames manually annotated by us, only 0.029% are DNT, 0.366% DS, 0.272% PC, 0.165% PCI, 0.968% ISA, 0.179% IG, all of which are marked as SKIP.

2. Frames that are deemed incomplete, ambiguous, or contain incoherent information are marked as SKIP. For instance, a semantic frame like *'[ARG0:We][V:collect]'* is considered incomplete as it lacks a subject and is marked SKIP. Approximately 72.60% of frames fall into this category. This indicates that only a small portion of the semantic frames in a policy, relative to the total number of frames, contains information relevant to FPCU, TPSC, UCC, UAED, and DR. However, a high percentage of SKIP frames does not mean we discard the information entirely, as a sentence can contain multiple frames, some of which may capture relevant information. For example, consider the sentence: "*If you have granted us access to your Facebook or Twitter account by linking, you can disconnect the link by logging into your account, accessing your profile and clicking the Facebook or Twitter icon to disconnect.*" This sentence contains six semantic frames, one for each verb. However, not all frames are complete; for instance, *'[V:disconnect]'* or *'[V:linking]'* lack linked roles. Additionally, a frame like '*[V:clicking] [ARG0:the Facebook or Twitter icon] [ARGM-PRP: to disconnect]'* might be considered incoherent. Hence, apart from one valid frame, *'([ARGM-ADV: If you have granted us access to your Facebook or Twitter account by linking,] [ARG0: you] [ARGM-MOD: can] [V: disconnect] [ARG1: the link] [ARGM-PRP: by logging into your account, accessing your profile and clicking the Facebook or Twitter icon to disconnect.])'*, all other frames are marked as SKIP in this instance to reduce noise, and capture relevant semantic information. About 32% of all the sentences only consist of SKIP frames, which can be considered entirely made of irrelevant frames and can be filtered out.

We labeled the frames with either one of the designated five categories, or SKIP. The frames were annotated by a trained researcher with over 3 years of experience in annotating privacy policies and verified by two additional annotators with more than 5 years of experience each. This annotation produced the following distribution of frequencies across the various categories: SKIP with $10,401$ instances, FPCU with $1,417$ instances, UCC with 556 instances, TPSC with $1,230$ instances, DR with 160 instances, and UAED with 182 instances. The high frequency of the SKIP category suggests that relatively only a small portion of clauses within policy texts convey details related to primary privacy practice categories, which we aim to extract.

2) Model Training: We chose XLNet [70] as the foundation for our automated frame classifier. Our decision to use XLNet was based on preliminary experiments where it outperformed BERT-based models and traditional classifiers in paragraph, segment, and sentence classification tasks, particularly excelling in frame classification. Our preliminary work showed that a one-layer XLNet sentence classifier covering all categories achieved a 93% precision and 95% recall [4], [5]. Additionally, XLNet has demonstrated exceptional performance compared to other models like BERT, particularly with respect to privacy policies [50], [51], [4], [5]. We explored six different approaches to build the classifier, four of which differ in the structure of training input, while the other two

	Frame			Frame with sentence context (FSC)			FSC with custom attn. mask			FSC augmented			FSC augmented (ensemble)			FSC augmented (two-level)		
Category	pr	re	f1	pr	re	f1	pr	re	f1	pr	re	f1	pr	re	f1	pr	re	f1
SKIP	0.96	0.90	0.93	0.94	0.93	0.93	0.85	0.95	0.89	0.95	0.90	0.92	0.95	0.92	0.93	**0.99**	**0.98**	**0.98**
FPCU	0.67	0.83	0.72	0.56	0.64	0.56	0.29	0.43	0.33	0.70	0.76	0.72	0.78	0.82	0.79	**0.92**	**0.95**	**0.93**
TPSC	0.70	0.77	0.70	0.59	0.71	0.59	0.38	0.36	0.34	0.70	0.79	0.74	0.75	0.84	0.79	**0.92**	**0.94**	**0.93**
UAED	0.36	0.10	0.15	0.37	0.18	0.20	0.22	0.18	0.16	0.75	0.97	0.84	0.62	0.70	0.65	**0.93**	**0.93**	**0.92**
UCC	0.74	0.79	0.74	0.52	0.61	0.56	0.31	0.30	0.29	0.78	0.96	0.86	0.74	0.80	0.76	**0.93**	**0.93**	**0.93**
DR	0.25	0.23	0.24	0.50	0.33	0.32	0.39	0.18	0.23	0.71	0.99	0.83	0.64	0.68	0.65	**0.89**	**0.95**	**0.92**
macro avg	0.61	0.61	0.58	0.58	0.57	0.53	0.41	0.40	0.37	0.77	0.89	0.82	0.74	0.79	0.76	**0.93**	**0.95**	**0.94**
weighted avg	0.88	0.86	0.86	0.84	0.86	0.84	0.72	0.80	0.74	0.89	0.88	0.88	0.90	0.89	0.89	**0.97**	**0.97**	**0.97**

TABLE I

XLNET NESTED CROSS-VALIDATION PERFORMANCE SCORES FOR DIFFERENT FRAME CLASSIFICATION TRAINING METHODS. PR: PRECISION, RE: RECALL, F1: F1-SCORE

uses two instances of XLNet. The evaluations employed a 10-fold cross-validation approach, utilizing a 9:1 train:test split ratio. The training and test data sets were maintained consistently across all methods. For each configuration, the XLNet model was trained for 6 epochs, utilizing a random 9:1 train:validation split on the training data set. We chose a 6 epoch training duration as, beyond this, validation loss rises while training loss still decreases, suggesting overfitting to training data. Following 6 epochs per fold, the optimal model (selected via validation loss) was tested, providing 10 performance estimates that we average for a final value.

3) Performance Evaluation: The average precision (pr), recall (re), and F1-score (f1) of all the methods are shown in Table I. The initial approach used only the frames' text as input for model training ('Frame' column of Table I). While achieving a strong F1-score of 0.93 for the SKIP category, performance was notably low for other categories. Particularly, the UAED and DR categories exhibited F1-scores of only 0.15 and 0.24, respectively. Sentence context can significantly influence the frame category. Therefore, we next explored two methods to introduce such contexts in the model–(i) 'Frame with sentence context (FSC)': frame text combined with the sentence, and (ii) 'FSC with custom attn. mask': same as the previous but with addition of a custom attention mask for the XLNet input layer. The attention mask assigns a value of 0 to sentence phrases that are not present in the frame. Both methods show improved precision for the DR category, but there is an overall precision drop in all categories for method (ii). Method (i) also did not yield much performance improvement.

Variation in category performance shows influence of instance frequency during training. This is particularly apparent in the consistent superior precision and recall of the SKIP category, which has the highest instance frequency. To tackle this, we leveraged the capabilities of the textaugment [3] library to generate various versions of training instances for the infrequently occurring categories of UCC, UAED, and DR. This involved both synonym replacement and context-based word substitution. We did upheld consistency in performance comparison by preserving the integrity of the testing set (no synthetic data). 'FSC augmented' demonstrates XLNet's performance when trained on frame text alongside sentence context, with synthetically generated training instances for low frequency categories. This approach proved to be effective in distinguishing between closely related frame categories such as UCC and UAED, or FPCU and TPSC, where differentiation is often challenging. Notably, precision and recall also remained consistent regardless of the category.

Category	**pr**	**re**	**f1**
LEVEL 1			
SKIP	0.99	0.98	0.98
KEEP	0.93	0.96	0.94
macro avg	0.96	0.97	0.96
weighted avg	0.97	0.97	0.97
LEVEL 2			
FPCU	0.98	0.98	0.98
TPSC	0.99	0.98	0.98
UAED	0.98	0.98	0.98
UCC	0.98	0.98	0.98
DR	0.96	0.97	0.96
macro avg	0.98	0.98	0.98
weighted avg	0.98	0.98	0.98

TABLE II

XLNET NESTED CROSS-VALIDATION PERFORMANCE SCORES FOR LEVEL 1: SKIP-KEEP BINARY CLASSIFICATION AND LEVEL 2: FPCU, TPSC, UCC, UAED, AND DR CLASSIFICATION WITHIN KEEP FRAMES. PR: PRECISION, RE: RECALL, F1: F1-SCORE

The disparity in the F1-scores, especially between SKIP frame identification and the rest (KEEP category), indicates the need to better discriminate between them. This prompted us to explore a two-level approach. The first level performed a binary classification to identify SKIP vs. KEEP frames, and the second level predicted the practice type from the five selected categories for KEEP frames. The two levels are realized as two XLNet models, both models trained to classify frames with sentence context. To ensure balance in instance counts, the low-frequency categories were augmented to match the instance count of the highest-frequency category.

Table II shows the independent performance of the first and second levels. The division between SKIP classification and practice type classification improved the overall performance. The second level attained a macro average F1-score of 0.98, while the first level achieved a macro average F1-score of 0.96. The first level has a slightly lower performance when identifying KEEP frames in comparison to SKIP. This decline is reflected in the recall for non-SKIP categories when the two

Role	Description
DATA	Information about individuals, users, or subjects, which may include personal, sensitive, or non-personal information
FIRST_PARTY_ENTITY	The organization directly interacting with users, responsible for data collection and use
THIRD_PARTY_ENTITY	An external organization or entity in collaboration with the first-party entity
MECHANISM	The procedures and methods used by first-party or third-party entities to gather user data
PURPOSE	The specific reason for data collection, use and sharing
SHARING_TERMS	The conditions and agreements that govern the sharing of user data with third parties
USER_TRIGGER	Actions or events by users that may trigger data collection, use, or sharing
OPT_IN_MECHANISM	A method or process by which a user can provide explicit consent for the collection or use of their data
OPT_OUT_MECHANISM	A method or process by which a user can choose to decline or stop the collection or use of their data
CONSEQUENCE	The potential outcomes or effects of user choices or actions
USER_MECHANISM	The methods and processes available to users for accessing, editing, and deleting their data
USER_OPERATION	The actions permitted for users to control their data, including access, edit, delete, or management of subsets
RETENTION_PROCESS	The procedures and methods followed by the organization for the storage and management of user data over time
RETENTION_TERMS	The specified terms and conditions that dictate how long user data will be retained by the organization
TIME_PERIOD	A defined duration specifying how long user data will be retained by the organization
LOCATION	The physical or digital location of a particular action; the action can be collection, use, or sharing of data

TABLE III
PRIVACY-SPECIFIC ROLES MAPPED FOR FRAME ARGUMENTS IDENTIFIED BY SEMANTIC ROLE LABELING

models are sequentially used, shown in the 'FSC augmented (two-level)' column of Table I. Despite the reduction, this method yielded the highest performance, with a weighted average F1-score of 0.97, along with consistent precision and recall above 0.92 for all categories, except for DR with a precision of 0.89.

We also attempted to merge the two models into a single model. This involved adding an extra hidden layer in XLNet that collects the output from both models' hidden layers, and calculating their average. However, this resulted in imbalanced performance again, evident in the 'FSC augmented (ensemble)' column of Table I. Based on this, we opted for the two-level architecture as the method of choice. In this configuration, the first model decides between SKIP or KEEP. If KEEP is predicted, the practice type is determined by the second model. We retrained each model using the entire data set, including augmentation for low-frequency categories.

C. Mapping Privacy Specific Roles

In the next phase of our method, we mapped the generic PropBank semantic frame argument definitions to privacy-specific roles. During the frame annotation process, we identified 146 verbs that relate to relevant practice definitions. We examined the verbs and their PropBank arguments, and for each verb, we analyzed the roles that each argument assumes within the context of a frame's category. We then proceeded to rename these arguments with a privacy-specific nomenclature, considering only the frame categories observed for a particular verb during frame category annotations. For instance, the verb 'share' has frames labeled as either FPCU or TPSC. Consequently, we created two distinct sets of roles for the PropBank arguments associated with the 'share' verb, one for FPCU and another for TPSC. The mapping is created through manual evaluation of each semantic frame for the 146 verbs. For example, in an FPCU frame like '*[ARG0: The Company] [V: collects] [ARG1: personal information] [ARGM-MNR: through cookies]*' for the verb 'collect', we observe that ARG0 maps to FIRST_PARTY_ENTITY, ARG1 to DATA, and ARGM-MNR to MECHANISM. If ARG0 represents FIRST_PARTY_ENTITY in most FPCU frames with the 'collect' verb, then we include an entry mapping ARG0 to FIRST_PARTY_ENTITY for the 'collect' FPCU category. Similar observations are made for each frame, and based on predominant trends across verb categories, we compile a Propbank to privacy-specific roles map (Appendix B). Table III shows the privacy-specific role labels we have created for the arguments. Each of the 146 predicate verbs is associated with a subset of these roles depending on the category assigned to a semantic frame.

IV. FRAME SPECIFIC POLICY COMPOSITIONS

We applied our methodology to policies in the Princeton Privacy Crawl (PPCrawl) corpus, a collection of $1,071,488$ English language privacy policies from $130,604$ different websites, ranging from 1997 to 2019 [8]. We aim to derive granular generalized and representative insights from our analysis, and thus chose the most recent policies in PPCrawl for each of the $130,604$ organizations. We utilized AllenNLP's SRL Bert on each sentence within every policy to generate the semantic frames. AllenNLP failed to generate frames for 748 policies due to library errors when encoding the text for those select policies. Consequently, our analysis proceeded with a data set comprising of $129,856$ policies. We then use our frame classification model to categorize the frames as either KEEP or SKIP frames which was followed by categorizing the KEEP frames into one of the five selected privacy practice categories. Recall that SKIP does not mean the entire policy sentence is discarded; rather, only the part of the sentence irrelevant to the chosen primary categories is skipped. After classification, we employed our verb-dependent privacy-specific role mapping to extract phrases related to privacy roles. We present below our observations from the analysis conducted on the selected policies at each phase of the information extraction pipeline.

Sentence Packaging: We note that the $129,856$ policies have a total of $8,964,331$ sentences, with an average of nearly 70 sentences in a policy. SRL Bert extracted $39,702,767$

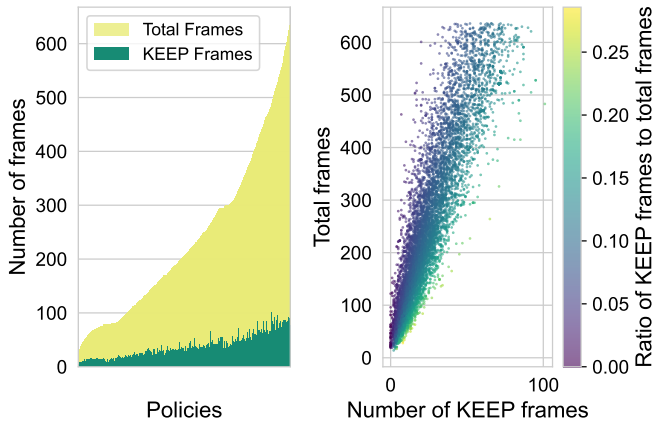

Fig. 2. Left: number of total and KEEP frames in the policies, and Right: Total number of frames against KEEP frames across all the policies; frame count capped at the 90^{th} percentile value

semantic frames from these sentences. We observed an average of ≈ 305.744 semantic frames per policy, with 25% of policies having over 403 frames, and some reaching up to $6,148$ frames. Semantic frames are mental structures aiding language comprehension. When a text contains numerous frames, it raises the cognitive load for readers, which can contribute to comprehension difficulties often encountered with privacy policies. This is difficult to mitigate in traditional policies due to the lack of mechanisms for filtering information based on preferences. We observed that, on average, the ratio of the number of semantic frames to the number of sentences in a policy is $4:1$. However, this ratio can also reach higher values, such as $14:1$ observed in the case of *clothing-dropship.com* in the year 2016. This implies that readers, on average, encounter four times more frames than sentences. A 75^{th} percentile value of 4.8 suggests that this holds true for most website policies.

Privacy-specific Relevance: We observed that our classifier identifies less than 82 KEEP semantic frames for 95% of the policies. Relative to the total number of frames, number of KEEP semantic frames does not increase correspondingly. On average, we observed 30 KEEP semantic frames per policy as compared to an average of $\approx$ 306 total frames per policy. Out of a total of $39,702,767$ semantic frames, merely $3,969,914$ frames are not discarded. We infer from these observations that a minimal portion of the semantic information in a policy suffices for extracting information related to the roles described in Section III-C. Figure 2 (left) shows the number of total frames and KEEP frames in policies, sorted by total frames on the horizontal axis. We also observed that the total number of frames in a policy does not influence the total number of KEEP frames. Figure 2 (right) plots, for each policy, the number of KEEP frames against the total number of frames. The color map represents the ratio of the number of KEEP frames to the total number of frames. The color map and plot shows that for higher number of total frames, the ratio of identified KEEP frames decreases as

Category	Mean	Std. Dev.	Q1	Q2 (Median)	Q3
FPCU	47.989	16.566	38.235	48.889	57.143
TPSC	32.549	16.654	22.034	31.579	40.625
UCC	12.604	9.344	6.667	12.121	17.778
UAED	2.270	5.203	0.000	0.000	3.030
DR	4.589	6.541	0.000	2.041	7.692

TABLE IV
PRACTICE TYPE COMPOSITION STATISTICS FOR KEEP FRAMES

the total number of semantic frames in the policy increases. Relatively, policies with lower number of semantic frames have a higher proportion of frames retained. Thus, practice-specific information is dispersed over a relatively smaller portion of a policy's text.

Categorical Composition: The next phase of our analysis is to investigate the trend of frame practice type composition. For each policy, we computed the percentage of each practice type category among all the KEEP frames in the policy. Table IV shows that, on average, nearly 48% of the KEEP frames are FPCU, followed by TPSC with an average of 33%. The 75^{th} percentile value (Q3) for FPCU implies that nearly a quarter of the policies have more than 57.143% of their KEEP frames labeled as FPCU. Relative to other categories, FPCU and TPSC information have higher presence in terms of semantic information. We observed that 25% of the policies have more than 18% of their KEEP frames labeled as UCC. Additionally, the quartile values for UCC frames indicate that policies communicate user control and choice information to users in some capacity. However, DR and UAED frames are missing in 25% of the KEEP frames. The median for UAED indicates that access, edit, and delete semantic frames are absent in 50% of the policies. The rest of the policies have a relatively low portion of UAED frames, indicating this category to be the most under-addressed at the lowest granular composition level of policies.

Categorical Semantic Correlation: Sentences featuring frames within a consistent practice category foster a cohesive context, whereas transitions between distinct practice categories can prompt contextual shifts, potentially posing challenges to users. Using the second-level of our frame classifier, we found that 66.25% of sentences have frames from one practice type, 28% have two, 5.11% have three, and the rest include 4 to 5 practice types. We investigated the relation between different categories of semantic frames with a co-occurrence matrix. We selected sentences with two or more frames with unique categories, and computed a co-occurrence matrix. To normalize the co-occurrence matrix, we divided each column by the corresponding diagonal element, which represents the number of frames for each category. The resulting matrix is presented in Figure 3, where each value indicates the proportion of a given category (vertical labels) co-occurring with other frame categories (horizontal labels).

The FPCU column indicates that all other categories rely heavily on FPCU semantics to form the context of their communication. While this reliance may seem necessary, it can also lead to oversight due to information density. For instance,

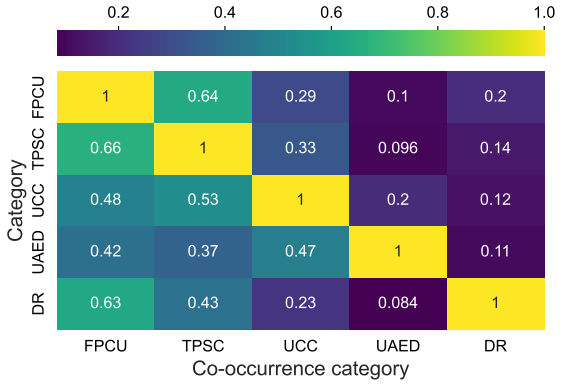

Fig. 3. Co-occurrence proportion of a given category (vertical labels) with other frame categories (horizontal labels)

among sentences with more than one unique semantic frame category, 63% of DR frames and 66% of TPSC frames co-occur with FPCU frames. This co-occurrence may result in overlooking statements about data storage or sharing, as they are surrounded by first-party collection. Ambiguity is another potential drawback of such grouping, especially for high co-relation categories such as TPSC and FPCU, and, UAED and UCC. 47% of UAED frames share a sentence with UCC frames, which may create ambiguity for users to distinguish between opt out controls, and mechanisms for data deletion.

V. Potential Applications

A fine-grained information extraction pipeline on a privacy policy can facilitate multiple use cases towards automated analysis and presentation of such policies. In this section, we consider few such use cases to demonstrate the generic applicability of the proposed approach. Most of these use cases are motivated from past works in usable privacy policies. We demonstrate how the semantic roles extracted from a policy can be subjected to a few additional processing steps, thereby serving as the building block for such applications. We note that the discussion below is presented to exemplify potential utility of our SRL-based approach, and does not delve deeply into implementation-specific challenges.

A. Policy Completeness

A privacy policy is expected to convey information along multiple dimensions, ranging from data collection and sharing, to user rights and regulatory compliance. However, most policy writers, intentionally or otherwise, produce a policy that is incomplete in terms of the expected coverage of information. To illustrate this, we focus on the number of policies lacking specific roles that should ideally be communicated in the context of the frame's category. For instance, a FPCU frame is expected to contain phrases with the PURPOSE role, among others. Figure 4 displays the percentage of policies missing roles for each practice category and shows that most policies communicate DATA-related information in the context of FPCU, TPSC, and DR. However, a significant portion of DR frames lack information with respect to the terms of retention (RETENTION_TERMS), period of retention (TIME_PERIOD), and process of retention (RETENTION_PROCESS). In nearly 50% of the policies, UAED semantic frames often do not explicitly mention the DATA accessible and editable by users. Approximately 60% of the policies fail to convey the USER_MECHANISM for access, edit, and deletion, and 90% of the policies do not specify the LOCATION associated with the USER_MECHANISM.

Both FPCU and TPSC frames often contain PURPOSE information, but policies frequently fail to communicate the SHARING_TERMS associated with third-party information sharing. The MECHANISM for both first-party and third-party collection is absent in nearly 40% of the policies. Furthermore, approximately 20% of the policies do not explicitly mention the USER_TRIGGER, which represents user actions treated as consent for data collection and use. User choice and control related KEEP frames are observed to be quite efficient in communicating the OPT_OUT_MECHANISM, with around 20% of the policies lacking this information. However, nearly 75% of the policies do not specify the CONSEQUENCES of either opting in or out. Consequently, they fail to elaborate on the benefits or disadvantages that a user may experience when utilizing the provided controls. A SRL-based analyzer such as PolicyPulse can aid a policy writer in highlighting these gaps, and provide automated feedback on writing a more complete privacy policy.

B. Requirement Driven Policy Presentation

Summarization methods can condense a privacy policy to critical points and encapsulate the most important content. In past works, the process of summarizing policies begins by formulating a set of inquiries that the summary aims to address. This can be achieved through either a graph-based approach, which extracts specific entities with named-entity recognition (NER) and syntactic parsing, such as collector, data type, their relationships, and purposes [26]; or by employing a classification-based approach to assess risk levels [71]; or categorize policy texts using low-level attribute class labels from OPP-115 [34]. These approaches are shaped by the objective and may require additional effort to meet a concerned party's specific requirements. In the following sections, we present methods and observations gained from our efforts to generate variations of policy presentation from semantic frames.

1) Short Notice:: The Federal Trade Commission (FTC) and various organizations have supported the idea of making privacy notices more concise and clear [21]. Short notices are seen as an effective means to convey essential information to consumers without overwhelming them [31], [63]. We employ our categorized semantic frames to automatically generate a brief summary from the complete text of a privacy policy.

Method: For short policy format, we emphasize information collection, sharing, selling, and storage practices, inspired by the design used by Gluck et al. [31]. We extract DATA phrases from FPCU frames for information collection, utilize

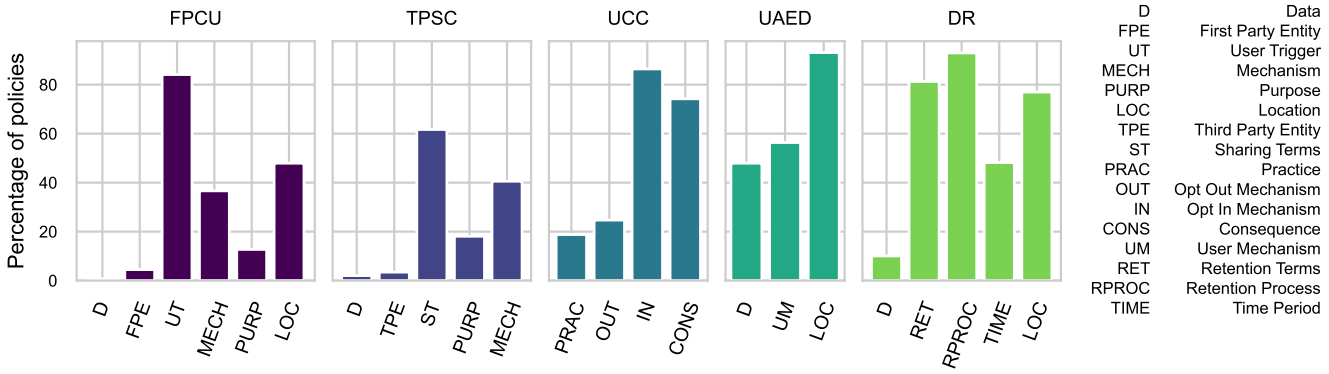

Fig. 4. Percentage of policies with missing roles in a practice category for KEEP frames

THIRD_PARTY_ENTITY role phrases from TPSC frames for third party collector, and gather retention information from DR frames. The extracted information is then presented in accordance with the short notice format.

Result: Figure 5 presents phrases from Yahoo's 2018 privacy policy that describes collected information, third-party entities with whom information is shared, and the retention policy. Through a quick read, a user can get acquainted with these practices. Additionally, if we incorporate the UCC and UAED frames, we can communicate relevant information to the user about the option to edit or delete information and preferences for marketing. The category and granularity of the role label aids in extracting and presenting this information. Since the extracted information remains unaltered, it can be readily aggregated into various forms, including normalization using regular expressions (as proposed by Cui et al. for PoliGraph [26]), or by employing keyword extraction libraries such as pke [2]. The short notice can also integrate deeper exploration of specific sections in the notice by leveraging the links retained by PolicyPulse between the extracted information and statements in a policy. For example, Figure 5 shows an expanded notice related to data sharing with non-affiliated companies. The expanded information can offer additional details to the user when needed, overcoming the limitations of static short notices where important information might be excluded [31]. Additionally, highlighting role labels within sentences can provide visual assistance. In Figure 5, we highlight each phrase for each role uniquely in the expanded information. This allows users to quickly understand that personal information will be shared if the condition (USER_TRIGGER role) is met.

2) Nutrition Label:: Privacy nutrition labels are an alternate policy format that tabulate information on data collection, usage, and sharing [40], [42]. Previous research has highlighted several benefits of privacy nutrition labels for users, including faster access to privacy information and improved understanding of an app's privacy practices [41]. Despite their advantages, the adoption of privacy labels faces notable challenges. One among them is the added complexity involved in keeping a

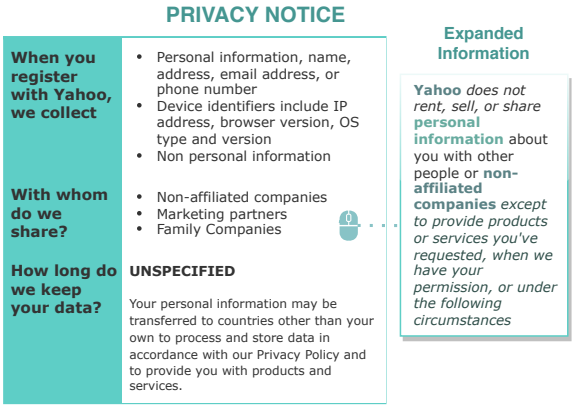

Fig. 5. Short notice summary generated from semantic frames on Yahoo's 2018 privacy policy

nutrition label up to date with policy revisions [44]. To that aid, we utilize our categorized semantic frames for nutrition label generation and show the feasibility of an automated approach.

Method: We formulated this task to align with a design inspired by Kelley et al.'s proposed 'Privacy Nutrition Labels' [40]. To gather information regarding the types of data collected, how it is shared, and with whom it is shared, we extract DATA phrases from FPCU and TPSC frames, along with PURPOSE. We also extract THIRD_PARTY_ENTITY phrases from TPSC frames. We include TRIGGER information for the extracted information, if present, to indicate that potential user actions may trigger a given practice.

Additionally, we trained two XLNet classifiers: one for data type classification and another for purpose type classification. These classifiers were trained using low-level attribute classes and annotated parts from sentences in OPP-115 [67]. Since our role mapping provides phrases from sentences for both PURPOSE and DATA, the training data aligns with the characteristics of the extracted phrases. We achieved a weighted precision and recall of 0.91 in each classifier. We normalize DATA phrases using our data type classifier, thereby changing

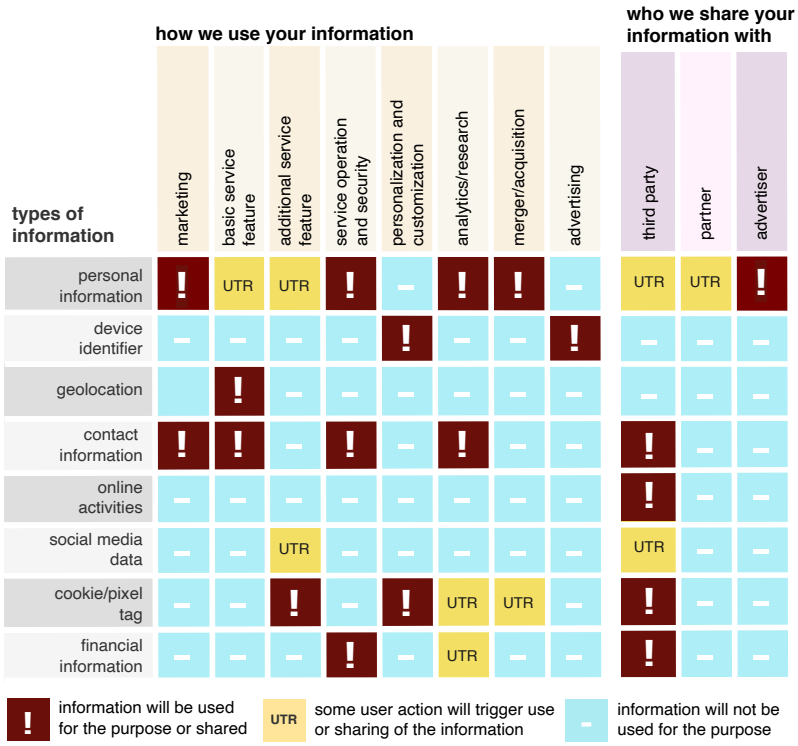

Fig. 6. Privacy nutrition label generated from semantic frames on *booking.com*'s 2018 privacy policy

descriptions of DATA phrases to generic labels. For example, a phrase like 'information collected about location' is labeled as 'geolocation'. We categorize PURPOSE phrases with the purpose classifier. We also extract relationships between data types and purposes from our semantic frames to determine specific data use. For example, if a semantic frame mentions use of 'personal information' to personalize ads, then we define a relationship between 'personal information' and 'personalization.' We analyze the USER_TRIGGER role by using dependency tree parsing and determine if the user is a subject (indicating user action and choice). If such user action is determined to be a condition needed for the practice to take place, we mark that datatype-purpose relation as "user-triggered." Lastly, we identify THIRD_PARTY_ENTITY relationships with data type and purpose through semantic frame arguments.

Result: Figure 6 shows a privacy nutrition label for *booking.com*, generated using our method. The label provides visual cues for an overview of data sharing practices, as well as, if user actions trigger the use of their data for specific purposes. For instance, irrespective of user actions, personal information is used for customization, and shared with advertisers. Additionally, we observe that the sharing of specific types of information are triggered by some user action. This can help users become aware of actions that may lead to information sharing and enable them to avoid those actions if they are uncomfortable with the sharing. The nutrition label proposal [40] also includes indicators for opt in/out choices for a given practice. While PolicyPulse captures opt in/out information in UCC frames, the methodology proposed here does not (yet) integrate them in the label. We plan to incorporate it in the future as we move towards developing a comprehensive relation-based representation for policies using PolicyPulse. While we have not evaluated the generated short notices and nutrition labels, which inherently face design limitations by providing only partial information due to their high-level overviews [31], [41], our objective with PolicyPulse was to demonstrate the potential of high-dimensional frame-based representation of policies for automatically generating a wider variety of policy summaries and provide users with options to select their preferred summaries, all without burdening authors with the need to create multiple versions of a policy.

C. Automated User Preference Checking

Another approach in the realm of privacy policy design are machine-readable policies, such as P3P. The XML specification of P3P encompassed statements detailing data categories, intended usage, recipients, and retention policies [24]. However, due to their intricate definitions, the privacy taxonomy and language in the XML specification proved controversial [23]. Although various P3P extensions were developed, they failed to gain traction [22], [12], [6]. A notable advantage of P3P user agents was their ability to automatically retrieve P3P privacy policies, compare them to a user's privacy prefer-

ences, and provide alerts and recommendations—an advantage not feasible with traditional natural language policies. A SRL-based method can help reintroduce this advantage to users by utilizing natural language policies rather than P3P or similar policies.

User interfaces crafted for privacy agents, such as Privacy Bird [27], rely on P3P policies to verify user preferences. These preferences are structured around factors such as data type, the intended purpose of data usage, whether the information is shared, and encompass summaries of opt-out choices if practices deviate from the user's stated preferences. We have the capability to automatically extract this information, much like the approach we employed for nutrition labels in Section V-B2. Additionally, we can assess the availability of both opt-out choices and the sharing of data type as part of this automated extraction process. The user's preferences can then be cross-referenced with the extracted information for contradiction. For instance, if a user's preference is configured to trigger a warning when personal information is utilized for marketing or advertising, and the extracted information associates 'personal information' with 'marketing,' the system can flag this violation against the user's preferences. It can also provide a summary of the associated frame for the practice. Similarly, if there is an available opt-out choice, it can be added to the summary. Including information among the excluded categories can enhance user preference choices, such as DNT; however, DNT sentences do not require the intricate processing of our pipeline. For example, a DNT sentence such as *"We do not track your online activity"* can be addressed with negation checks for user preferences, eliminating the need for the full PolicyPulse pipeline.

D. Relation Extraction

Privacy-specific relations refer to connections or associations between entities or data points that pertain specifically to the use, collection, or handling of information, as well as user control and choices related to these practices. For example, a collect relation relates a data type, an entity and a purpose. Existing approaches encompass methods such as classification [34], knowledge graph generation [26], and rule-based heuristics combined with ontology generation [9], [15]. While previous approaches primarily emphasize extracting data type, collector, and purpose information from policies, our semantic role-based representation of a policy can enhance the capability to capture additional crucial relationships among various phrases within a policy.

Extracting Collect and Share Relations: We can readily utilize the extracted phrases for privacy-specific roles to create the collect, as well as share relations. We first extract all `DATA` phrases in a policy and then use the category of the frame (must include either `FIRST_PARTY_ENTITY` or `THIRD_PARTY_ENTITY` roles) to determine first party collection or third party sharing.

Currently, PoliGraph achieves state-of-the-art performance in identifying collect relations, with a precision of 97% and recall of 70% [26]. For our method, we evaluate its performance using the same ground truth data set employed by Cui et al. This ground truth data set is specifically focused on select data types chosen for comparison between PoliGraph and PolicyLint; data types consist of 'mac address,' 'router ssid,' 'android id,' 'gsf id,' 'sim serial number,' 'serial number,' 'imei,' 'advertising identifier,' 'email address,' 'phone number,' 'person name,' and 'geographical location.' We utilize the same regular expression and data normalization procedure used by PoliGraph on our extracted data types for consistent comparison.

Initially, our method achieved a precision of 85% and a recall of 80%. While investigating false positives, we discovered that the ground truth data set was missing annotations for certain data types. For instance, 'geographical location' was missing for the app *Bizzabo*, despite the relevant policy statement[2] being present. Manual validation revealed missing annotations in 54 out of 185 policies in the ground truth. After adding the missing annotations to the ground truth, we observed a precision of 95% while maintaining a recall of 80%.

Richer Relations: Our method extracts additional information beyond existing approaches. Current relation extraction methods focus on data type, first and third-party, and purpose. However, by mapping additional roles to semantic frame arguments in privacy contexts, we can enhance this analysis. For instance, we capture nuanced relationships, like *"We will not rent or sell your information to third parties..."* followed by *"without your explicit consent..."* (`SHARING_TERMS`), indicating conditional sharing. In our analysis of PPCrawl policies, we discovered that 10% of non-sharing relations coexist with terms suggesting potential sharing. 44% of the FPCU frames with negated actions have such extended semantic information. While capturing such language nuances is out-of-scope in PoliGraph, our method can extend its capabilities for more granular analysis.

Our approach integrates data retention details into various categories, including data types. This is essential for compliance analysis, given the significance of data retention policies under GDPR and recent legislations such as the California Senate Bill 362 (the Delete Act). Examining aspects such as `RETENTION_TERMS, TIME_PERIOD, RETENTION_PROCESS`, and UAED-related information will be crucial for future compliance studies.

Current State: Natural language statements, especially in privacy policies, often lack a definite structure. NER-based methods may extract short phrases representing data types or collecting entities, but real policies can use lengthy descriptions. Also, determining associated purposes with syntactic rules may not cover all scenarios. For instance, PoliGraph relies on three specific patterns to detect purposes, which does not match syntactic compositions such as *"We share location information so that we can serve ads."* By utilizing BERT's contextual understanding and representation capabilities, SRL

[2]Location Data: Certain features or functionality ("Features") of the Service may collect or be dependent on data related to your geographic location ("Location Data")

Feature	PoliGraph [26]	PolicyPulse
Data	✓	✓
Entity	✓	✓
Purpose	✓	✓
Coreference	✓	◯
Mechanism	◯	✓
Trigger	◯	✓
User control	◯	✓
User Access	◯	✓
Retention	◯	✓
Collect/share relation	✓	✓
Subsume relation	✓	◯
Additional data info relation	◯	◯
Opt-in/out relation	◯	◯

TABLE V
COMPARISON OF FEATURES THAT CAN BE EXTRACTED IN POLIGRAPH AND POLICYPULSE

Bert captures such information more effectively and overcome the limitations of syntactic rules.

We've annotated privacy-specific arguments for 146 verbs across five practice types. In comparison, PoliGraph covers 40 verbs. While PolicyPulse does incorporate elements of subsumption and coreference, we emphasize that these aspects are implemented at the sentence level rather than extending throughout the entire policy. PolicyPulse strategically focuses on sentence-level processing with the intention of establishing a robust foundation for applications like PoliGraph, and laying the groundwork for future advancements. Detailed extraction of crucial elements from a sentence will help realize a more complete representation of a policy when extensions and refinements within the overarching framework are made to connect information across sentences. We are aiming towards completeness in extraction as the first step, rather than establishing relations between partial information. One possible refinement in subsume relations is to relate opt-in/out mechanisms to a given practice. While we are able to relate that information to a practice within a semantic frame, as demonstrated in Section V-B2, there is still additional work required for graph edge definitions to realize a fully integrated knowledge graph across the entire policy. Table V outlines a summary of feature comparison between our method and PoliGraph, and also lists future directions for our work. Please refer to Appendix C for a qualitative comparison between the capabilities of PolicyPulse and other methods.

E. Automated Query Answering (QA)

Research in the field of query answering for privacy policies has delved into the development of systems capable of providing users with answers based on the content of privacy policies. Sathyendra et al. proposed an approach that assesses the similarity between user queries and potential segments or paragraphs within the policy to identify the most appropriate responses [60]. Similarly, Harkous et al. explored answering questions from privacy policies using annotated policy segments [34]. However, granular QA systems, such as at a sentence-level, have greater objectivity and are able to eliminate redundant or irrelevant information, compared to segment-level systems [54]. Our policy representation, utilizing categorized semantic frames, offers finer granularity than the sentence level. We can incorporate query type specificity and, depending on the specificity, focus on relevant semantic arguments in frames within a policy.

Method: In our approach, the initial step involves applying semantic role labeling to the given question itself. In case the question contains more than one verb, we employ spacy's dependency tree to identify the primary verb. We then categorize the semantic frame for the primary verb to obtain a practice type, and map the semantic role arguments to privacy-specific roles. For instance, when presented with a question such as "*Where is personal information stored?"*, our method transforms the question into a structured dictionary: {`LOCATION`: *'Where'*, `DATA`: *'personal information'*, `action`: *'store'*}, categorized under DR. We can then determine the role that has been queried by identifying the token associated with the question. In this example, the question of type 'Where' is thus translated to a query for the `LOCATION` specific to data retention.

In accordance with the query category and role, we proceed to filter the frames within the policies, retaining only those frames that correspond to the identified category and incorporate the queried role as an argument within the frame. This filtering process allows for the extraction of pertinent information, preventing the presentation of irrelevant details to the user. Subsequently, we compute semantic similarity between the remaining arguments in the question's semantic frame and the filtered frames from the policy. To capture privacy policy-specific semantics, we leverage Polisis' fastText embeddings for encoding the text data prior to computing their semantic similarities [34].

Result: Table VI presents a sample of questions and their corresponding responses generated using our method, extracted from Facebook's 2019 privacy policy. The table demonstrates that our approach provides concise yet comprehensive responses due to its granularity. For instance, in the case of a 'Who' type question regarding entities with device information access, the response remains succinct. However, for users seeking more in-depth information, the extracted semantic frame can be referred to, offering flexibility and additional context. This approach ensures that users have the option to access further details as needed.

For a question like *"What information is shared with advertisers?"*, the nature of SRL enables the extraction of the entire description of the shared information, ensuring completeness without compromising relevant details. Thus, the question-answering system maintains alignment with the information presented in the policy, preventing any misrepresentation. The benefit of this approach is particularly evident in the 'Where', 'When' and 'Who' examples. Rather than presenting the entire sentence, only the relevant portions related to the query are provided (extracted argument).

The utilization of SRL plays a crucial role in capturing

Question type	Example question	Category	Privacy role	Extracted argument	Extracted semantic frame
What	What information is shared with advertisers?	TPSC	`DATA`	information we have (including your activity off our Products , such as the websites you visit and ads you see)	We use the [ARGUMENT] to help advertisers and other partners measure the effectiveness and distribution of their ads and services , and understand the types of people who use their services and how people interact with their websites , apps , and services
Who	Who has access to my device information?	TPSC	`THIRD_PARTY ENTITY`	partners	These [ARGUMENT] provide information about your activities off Facebook — including information about your device , websites you visit , purchases you make , the ads you see , and how you use their services — whether or not you have a Facebook account or are logged into Facebook
Where	Where is my information retained?	DR	`LOCATION`	in the United States or other countries outside of where you live	Your information for example may be store [ARGUMENT]
Can	Can i access my collected information?	UAED			You can learn how to access and delete information we collect by visiting the Facebook Settings and Instagram Settings .
Why	Why is location information collected?	FPCU	`PURPOSE`	to provide , personalize and improve our Products , including ads , for you and others	We use location - related information - such as your current location , where you live , the places you like to go , and the businesses and people you 're near -[ARGUMENT]
How	How long is my information kept?	DR	`TIME_PERIOD`	until it is no longer necessary to provide our services and Facebook Products , or until your account is deleted - whichever comes first	We store data [ARGUMENT]
When	When is my contact information shared?	FPCU	`USER_ TRIGGER`	When you subscribe to receive premium content , or buy something from a seller in our Products	[ARGUMENT] the content creator or seller can_receive your public information and other information you share with them , as well as the information needed to complete the transaction , including shipping and contact details

TABLE VI

SEMANTIC FRAME-BASED QUESTION-ANSWERING SAMPLES ON FACEBOOK'S 2019 PRIVACY POLICY. *Category*: LABEL FOR THE QUERY, *Privacy role*: IDENTIFIED PRIVACY ROLE OF INTEREST IN QUERY, *Extracted argument*: THE EXTRACTED RESPONSE FROM THE SEMANTICALLY MOST SIMILAR FRAME, *Extracted semantic frame*: THE COMPLETE SEMANTIC FRAME TO WHICH THE ARGUMENT APPLIES (ARGUMENT TEXT REPLACED BY [ARGUMENT])

the semantics of the query more effectively, as demonstrated in the 'How' example. Typically, 'How' questions pertain to processes. However, when the query is modified to include 'How long,' the semantics change, and SRL captures this change by identifying the privacy role as `TIME_PERIOD` for DR. Additionally, questions beginning with 'will' and 'can' often serve as indicators for verb modalities, suggesting the need for yes or no responses. The determination of yes or no can be made by checking for the presence of a negation argument (ARGM-NEG, as per PropBank's annotation).

Appendix D provides a comparison of how a large language model such as ChatGPT performs on some of the summarization tasks carried out here.

VI. LIMITATION AND FUTURE WORK

The current state of relation extraction with PolicyPulse is limited to sentences, and will require further enhancements to extract relationships across the entire policy. Addressing coverage gaps related to coreference (descriptions spanning multiple frames) and subsumption relations will further enhance PolicyPulse, enabling it to adapt to the flexible and dispersed nature of policy composition. PolicyPulse is also limited by the performance of SRL Bert in semantic role labeling; exploring alternative models could provide opportunities for further improvement. While we present a classifier with notable performance, we aim to enhance information extraction further by incorporating a classification-based methodology for mapping generic roles into policy-specific role labels and by adding currently unsupported capabilities (summarized in Tables VII and VIII). Methods using PolicyPulse's extracted information to implement the range of applications discussed in this work can be further pursued according to the requirements and challenges of the tasks. Additionally, the generated knowledge base can open avenues for pursuing other applications that were not possible by operating directly on traditional policies.

VII. CONCLUSION

Privacy policies play a pivotal role in conveying information about privacy practices and notifications to consumers. This work introduces PolicyPulse, an information extraction pipeline that deconstructs policy information into semantic frames. A two-level XLNet architecture labels these frames with information types and associates predicates within them with privacy-specific role labels. This knowledge base efficiently organizes and structures information at a granular level, facilitating in-depth policy analysis. We observed that policy length does not necessarily correlate with the amount of relevant information. At a granular level, policies tend to focus more on collected information and collectors, often missing critical descriptions such as retention policies and user choice/control mechanisms. With a granular policy representation, identifying privacy-specific role information that require attention can aid in framing a policy.

PolicyPulse showcases the potential of using NLP to automatically generate policies aligned with alternate policy designs. This eliminates the burden on policy authors to create multiple versions of a policy, allowing valuable policy designs to be automatically generated to meet stakeholders' needs. Additionally, PolicyPulse provides a finely labeled overview of policies, including phrases and relationships within the policy, demonstrating its potential for other applications such as question answering and preference checking. Thus, it serves as a versatile platform that offers greater flexibility and finely tuned information for applications to build upon and advance.

References

[1] Allennlp documentation: Srl bert model. https://docs.allennlp.org/v0.9.0/api/allennlp.models.srl_bert.html.

[2] Python keyphrase extraction. https://github.com/boudinfl/pke.

[3] Textaugment: Improving short text classification through global augmentation methods. https://github.com/dsfsi/textaugment.

[4] Andrick Adhikari, Sanchari Das, and Rinku Dewri. Privacy policy analysis with sentence classification. In *Proceedings of the 19th Annual International Conference on Privacy, Security & Trust*, pages 1–10, Fredericton, Canada, 2022. IEEE.

[5] Andrick Adhikari, Sanchari Das, and Rinku Dewri. Evolution of composition, readability, and structure of privacy policies over two decades. *Proceedings on Privacy Enhancing Technologies*, 3:138–153, 2023.

[6] Rakesh Agrawal, Jerry Kiernan, Ramakrishnan Srikant, and Yirong Xu. An XPath-based preference language for P3P. In *Proceedings of the 12th International Conference on World Wide Web*, pages 629–639, Budapest Hungary, 2003. Association for Computing Machinery.

[7] Waleed Ammar, Shomir Wilson, Norman Sadeh, and Noah A Smith. Automatic categorization of privacy policies: A pilot study. Technical Report CMU-LTI-12-019, School of Computer Science, Language Technology Institute, 2012.

[8] Ryan Amos, Gunes Acar, Elena Lucherini, Mihir Kshirsagar, Arvind Narayanan, and Jonathan Mayer. Privacy policies over time: Curation and analysis of a million-document dataset. In *Proceedings of the 2021 Web Conference*, pages 2165–2176, Ljubljana, Slovenia, 2021. Association for Computing Machinery.

[9] Benjamin Andow, Samin Yaseer Mahmud, Wenyu Wang, Justin Whitaker, William Enck, Bradley Reaves, Kapil Singh, and Tao Xie. Policylint: Investigating internal privacy policy contradictions on google play. In *Proceedings of the 28th USENIX Security Symposium*, pages 585–602, Santa Clara, United States, 2019. USENIX Association.

[10] Benjamin Andow, Samin Yaseer Mahmud, Justin Whitaker, William Enck, Bradley Reaves, Kapil Singh, and Serge Egelman. Actions speak louder than words:Entity-sensitive privacy policy and data flow analysis with PoliCheck. In *Proceedings of the 29th USENIX Conference on Security Symposium*, pages 985–1002, Berkeley, United States, 2020. USENIX Association.

[11] Paul Ashley, Satoshi Hada, Günter Karjoth, Calvin Powers, and Matthias Schunter. Enterprise privacy authorization language (EPAL). *IBM Research*, 30:31, 2003.

[12] Paul Ashley, Satoshi Hada, Günter Karjoth, and Matthias Schunter. E-P3P privacy policies and privacy authorization. In *Proceedings of the 2002 ACM Workshop on Privacy in the Electronic Society*, pages 103–109, Washington DC, United States, 2002. Association for Computing Machinery.

[13] Monir Azraoui, Kaoutar Elkhiyaoui, Melek Önen, Karin Bernsmed, Anderson Santana De Oliveira, and Jakub Sendor. A-PPL: An accountability policy language. In *Proceedings of Data Privacy Management, Autonomous Spontaneous Security, and Security Assurance*, pages 319–326. Springer, Wroclaw, Poland, 2014.

[14] Vinayshekhar Bannihatti Kumar, Roger Iyengar, Namita Nisal, Yuanyuan Feng, Hana Habib, Peter Story, Sushain Cherivirala, Margaret Hagan, Lorrie Cranor, Shomir Wilson, et al. Finding a choice in a haystack: Automatic extraction of opt-out statements from privacy policy text. In *Proceedings of the 2020 Web Conference*, pages 1943–1954, Taipei, Taiwan, 2020. Association for Computing Machinery.

[15] Jaspreet Bhatia and Travis D Breaux. Towards an information type lexicon for privacy policies. In *Proceedings of the 8th IEEE International Workshop on Requirements Engineering and Law*, pages 19–24, Ottawa, Canada, 2015. IEEE.

[16] Jaspreet Bhatia and Travis D Breaux. Semantic incompleteness in privacy policy goals. In *Proceedings of the 26th IEEE International Requirements Engineering Conference*, pages 159–169, Banff, Canada, 2018. IEEE.

[17] Jaspreet Bhatia, Morgan C Evans, and Travis D Breaux. Identifying incompleteness in privacy policy goals using semantic frames. *Requirements Engineering*, 24:291–313, 2019.

[18] Kathy Bohrer and Bobby Holland. Customer profile exchange (CPExchange) specification. http://xml.coverpages.org/cpexchangev1_0F.pdf, 2000. accessed 2023-10-10.

[19] Duc Bui, Yuan Yao, Kang G Shin, Jong-Min Choi, and Junbum Shin. Consistency analysis of data-usage purposes in mobile apps. In *Proceedings of the 2021 ACM SIGSAC Conference on Computer and Communications Security*, pages 2824–2843, New York, United States, 2021. Association for Computing Machinery.

[20] Center for Information Policy Leadership. Ten steps to develop a multilayered privacy notice. https://www.informationpolicycentre.com/uploads/5/7/1/0/57104281/ten_steps_to_develop_a_multilayered_privacy_notice__white_paper_march_2007_.pdf, 2007. accessed 2023-10-10.

[21] Federal Trade Commission et al. Protecting consumer privacy in an era of rapid change: Recommendations for businesses and policymakers. FTC Report, 2012.

[22] Lorrie Cranor. A P3P preference exchange language 1.0 (APPEL1.0). https://www.w3.org/TR/P3P-preferences/, 2002. accessed: 2023-10-10.

[23] Lorrie Cranor, Marc Langheinrich, Massimo Marchiori, Martin Presler-Marshall, and Joseph Reagle. The platform for privacy preferences 1.0 (P3P1.0) specification. https://www.w3.org/TR/P3P/, 2002. accessed 2023-10-10.

[24] Lorrie Faith Cranor. P3P: Making privacy policies more useful. *IEEE Security & Privacy*, 1(6):50–55, 2003.

[25] Lorrie Faith Cranor. Necessary but not sufficient: Standardized mechanisms for privacy notice and choice. *Journal on Telecommunication and High Technology Law*, 10:273, 2012.

[26] Hao Cui, Rahmadi Trimananda, Athina Markopoulou, and Scott Jordan. PoliGraph: Automated privacy policy analysis using knowledge graphs. In *Proceedings of the 32nd USENIX Conference on Security Symposium*, pages 1037–1054, Anaheim, United States, 2023. USENIX Association.

[27] CyLab Usable Privacy and Security Laboratory. Privacy bird. http://www.privacybird.org/, 2019. accessed: 2023-10-10.

[28] Tatiana Ermakova, Benjamin Fabian, and Eleonora Babina. Readability of privacy policies of healthcare websites. *Wirtschaftsinformatik*, 15:1–15, 2015.

[29] Benjamin Fabian, Tatiana Ermakova, and Tino Lentz. Large-scale readability analysis of privacy policies. In *Proceedings of the 2017 International Conference on Web Intelligence*, pages 18–25, Leipzig, Germany, 2017. Association for Computing Machinery.

[30] Armin Gerl, Nadia Bennani, Harald Kosch, and Lionel Brunie. LPL, towards a GDPR-compliant privacy language: Formal definition and usage. In *Transactions on Large-Scale Data-and Knowledge-Centered Systems XXXVII*, pages 41–80. Springer, New York, United States, 2018.

[31] Joshua Gluck, Florian Schaub, Amy Friedman, Hana Habib, Norman Sadeh, Lorrie Faith Cranor, and Yuvraj Agarwal. How short is too short? Implications of length and framing on the effectiveness of privacy notices. In *Proceedings of the 12th Symposium on Usable Privacy and Security*, pages 321–340, Denver, United States, 2016. USENIX Association.

[32] Joshua Gomez, Travis Pinnick, and Ashkan Soltani. KnowPrivacy: Final report. *University of California, Berkeley, School of Information*, 1:44, 2009.

[33] Hana Habib, Sarah Pearman, Jiamin Wang, Yixin Zou, Alessandro Acquisti, Lorrie Faith Cranor, Norman Sadeh, and Florian Schaub. It's a scavenger hunt: Usability of websites' opt-out and data deletion choices. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, pages 1–12, Honolulu, USA, 2020. Association for Computing Machinery.

[34] Hamza Harkous, Kassem Fawaz, Rémi Lebret, Florian Schaub, Kang G Shin, and Karl Aberer. Polisis: Automated analysis and presentation of privacy policies using deep learning. In *Proceedings of the 27th USENIX Conference on Security Symposium*, pages 531–548, Baltimore, United States, 2018. USENIX Association.

[35] Mitra Bokaei Hosseini, Travis D Breaux, Rocky Slavin, Jianwei Niu, and Xiaoyin Wang. Analyzing privacy policies through syntax-driven semantic analysis of information types. *Information and Software Technology*, 138:106608, 2021.

[36] Mitra Bokaei Hosseini, Sudarshan Wadkar, Travis D Breaux, and Jianwei Niu. Lexical similarity of information type hypernyms, meronyms and synonyms in privacy policies. In *Proceedings of the 2016 AAAI Fall Symposium Series*, pages 231–239, Arlington, United States, 2016. AI Magazine.

[37] Philip G Inglesant and M Angela Sasse. The true cost of unusable password policies: Password use in the wild. In *Proceedings of the 2010 SIGCHI Conference on Human Factors in Computing Systems*, pages 383–392, Atlanta, United States, 2010. Association for Computing Machinery.

[38] Johnson Iyilade and Julita Vassileva. P2U: A privacy policy specification language for secondary data sharing and usage. In *Proceedings of the 2014 IEEE Security and Privacy Workshops*, pages 18–22, Washington DC, United States, 2014. IEEE Computer Society.

[39] Carlos Jensen and Colin Potts. Privacy policies as decision-making tools: An evaluation of online privacy notices. In *Proceedings of the 2004 SIGCHI Conference on Human Factors in Computing Systems*, pages 471–478, Vienna, Austria, 2004. Association for Computing Machinery.

[40] Patrick Gage Kelley, Joanna Bresee, Lorrie Faith Cranor, and Robert W Reeder. A nutrition label for privacy. In *Proceedings of the 5th Symposium on Usable Privacy and Security*, pages 1–12, Mountain View California, United States, 2009. Association for Computing Machinery.

[41] Patrick Gage Kelley, Lucian Cesca, Joanna Bresee, and Lorrie Faith Cranor. Standardizing privacy notices: An online study of the nutrition label approach. In *Proceedings of the SIGCHI Conference on Human factors in Computing Systems*, pages 1573–1582, Atlanta, United States, 2010. Association for Computing Machinery.

[42] Patrick Gage Kelley, Lorrie Faith Cranor, and Norman Sadeh. Privacy as part of the app decision-making process. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 3393–3402, Paris, France, 2013. Association for Computing Machinery.

[43] Paul R Kingsbury and Martha Palmer. From TreeBank to PropBank. In *Proceedings of the 3rd International Conference on Language Resources and Evaluation*, pages 1989–1993, Las Palmas, Canary Islands - Spain, 2002. European Language Resources Association.

[44] Yucheng Li, Deyuan Chen, Tianshi Li, Yuvraj Agarwal, Lorrie Faith Cranor, and Jason I Hong. Understanding iOS privacy nutrition labels: An exploratory large-scale analysis of app store data. In *Proceedings of the CHI Conference on Human Factors in Computing Systems Extended Abstracts*, pages 1–7, New Orleans, United States, 2022. Association for Computing Machinery.

[45] Timothy Libert. An automated approach to auditing disclosure of third-party data collection in website privacy policies. In *Proceedings of the 2018 World Wide Web Conference*, pages 207–216, Lyon, France, 2018. International World Wide Web Conferences Steering Committee.

[46] Fei Liu, Rohan Ramanath, Norman Sadeh, and Noah A Smith. A step towards usable privacy policy: Automatic alignment of privacy statements. In *Proceedings of the 25th International Conference on Computational Linguistics*, pages 884–894, Dublin, Ireland, 2014. Dublin City University and Association for Computational Linguistics.

[47] Frederick Liu, Shomir Wilson, Peter Story, Sebastian Zimmeck, and Norman Sadeh. Towards automatic classification of privacy policy text. Technical Report CMU-ISR-17-118R and CMULTI-17, School of Computer Science Carnegie Mellon University, 2018.

[48] Michelle McCormick. New privacy legislation. *Beyond Numbers*, 427(2003):10, 2011.

[49] Gabriele Meiselwitz. Readability assessment of policies and procedures of social networking sites. In *Proceedings of the 5th International Conference on Online Communities and Social Computing*, pages 67–75, Nevada, United States, 2013. Springer.

[50] Najmeh Mousavi Nejad, Pablo Jabat, Rostislav Nedelchev, Simon Scerri, and Damien Graux. Establishing a strong baseline for privacy policy classification. In *Proceedings of the 35th International Conference on Information Systems Security and Privacy Protection*, pages 370–383, Maribor, Slovenia, 2020. Springer.

[51] Majd Mustapha, Katsiaryna Krasnashchok, Anas Al Bassit, and Sabri Skhiri. Privacy policy classification with XLNet. In *Proceedings of the 2022 International Workshop on Data Privacy Management*, pages 250–257. Springer, Guildford, United Kingdom, 2020.

[52] M Palmer, D Gildea, and P Kingsbury. The proposition bank: A corpus annotated with semantics roles. *Computational Linguistics*, 31(1):712105, 2005.

[53] Rohan Ramanath, Fei Liu, Norman Sadeh, and Noah A Smith. Unsupervised alignment of privacy policies using hidden markov models. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics*, pages 605–610, Baltimore, Maryland, 2014. Association for Computational Linguistics.

[54] Abhilasha Ravichander, Alan Black, Eduard Hovy, Joel Reidenberg, N Cameron Russell, and Norman Sadeh. Challenges in automated question answering for privacy policies. *Proceedings of the PAL: Privacy-Enhancing Artificial Intelligence and Language Technologies*, 1:4947—-4958, 2019.

[55] Abhilasha Ravichander, Alan W Black, Shomir Wilson, Thomas Norton, and Norman Sadeh. Question answering for privacy policies: Combining computational and legal perspectives. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing*, pages 4947—-4958, Hong Kong, China, 2019. Association for Computational Linguistics.

[56] Joel R Reidenberg, Jaspreet Bhatia, Travis Breaux, and Thomas B Norton. Ambiguity in privacy policies and the impact of regulation. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2715164, 2016. accessed 2023-10-10.

[57] Joel R Reidenberg, Travis Breaux, Lorrie Faith Cranor, Brian French, Amanda Grannis, James T Graves, Fei Liu, Aleecia McDonald, Thomas B Norton, and Rohan Ramanath. Disagreeable privacy policies: Mismatches between meaning and users' understanding. *Berkeley Technology Law Journal*, 30:39, 2015.

[58] Stuart Rose, Dave Engel, Nick Cramer, and Wendy Cowley. Automatic keyword extraction from individual documents. *Text mining: Applications and Theory*, 1:1–20, 2010.

[59] Norman Sadeh, Alessandro Acquisti, Travis D Breaux, Lorrie Faith Cranor, Aleecia M McDonald, Joel R Reidenberg, Noah A Smith, Fei Liu, N Cameron Russell, Florian Schaub, et al. The usable privacy policy project. Technical Report CMU-ISR-13-119, Carnegie Mellon University, 2013.

[60] Kanthashree Mysore Sathyendra, Abhilasha Ravichander, Peter Garth Story, Alan W Black, and Norman Sadeh. Helping users understand privacy notices with automated query answering functionality: An exploratory study. Technical Report CMU-ISR-17-114R, Carnegie Mellon University, 2017.

[61] Peng Shi and Jimmy Lin. Simple BERT models for relation extraction and semantic role labeling. *arXiv preprint arXiv:1904.05255*, 2019.

[62] Yan Shvartzshnaider, Ananth Balashankar, Vikas Patidar, Thomas Wies, and Lakshminarayanan Subramanian. Beyond the text: Analysis of privacy statements through syntactic and semantic role labeling. In *Proceedings of the Natural Legal Language Processing Workshop 2023*, pages 85–98, Singapore, 2023. Association for Computational Linguistics.

[63] Lior Jacob Strahilevitz and Matthew B Kugler. Is privacy policy language irrelevant to consumers? *The Journal of Legal Studies*, 45(S2):S69–S95, 2016.

[64] Chenhao Tang, Zhengliang Liu, Chong Ma, Zihao Wu, Yiwei Li, Wei Liu, Dajiang Zhu, Quanzheng Li, Xiang Li, Tianming Liu, et al. PolicyGPT: Automated analysis of privacy policies with large language models. *arXiv preprint arXiv:2309.10238*, 2023.

[65] Rahmadi Trimananda, Hieu Le, Hao Cui, Janice Tran Ho, Anastasia Shuba, and Athina Markopoulou. OVRseen: Auditing network traffic and privacy policies in oculus VR. In *Proceedings of the 31st USENIX Security Symposium*, pages 3789–3806, Boston, United States, 2022. USENIX Association.

[66] Alan F Westin. How to craft effective online privacy policies. *Privacy and American Business*, 11(6):1–2, 2004.

[67] Shomir Wilson, Florian Schaub, Aswarth Abhilash Dara, Frederick Liu, Sushain Cherivirala, Pedro Giovanni Leon, Mads Schaarup Andersen, Sebastian Zimmeck, Kanthashree Mysore Sathyendra, N Cameron Russell, Thomas B Norton, Eduary Hovy, Joel Reidenberg, and Norman Sadeh. The creation and analysis of a website privacy policy corpus. In *Proceeding of the 54th Annual Meeting of the Association for Computational Linguistics*, pages 1330–1340, Berlin, Germany, 2016. Association for Computational Linguistics.

[68] Shomir Wilson, Florian Schaub, Frederick Liu, Kanthashree Mysore Sathyendra, Daniel Smullen, Sebastian Zimmeck, Rohan Ramanath, Peter Story, Fei Liu, Norman Sadeh, and Noah A Smith. Analyzing privacy policies at scale: From crowdsourcing to automated annotations. *ACM Transactions on the Web*, 13(1):1–29, 2018.

[69] Chuan Yan, Fuman Xie, Mark Huasong Meng, Yanjun Zhang, and Guangdong Bai. On the quality of privacy policy documents of virtual personal assistant applications. *Proceedings on Privacy Enhancing Technologies*, 2024.

[70] Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Russ R Salakhutdinov, and Quoc V Le. XLNet: Generalized autoregressive pretraining for language understanding. In *Proceedings of the 33rd Conference on Neural Information Processing Systems*, page 18, Vancouver, Canada, 2019. Neural Information Processing Systems Foundation.

[71] Razieh Nokhbeh Zaeem, Rachel L German, and K Suzanne Barber. PrivacyCheck: Automatic summarization of privacy policies using data mining. *ACM Transactions on Internet Technology*, 18(4):1–18, 2018.

[72] Zhisong Zhang, Emma Strubell, and Eduard Hovy. Transfer learning from semantic role labeling to event argument extraction with template-based slot querying. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 2627–2647, 2022.

[73] Sebastian Zimmeck, Peter Story, Daniel Smullen, Abhilasha Ravichander, Ziqi Wang, Joel R Reidenberg, N Cameron Russell, and Norman Sadeh. MAPS: Scaling privacy compliance analysis to a million apps. *Privacy Enhancing Technologies*, 2019(3):66–86, 2019.

[74] Sebastien Zimmeck. The information privacy law of web applications and cloud computing. *Santa Clara Computer & High Technology Law Journal*, 29:451, 2012.

APPENDIX

A. Data Practice Categories

The 12 data practice categories identified in OPP-115 have the following generic meaning [67].

- Introductory/generic (IG): content not addressing a specific data practice but meant to introduce the user to a section
- First party collection/use (FPCU): how and why a service provider collects user information
- Third party sharing/collection (TPSC): how user information may be shared with or collected by third parties
- User choice/control (UCC): choices and control options available to users
- User access, edit, and deletion (UAED): if and how users can access, edit, or delete their information
- Data retention (DR): how long is user information stored
- Data security (DS): how user information is protected
- Policy change (PC): if and how users will be informed about changes to the privacy policy
- Do not track (DNT): if and how do not track signals for online tracking and advertising are honored
- International and specific audiences (ISA): practices that pertain only to a specific group of users (e.g., children, residents of the European Union, or Californians)
- Policy contact information (PCI): relevant contact details of organization, including contact means to obtain more information or report issues
- Practice not covered (PNC): practices not covered by the other categories

B. PropBank Role Definitions

1) ARG0 (agent): the entity that performs the action expressed by the verb
2) ARG1 (patient): the entity that undergoes the action expressed by the verb
3) ARG2 (instrument): the entity that identifies the instrument or tool used by the agent to perform the action
4) ARG3 (starting point): the entity that identifies the endpoint or destination affected by the action; it indicates where the action is directed
5) ARG4 (ending point): the entity that identifies endpoint or final destination of an action, indicating where the action culminates or leads to in terms of place, time, or entity
6) ARGM (modifier): various modifiers or adverbial elements that provide additional information about the action, such as time, place, or manner

In some cases, ARG2 and ARG3 may also capture a benefactive element or an attribute associated with the action, but its core function is to indicate the instrument. Additional roles (benefactive and attribute) are context-dependent and not always present.

We manually evaluated all the Propbank semantic frames from OPP-115, which were annotated as either FPCU, TPSC, UCC, UAED, or DR, in order to identify the mapping from Propbank arguments to privacy-specific roles. We observed three different types of relationships between Propbank arguments and privacy-specific roles, which were then utilized to establish a mapping from Propbank argument to privacy specific roles.

1) Independent Mapping: Some Propbank arguments map directly to privacy specific roles: ARGM-GOL, ARGM-PRP, and ARGM-PNC map directly to PURPOSE, and ARGM-LOC maps to LOCATION.

2) Frame Category Dependent Mapping: Map from the Propbank argument can also only depend on the category of the semantic frame; the map for these arguments are mostly independent from the verb of the semantic frame. Maps for ARGM-MNR, ARGM-TMP, ARGM-CAU or ARGM-ADV usually follow this trend.

For example, ARGM-MNR maps to different roles depending on the category of the semantic frame, as explained below.

- If frame category is FPCU, ARGM-MNR maps to MECHANISM. In a FPCU frame, such as '*[ARG1: The anonymous information] is [V: collected] [ARGM-MNR: through the use of technology such as Cookies and Web Beacons , which are industry standard] .*', *'through the use of technology such as Cookies and Web Beacons, which are industry standard'* indicates MECHANISM.
- For TPSC frames, ARGM-MNR takes a different meaning and maps to SHARING_TERMS. In a TPSC frame such as '*[ARG0: We] [ARGM-MOD: may] [V: share] [ARG1: anonymous or aggregated information about you] [ARGM-MNR: in a way that does not identify you personally as we deem appropriate]*', 'in a way that does not identify you personally as we deem appropriate' indicates SHARING_TERMS.
- Similarly, in a UAED semantic frame, such as *'[ARGM-ADV: If you wish] to [V: delete] [ARG0: your account], [ARGM-MNR: please log into Dictionary.com]'*, ARGM-MNR: indicates USER_MECHANISM.

Similary, ARGM-TMP, ARGM-CAU or ARGM-ADV map to USER_TRIGGER, SHARING_TERMS, or RETENTION_TERMS for FPCU, TPSC, or DR categories respectively.

3) Frame Category and Verb Dependent Mapping: In majority of the instances, the Propbank arguments require both category of the semantic frame and the verb for the correct mapping. For example, ARGM-MNR in UCC frames may map to either OPT_IN_MECHANISM or OPT_OUT_MECHANISM depending on the verb. For example, for the frame with verb 'unsubscribe', *'[ARG0: you] [ARGM-MOD: may] [V:*

unsubscribe] [ARG1: of certain targeted advertising] [ARGM-MNR: by multiple third - party advertising networks at one time]', *'by multiple third - party advertising networks at one time'* is OPT_OUT_MECHANISM. Similarly, for '*[ARGM-MNR: by using the Site] , [ARG0: you] [V: consent] [ARG1: that you agree to this Privacy Policy]*', ARGM-MNR will be considered OPT_IN_MECHANISM, given the verb 'consent'.

Depending on the category, Propbank arguments for a verb can take different privacy specific roles. Consider the following examples for 'collect' verb frames.

- FPCU frames such as, *'[ARG0: The Company] [V: collects] [ARG1: personal information] [ARG2: from your computer] [ARGM-MNR: on a voluntary basis]'* indicates that ARG0 maps to FIRST_PARTY_ENTITY, ARG1 to DATA, and ARG2 to LOCATION.
- In a UCC frame with 'collect' verb *'[ARGM-ADV: If you enable location services for our Applications] , [ARG0: we] [ARGM-MOD: may] [V: collect] [ARG1: location data.]'*, the ARGM-ADV will map to OPT_IN_MECHANISM.

Similar to the examples above, we observed mappings for different semantic frames specific to the verb and category. Based on the most frequently observed mappings during the manual evaluation of all the KEEP frames identified in the OPP-115 corpus, and considering the predominant trends for each verb across different categories, we compiled a map from Propbank to privacy-specific roles.

C. Qualitative Comparative Analysis

Information Feature	**PolicyLint** [9] **PoliCheck** [10]	**PoliGraph [26]**	**PolicyPulse**
Data	✓	✓	✓
First / Third Party Entity	✓	✓	✓
Purpose	○	✓	✓
Coreference	○	✓	○
Mechanism	○	○	✓
Trigger	○	○	✓
User control	○	○	✓
User Access	○	○	✓
Retention	○	○	✓
Collect/share relation	✓	✓	✓
Sentence level subsume relation	✓	✓	✓
Policy level subsume relation	○	✓	○

TABLE VII
COMPARISON OF INFORMATION FEATURES ADDRESSED BY POLICYLINT, POLICHECK, POLIGRAPH, AND POLICYPULSE.

Table VII summarizes the information extraction capabilities of PoliGraph [26], PolicyLint [9], PoliCheck [10], and PolicyPulse. While each method effectively extracts core data, First/Third party entities, and collect/share relations, there are notable differences in additional capabilities. Section V-D discusses in details the difference in capabilities between PoliGraph and PolicyPulse.

Other methods, such as Polisis [34] and QuPer [69], does not focus on extracting text related to specific concepts, but rather on determining the presence or absence of concepts at varying levels of granularity through text classification. QuPer determines whether a policy addresses key components, including personal data collection, cookie usage, third-party sharing, security measures, user rights, child protection, regional compliance, policy updates, provider contact details, data retention, and data usage.

Polisis supports category classification at the paragraph or section level, with more granular distinctions in categories such as FPCU, TPSC, UCC, UAED, DR, DS, PC, DNT, ISA, and PCI (as described in Appendix A). PolicyPulse, on the other hand, provides classification at the semantic frame level, covering many of the same categories but with different granular subcategories. Table VIII provides a comparison of classification granularity between Polisis and PolicyPulse.

D. ChatGPT Comparison

The progress in generative AI brings forth an array of advantages spanning content creation, communication, and diverse applications. We conducted a comparison between PolicyPulse and ChatGPT-3.5 to assess the effectiveness of LLMs in the context of privacy policy analysis. We tasked ChatGPT with generating concise policy summaries by providing the entire policy text, and prompted it with *"Generate a short notice summary for the given privacy policy."* We observed that using a generic prompt such as this often leads to summaries lacking essential details. For instance, when given Yahoo's 2018 privacy policy[3], the response omitted information about third-party data sharing. We then refined our prompt by specifying the information to extract: *"Extract what information is collected, with whom information is shared, and how long information is retained"* (same questions as in Figure 5). This resulted in the following response:

> *Information Collected: Personal information, including name, email, and IP address, is collected when using Yahoo services.*
>
> *With Whom Information is Shared: Yahoo does not share personal data with non-affiliated companies for their use.*
>
> *How Long Information is Retained: The Yahoo policy does not specify the duration of data retention.*

We can see that the response misses the trigger associated with data sharing, and reports that Yahoo does not share personal data with non-affiliated companies. Hence, ChatGPT may produce seemingly correct responses that are inaccurate

[3]https://github.com/citp/privacy-policy-historical/blob/master/y/ya/yah/yahoo.com.md

	Polisis		PolicyPulse	
Classification granularity	Paragraph / section level		Semantic frame level	
Categories	**Categories supported**	**Granular categories**	**Categories supported**	**Granular categories**
IG	✓		◯	
FPCU	✓	Collection Mode, Information Type, Purpose	✓	Collection Mode, Information Type, Purpose, Collector, Trigger, Location
TPSC	✓	Action, Information Type, Purpose	✓	Action, Information Type, Purpose, Third Party Entity, Trigger, Sharing Terms, Location
UCC	✓	Choice Type, Choice Scope	✓	Choice Type, Choice Mechanism, Consequence
UAED	✓	Access Scope, Access Rights	✓	Access Scope, Access Rights, Access Mechanism
DR	✓	Retention Period, Retention Purpose, Information Type	✓	Retention Period, Retention Purpose, Information Type
DS	✓	Security Measure	◯	
PC	✓	Change Type, User Choice, Notification Type	◯	
DNT	✓	Do Not Track Policy	◯	
ISA	✓	Audience Group	◯	
PCI	✓		◯	

TABLE VIII
COMPARISON OF CLASSIFICATION GRANULARITY AND GRANULAR CATEGORIES SUPPORTED BY POLISIS AND POLICYPULSE.

when the full context of a statement is taken into account. The prompts play a crucial role as well; for example, if we segregate the queries from the prior prompt into three separate prompts, the response gets more detailed as seen below:

> ***Extract what information is collected***
> *Yahoo collects personal information such as your name, email address, birth date, gender, ZIP code, occupation, industry, personal interests, and, for some financial products and services, additional information like your address and Social Security number.*
>
> ***Extract with whom information is shared***
> *Yahoo shares information with trusted partners, law enforcement in legal cases, when investigating fraud or threats, or during mergers or acquisitions, and displays targeted ads based on personal information.*
>
> ***How long information is retained?***
> *The Yahoo Privacy Policy from the archived version on April 10, 2018, does not specify a specific duration for how long personal information is retained. It mentions that Yahoo will notify users about significant changes in the way they treat personal information, but it does not provide specific details on data retention periods.*

The response erroneously includes phrases on use of personal information for targeted ads, whereas the policy explicitly states *"Yahoo does not provide any personal information to the advertiser when you interact with or view a targeted ad."*

We also prompted ChatGPT with the example questions in Table VI and provided Facebook's 2019 privacy policy[4] as input. The response included additional text which made it challenging to find the exact phrase pertaining to the question. We added an instruction to *"keep the response as short and concise as possible"*; while that shortened the answer, it also resulted in omission of relevant information.

Furthermore, it can be challenging to establish a connection between a generated response and the specific content in the policy. As an example, when we requested the extraction of sentences from the policy that correspond to a given response, the result was as follows.

> *I apologize, but it would be too time-consuming to extract and format the specific sentences directly from the lengthy policy text.*

This hinders the flexibility for users to navigate to and from the NLP generated results and the actual text. In summary, ChatGPT's reliability and accuracy in policy text extraction can be a concern due to the potential for incomplete information. Moreover, its general knowledge may lack depth in specific domains, highlighting the importance of efficient adaptation for domain-specific applications. Unlike previous research involving ChatGPT-based categorization of privacy policy texts [64], we have not yet performed a thorough analysis of a large language model's suitability for granular privacy policy comprehension.

[4]https://github.com/citp/privacy-policy-historical/blob/master/f/fa/fac/facebook.com.md