



# Privacy2Practice: Leveraging Automated Analysis for Privacy Policy Transparency and Compliance

Saja Alqurashi<sup>1,2</sup> <sup>a</sup> and Indrakshi Ray<sup>1</sup>  <sup>b</sup>

<sup>1</sup>Department of Computer Science, Colorado State University, U.S.A.

<sup>2</sup>Department of Information Technology, King Abdulaziz University, Saudi Arabia

**Keywords:** Privacy Policy, General Data Protection Regulation (GDPR), California Consumer Privacy Act (CCPA), Natural Language Processing (NLP).


**Abstract:** Privacy policies play a critical role in safeguarding information systems, yet they are frequently expressed in lengthy, complex natural language documents. The intricate and dense language of these policies poses substantial challenges, making it difficult for both novice users and experts to fully comprehend data collection, sharing practices, and the overall transparency of data handling. This issue is particularly concerning given the necessity of disclosing data practices to users, as mandated by privacy regulations such as the General Data Protection Regulation (GDPR) and the California Consumer Privacy Act (CCPA). To address these challenges and improve data transparency, this paper introduces *Privacy2Practice*, a comprehensive automated framework leveraging Natural Language Processing (NLP) techniques to extract and analyze key information from privacy policies. By automating the identification of data practices mandated by privacy regulations, the framework assesses how transparently these practices are disclosed, ensuring better alignment with regulatory requirements. The proposed approach significantly enhances the transparency and the compliance of privacy policies by identifying entities (F1-scores: 97% for first-party and 93% for third-party entities), data types (F1-score: 82%), and purposes of data collection and sharing (F1-score: 90%). These results underscore the importance of transparency, particularly when data is shared with external parties, and highlight the challenges associated with automating privacy policy analysis. The results highlight significant challenges, such as undisclosed third-party sharing, while showcasing the potential of automation to be more comprehensive, transparent and compliant with regulatory standards.


## 1 INTRODUCTION

A privacy policy is a legal document that outlines how an organization collects, uses, stores, and shares user data while establishing the terms of data privacy. As a legally binding mechanism, it ensures transparency by informing users of an organization's data practices. Privacy policies are crucial for regulatory compliance with frameworks such as the General Data Protection Regulation (GDPR) (Tankard, 2016), the California Consumer Privacy Act (CCPA) (de la Torre, 2018), and the Health Insurance Portability and Accountability Act (HIPAA) (Annas, 2003). These regulations mandate strict guidelines for transparency and accountability in handling personal data (Cohen, 2008; Zaem and Barber, 2020; Erkkilä, 2020). An effective privacy policy provides a clear and compre-

hensive explanation of how customer personal data is collected, used, stored, and shared, while explicitly detailing their privacy rights. To truly serve its purpose, the policy must be transparent and articulated in a way that is easy for users to understand.

Under CCPA, businesses must disclose what personal information is collected, how it is used, whether it is shared or sold, and how long it will be retained. This information is typically provided in a notice linked on the business's website before or at the time of data collection. Consumers are also informed of their rights, including the right to know, delete, opt out of the sale of their data, and be free of discrimination for exercising these rights (de la Torre, 2018). Similarly, GDPR requires data controllers to provide clear and accessible information to data subjects, including the identity of the data controller, the purpose of processing, the legal basis for processing, retention periods, and any potential data transfers. Fur-

<sup>a</sup>  <https://orcid.org/0009-0009-9118-5533>

<sup>b</sup>  <https://orcid.org/0000-0002-0714-7676>

thermore, privacy notices must explain the rights of individuals, such as access, rectification, erasure, data portability, and the right to object or withdraw consent where applicable (Tankard, 2016).

However, studies indicate that privacy policies present significant obstacles to transparency. Their complexity, dense legal jargon, and excessive length make them difficult for novice and expert users to understand, affecting their effectiveness in clearly communicating data practices (Antón et al., 2004; Jensen and Potts, 2004). Although the primary purpose of privacy policies is to inform users about their rights and the data practices of an organization, most individuals do not read them. Even those who attempt to engage with these documents often find the information overwhelming. Fully understanding the privacy policies of all the services one uses would require an impractical amount of time and effort, further discouraging users from engaging with them. This complexity exacerbates the gap in privacy policy transparency, leaving users poorly informed about how their personal information is collected, used, and protected.

To address these challenges, recent research has focused on developing methods to extract and analyze critical information from privacy policies automatically. These efforts aim to make privacy policies more understandable while ensuring compliance with legal frameworks (Harkous Hamza et al., 2018; Andow et al., 2019; Andow et al., 2020; Elluri et al., 2020). Despite these advancements, ensuring transparency in privacy policies remains a challenge. The length and complexity of these policies continue to make it difficult for consumers to fully understand how their personal information is used by the organizations.

Current studies have not thoroughly examined the specific types of data extracted from privacy policies, such as personal information categories or the stated purposes for data collection and sharing, including marketing activities or service provision (Andow et al., 2019; Andow et al., 2020; Bui et al., 2021). To address these gaps, we introduce *Privacy2Practice*, a framework designed to extract and classify data practices outlined in privacy policies, along with their corresponding purposes. We provide a methodology of transforming intricate natural language requirements into structured and actionable formats pertaining to privacy policies. This approach emphasizes systematic analysis and alignment of privacy policies with regulatory mandates, ensuring clarity, transparency, and compliance, thereby providing a comprehensive understanding of how organizations manage user data and the intent behind these practices.

This paper presents *Privacy2Practice*, an auto-

mated framework designed to analyze privacy policies articulated in natural language documents. *Privacy2Practice* is designed to identify and extract data practices embedded within these policies that describe the procedures organizations follow when collecting, storing, sharing, and protecting personal information. Our approach focuses on three essential components of data practices: *entities*, *data types*, and *purpose types*, each representing key aspects of data collection and sharing.

*Privacy2Practice* leverages a Natural Language Processing (NLP)-driven approach to systematically analyze privacy policies. It begins with (i) *Entity Type Identification*, classifying data controllers as first party, that is, entities that directly collect user data, or third party, that is, external entities that receive shared data. This classification relies on binary classification and dependency parsing. Next, (ii) *Data Type Identification* extracts and categorizes the specific types of data collected, such as personal or financial information, using a multiclass classification approach. Following this, (iii) *Purpose Type Identification* determines the intent behind data collection, mapping it to categories like advertising or analytics through multiclass classification. Finally, (iv) *Policy Policy Analysis* evaluates how clearly and transparently privacy policies align with regulatory frameworks such as GDPR and CCPA.

The results of this approach are highly promising, with the framework achieving an impressive F1-score of 97% for first-party entity identification and 93% for third-party entities. Additionally, the framework demonstrated its effectiveness in accurately identifying data types, attaining an F1-score of 82%. Furthermore, it excelled in determining the purposes of data collection and sharing, achieving a strong F1-score of 90%. These results underscore the framework's potential to enhance transparency and regulatory compliance in privacy policy analysis. In our analysis of multiple privacy policies, we identified frequent instances of non-compliance, often linked to undisclosed third-party data sharing, particularly with advertisers. These findings emphasize the importance of transparency, especially when personal data is shared externally.

The paper is organized as follows: Section 2 reviews related work, while Section 3 details the proposed methodology. Section 4 describes the datasets. Section 5 covers NLP tasks such as entity, data, and purpose type identification, while Section 6 focuses on privacy policy analysis. Section 7 concludes with key findings and future directions.

## 2 LITERATURE REVIEW

Sunkle et al. (Sunkle et al., 2015) developed a compliance-checking approach based on the Semantics of Business Vocabulary and Rules (SBVR). Their methodology translates regulatory requirements into operational procedures by creating semantic vocabularies and constructing logical expressions for policy rules. While this approach ensures alignment with organizational processes, its reliance on semi-automated processes and semantic similarity models poses scalability and adaptability challenges, particularly when addressing complex regulatory language.

Elluri et al. (Elluri et al., 2020) introduced an automated framework for aligning privacy policies with GDPR guidelines. Using Doc2Vec for semantic similarity, they constructed an ontology to compare compliance levels between organizational policies and GDPR requirements. Though effective in identifying gaps, the approach's reliance on semantic similarity limits its ability to fully capture legal nuances and contextual intricacies, especially for complex or lengthy documents.

Mousavi et al. (Mousavi Nejad et al., 2018) proposed *KnIGHT*, an automated tool that maps privacy policy statements to corresponding GDPR articles. Using NLP techniques, *KnIGHT* identifies semantically significant sentences and aligns them with GDPR clauses. Despite its potential for improving compliance efforts, expert evaluations highlight its partial accuracy and inability to address false negatives, underscoring the need for enhanced precision.

Hamdani et al. (Hamdani et al., 2021) presented a GDPR compliance framework that combines machine learning and rule-based approaches to evaluate privacy policies. Their system extracts data practices using multi-label classification and applies predefined rules derived from GDPR Articles 13 and 14. While promising, the incomplete coverage of GDPR concepts within their taxonomy and moderate performance in data practice extraction indicate room for improvement in achieving comprehensive compliance verification.

Amaral et al. (Amaral et al., 2023) introduced DERECHA, a machine-learning-based method for verifying the compliance of Data Processing Agreements (DPAs) with GDPR. Using a conceptual model of compliance requirements, DERECHA assesses compliance through semantic parsing and NLP techniques, achieving high precision and recall. However, its reliance on predefined models limits adaptability to nuanced or context-specific GDPR interpretations.

Andow et al. (Andow et al., 2019) developed *PolicyLint*, a tool for detecting contradictions in privacy

policies of mobile applications. By analyzing over 11,000 apps, *PolicyLint* identified contradictions in 14.2% of policies, including misleading statements and redefined terms. While effective in detecting inconsistencies, the tool's inability to distinguish between data-collecting entities limits its analytical accuracy.

Bui et al. (Bui et al., 2021) proposed *PurPliance*, an automated approach for detecting inconsistencies between privacy policy statements and actual data usage behaviors. By analyzing semantic structures and leveraging predicate-argument patterns, *PurPliance* significantly improved detection precision and recall compared to previous methods. However, its reliance on context-insensitive patterns may hinder its application in more complex scenarios.

Nguyen et al. (Nguyen et al., 2021) examined compliance issues related to GDPR's consent management requirements, analyzing how applications handle user consent for data processing activities. Their study provided valuable insights into violations and recommendations for enhancing user privacy. However, the research was limited as it assessed consent without executing the applications, overlooking potential violations occurring later in the data life-cycle.

Neupane et al. (Neupane et al., 2022) conducted an extensive analysis of the privacy and security risks in mobile companion apps, uncovering significant vulnerabilities and threats to user privacy. They emphasized the importance of stronger regulations and enforcement to protect user data. The study evaluated Android companion apps across three dimensions: data privacy, security, and risk. However, it did not incorporate GDPR considerations within these pillars.

Alfawzan et al. (Alfawzan et al., 2022) reviewed privacy, data sharing, and security policies in women's mHealth apps under EU GDPR compliance. The study highlighted concerns such as inadequate security measures, unclear third-party data sharing, and insufficient consent mechanisms. Despite these findings, the study's scope was limited to 23 apps, reducing its generalizability.

## 3 OVERVIEW OF OUR APPROACH

We developed *Privacy2Practice*, an NLP pipeline designed to overcome the complexities of identifying and analyzing data practices embedded in privacy policy documents. Using Few-Shot Learning techniques, *Privacy2Practice* facilitates efficient privacy policy

analysis, ensuring that organizations remain aligned with regulatory frameworks such as GDPR and CCPA while enhancing transparency and compliance. The pipeline is specifically engineered to extract critical data practices mandated by the major privacy regulations, including GDPR and CCPA (Harkous Hamza et al., 2018). These practices are categorized into three key dimensions as follows:

- **Entities:** This dimension identifies the parties involved in data collection and sharing processes:
  - **First Party:** The organization or entity directly responsible for collecting user data.
  - **Third Party:** External entities or organizations that receive data from the first party for various purposes.
- **Data Type:** This dimension defines the specific categories of data collected or processed by the organization, such as personally identifiable information (PII), financial details, browsing history, and other relevant data types.
- **Purpose Type:** This dimension captures the underlying reasons for data collection and sharing, including objectives like service improvement, targeted advertising, or compliance with legal obligations.

Figure 1 illustrates the architecture of the *Privacy2Practice* pipeline to identify and classify *entities*, *data*, and *purposes* types within privacy policies.

*Privacy2Practice* analyzes privacy policies leveraging advanced natural language processing (NLP) techniques, including Few-Shot learning, a subfield of machine learning in which models are trained to perform tasks with only a few labeled examples. This is beneficial in privacy policy analysis, where annotated datasets are scarce. Traditional supervised learning approaches require large volumes of labeled data to generalize well, which is often impractical in the privacy domain. Few-Shot Learning models (Parnami and Lee, 2022) overcome this limitation by using pre-trained sentence transformers and contrastive learning to quickly adapt to new tasks.

In this work, we adopted SetFit, a state-of-the-art Few-Shot Learning algorithm proposed by Tunstall et al. (Tunstall et al., 2022), which has demonstrated high performance in a range of text classification applications.

Utilizing the Few-Shot Learning approach, the implementation of *Privacy2Practice* consists of four downstream tasks aimed at automating the identification and categorization of data practices (entity, data, and purpose types). These tasks include:

**Task 1: Entity Type Identification:** This task extracts essential entities, including first-party and third-

party entities, from natural language text. Binary classification, enhanced with dependency parsing, is employed to identify specific entities in the collection or sharing of data.

**Task 2: Data Type Identification:** This task detects the presence of data within policy segments and accurately classifies the specific data types mentioned. To achieve precise categorization, a multi-class classification approach is utilized.

**Task 3: Purpose Type Identification:** This task analyzes policy segments to determine whether they specify the purposes of the data collection and sharing activities, categorizing identified purposes accordingly using a multi-class classification approach.

**Task 4: Privacy Policy Analysis:** This task evaluates the transparency of a privacy policy by verifying whether it adequately discloses all identified data practices, namely entities, data types, and purposes of data collection and sharing. If any of the expected data practices are missing or vague, the policy is flagged for further review. This review should be conducted by the appropriate compliance personnel, such as a privacy analyst or auditor.

The following sections of this paper provide a detailed overview of the dataset used for our approach, a comprehensive explanation of each task, a description of the experimental methodology, and key insights gained from the implementation.

## 4 DATASETS

Access to accurately labeled data is crucial to enhance our supervised NLP algorithms for the initial three tasks in our pipeline. After conducting an exhaustive review of existing literature, we have pinpointed a meticulously curated dataset concerning privacy policies (OOP-115) dataset (Wilson et al., 2016), which we denote as Dataset 1. Moreover, to enrich our analysis for the final task (Policy Analysis), we collected various policies from various technology companies, which we denote as Dataset 2.

### 4.1 Dataset 1

In this study, we used the publicly available OOP-115 dataset (Wilson et al., 2016), a comprehensive resource on privacy policy. Designed by Wilson et al. (Wilson et al., 2016), this dataset comprises a diverse collection of 115 privacy policies spanning 15 sectors, including arts, shopping, business, and news. It encapsulates a wealth of information, including the data practices shown in Table 1, and each data practice category has attributes such as data type, purpose



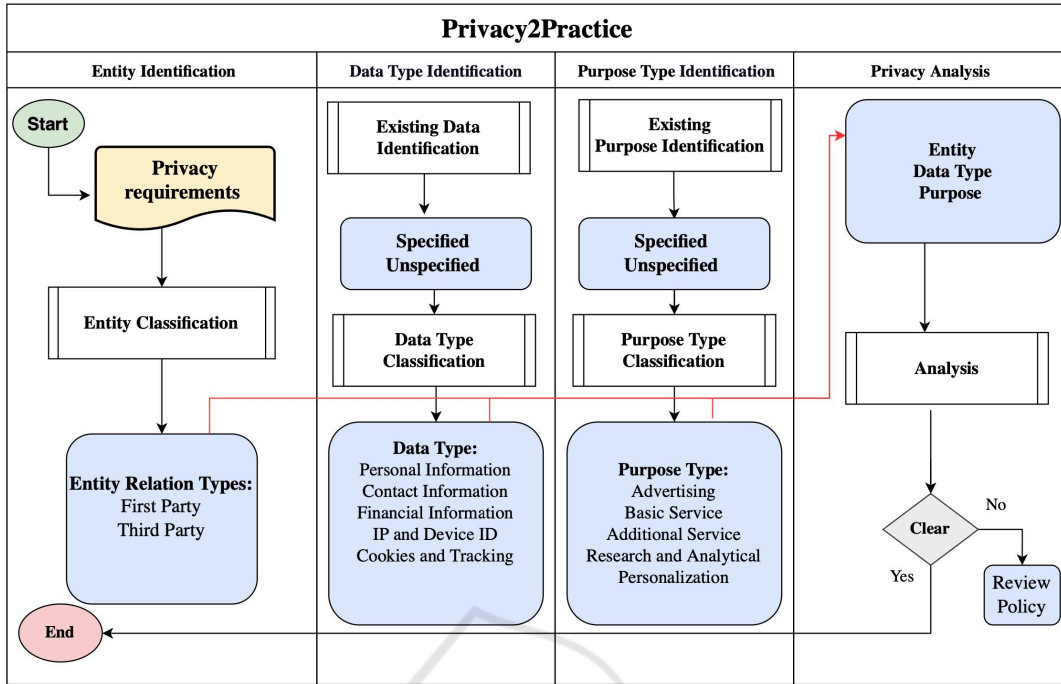


Figure 1: A Cross Functional Flowchart Presents the Automated Privacy Policy Analysis (*Privacy2Practice*).

type, and others meticulously curated by proficient annotators (Wilson et al., 2016). The final annotation scheme comprises a comprehensive breakdown into ten distinct categories of data practices, as illustrated in Table 1 below. Consequently, we employ this dataset as the cornerstone of our investigation, focusing on tasks that aim to identify entities, data types, and purpose types, as elaborated in Sections 5.1,5.2,5.3.

The OPP-115 data set comprises a wide array of data practice categories as shown in Table 1. However, our focus was specifically on categories and attributes aligning with our research objectives. In particular, we honed in on entity categories, specifically first and third parties, and the attributes of data type and purpose type. These chosen categories were carefully extracted from the OPP-115 dataset to form the basis of our training data.

Table 2 provides a detailed breakdown of the category distribution within the training subset of the OPP-115 dataset. It presents the number of samples associated with three key aspects: *entities*, *data type*, and *purpose type*. The *entities* category, which includes both *first party* and *third party*, comprises 1,790 samples. The *data type* category contains 801 samples, while the *purpose type* category consists of 827 samples. This distribution offers valuable insights into the data practices utilized from the OPP-115 dataset, serving as a foundational dataset for models training and evaluation.

Table 1: Categories of Data Practices Defined by Wilson et al. (Wilson et al., 2016).

Category	Description
First Party (Collection)	Clarifies the entity who processes and collects the data and the rationale behind the service provider’s acquisition of user information.
Third Party (Sharing)	Elaborates on the mechanisms through which user information is shared with or gathered by third-party entities.
User Choice and Control	Outlines the choices and control options available to users regarding their information.
User Access, Edit, and Deletion	Describes whether and how users can access, modify, or delete their information.
Data Retention	Indicates the duration for which user information is retained.
Data Security	Explains the measures in place to safeguard user information.
Policy Change	Addresses how users will be notified about changes to the privacy policy.
Do Not Track	Specifies if and how Do Not Track signals for online tracking and advertising are respected.
International and Specific Audiences	Covers practices relevant to specific user groups such as children, Europeans, or California residents.
Other	Includes additional sub-labels for introductory or general text, contact information, and practices not falling under the other defined categories.

## 4.2 Dataset 2

We collect comprehensive privacy policies from various companies operating within the technology sec-

Table 2: Distribution of the Training Dataset.

Data Practices	Sample Count
Entities ( First Party- Third Party)	1790
Data Types	801
Purpose Types	827

tor. This data set forms the cornerstone of our analysis. To acquire these data, we developed a crawler capable of extracting text from websites that host privacy policy statements. Subsequently, we systematically parsed each document, breaking them into discrete paragraphs and statements to facilitate our analysis.

We followed a set of predefined criteria to guide the selection of privacy policies throughout the data collection process. Based on these criteria, companies were categorized into two distinct groups. The first group comprises well-established organizations that are generally recognized for demonstrating compliance with regulations such as the GDPR and the CCPA, which may contribute to more detailed disclosures of their data practices (Wong et al., 2023). In contrast, the second group consists of smaller or less prominent companies that may not provide the same level of detail regarding their data practices. To preserve the confidentiality of these organizations, their names have been replaced with symbolic identifiers. We must emphasize that we used this data set in our analysis task, as discussed in Section 6. Table 3 provides a comprehensive breakdown, illustrating the number of statements collected from each company.

Table 3: Sentence Count in Companies’ Privacy Policies.

Group	Companies	Sentences Count
Group 1	Google	239
	Microsoft	243
Group 2	Company_X	59
	Company_Y	33

## 5 DOWN-STREAM NLP TASKS

The tasks for automating the analysis of privacy policies using the *Privacy2Practice* framework are structured into three primary downstream NLP tasks, each addressing critical aspects of data practices outlined in privacy policies. Below, we delve into the methodologies and approaches used for these tasks.

### 5.1 Entity Type Identification

Our initial focus revolves around identifying who is collecting the data and with whom it is shared, referred to as first party and third party entities, respectively. Thus, this task primarily centers on identify-

ing the entities engaged in data collection and sharing, as delineated within privacy policies, commonly referred to as first party and third party entities.

In this task, we propose a two-level classification approach to precisely identify entities within privacy policy sentences. Firstly, we classify the entity type, aiming to differentiate between first party and third party entities, a process we term as *Entity Identification*. Once this foundational classification is established, we develop an algorithm to identify the specific entities mentioned within each sentence, a process we refer to as *Entity Detection*.

*Entity Identification*: The objective of this task is to categorize segments of text based on whether they refer to the entity responsible for collecting the data (referred to as the first party) or if they pertain to the external entity from which the data is sourced and subsequently obtained by the first party (referred to as the third party). To accomplish this goal, we adopt a binary text classification methodology. This entails designing a model that can effectively classify each text segment into one of two categories: referencing the first party or the third party. Given the inherent challenge posed by the limited availability of training data, we employ a Few-Shot Learning approach. By utilizing this approach, we develop a binary classification model that effectively classifies text segments despite having access to only a limited dataset for training. In our approach, we leverage the *SetFit* model to build a classification model capable of accurately categorizing text segments into either first party or third party entities. The *SetFit* model has demonstrated notable efficiency and effectiveness in similar text classification tasks, making it a suitable choice for our purpose. To evaluate the performance of our entity identification process, we rely on standard evaluation metrics, including *precision*, *recall*, and *F1-score*.

*Entity Detection*: In this step, our focus is on defining that *first party* and *third party* entities within text segments. Our approach entails identifying *first party* entities by recognizing them as the subject or first person within a sentence, and it is followed by discerning their associated actions. This involves analyzing the syntactic structure of the text to locate instances where the subject of the sentence corresponds to the first party mentioned. Conversely, for identifying *third party* entities, our methodology centers around pinpointing the name of organizations within the text. These organizations typically represent the third parties with whom data sharing agreements are established. To carry out this analysis, we employ a dependency parser applied to the preprocessed text data. The dependency parser enables us to analyze the grammatical structure and relationships between

words in the text, providing valuable insights into how different elements within the text are connected and interact. We utilized *SpaCy* (Honnibal and Montani, 2017), which is an NLP library, to apply the dependency parser. *SpaCy* offers robust capabilities for linguistic analysis, including dependency parsing. By leveraging *SpaCy*, we can extract detailed syntactic information from the text data, facilitating the identification of the *first party* and *third party* entities within text segments.

**Experimental Results.**

*Entity Classification:* Fine-tuning large language models for text classification involves training them on a curated dataset specifically tailored to the task. In our case, the goal was to classify text into two distinct categories: first-party and third-party entities. To achieve this, we conducted an experiment by fine-tuning the *SetFit* model (Tunstall et al., 2022), configuring it as a binary classifier using labeled data representative of first-party and third-party entities.

The fine-tuning process utilized the following hyperparameters:

- Training Data: 1,000 samples
- Testing Data: 500 samples
- Sentence Transformer: paraphrase-MiniLM-L3-v
- Number of Epochs: 10

Based on the experimental results shown in Table 4, our model demonstrated an exceptional ability to accurately identify first-party entities within the text, achieving an impressive F1 score of 97%. In addition, it exhibited notable proficiency in the detection of third-party entities, achieving an F1 score of 93%. These high-performance metrics underscore the effectiveness of the model and highlight its promising potential for reliable text classification tasks.

Table 4: Performance of the Entity Classification.

Entity Type	P(%)	R(%)	F <sub>1</sub> (%)
First Party	100%	95%	97%
Third Party	87%	100%	93%

*Entity Extraction:* To extract relevant entities, we utilized a dependency parser, leveraging the advanced capabilities of *SpaCy*. Our objective was to identify the *first party* responsible for data collection and the *third party* with whom the data is shared. The process was carried out using the following steps:

1. *Tokenization and Iteration:* The input sentence is first tokenized into individual tokens, forming the foundation for detailed analysis. Each token is then iterated over to facilitate thorough entity identification and extraction.

2. *Identifying First Party:* Tokens are inspected to determine whether they correspond to first-party pronouns (e.g., I, we, us, it). These pronouns typically represent the speaker or the entity collecting the data. Dependency relationships between tokens are analyzed to verify whether the first-party pronoun functions as the subject of a verb (nsubj). If the pronoun is not directly linked as the subject of a verb, additional analysis is conducted to determine whether it serves as the subject of an auxiliary verb (aux) that functions as the head of a main verb. This approach ensures accurate identification of first-party entities in varying syntactic contexts.

3. *Identifying Third Party:* Each token is examined to identify entities tagged as organizations (labeled "ORG") using Named Entity Recognition (NER). The corresponding text for these identified organization entities is then extracted for further analysis.

The following example demonstrates entity extraction from privacy policies using *SpaCy*. A sample privacy policy sentence is analyzed to identify the *first party* (the entity collecting the data) and the *third party* (the entity with whom the data is shared). Consider the statement:

"We may collect data that include, but are not limited to, weight, steps, and height. Based on the initial setup on iOS, enable us to link to Apple Health."

In this example: - **First Party:** We, us - **Third Party:** Apple Health

**Comparative Performance Analysis.**

A comparison of our model with prior research, including methodologies employing CNN and XLNET models (Hamdani et al., 2021; Harkous Hamza et al., 2018), highlights the superiority of the proposed approach. As shown in Table 5, the *SetFit* model outperformed these approaches, achieving significantly higher F1-scores for both *first-party* and *third-party* entity identification.

By systematically analyzing the dependency structures of sentences and identifying key linguistic patterns, our algorithm effectively pinpoints instances of both *first-party* and *third-party* entities. These results underscore the robustness of the *SetFit* model, demonstrating its capability as a few-shot text classifier tailored to the complexities of privacy policy analysis.

**5.2 Data Type Identification**

This task focuses on classifying the diverse data instances mentioned in privacy policies into specific

Table 5: F1-score comparison between our approach and prior methods.

Entity Type	CNN (Ham-dani et al., 2021)	XLENT (Ham-dani et al., 2021)	Polisis (CNN) (Hark- ous Hamza et al., 2018)	Our Ap- proach (SetFit)
First Party	76%	83%	80%	97%
Third Party	78%	81%	81%	93%

categories, each representing a distinct type of information. To achieve this, we leverage a taxonomy inspired by OPP-115 [26], categorizing the data into key groups such as personal information, financial details, device-related data, cookies and tracking technologies, user online activities, and unspecified instances. The detailed classification categories are outlined below:

1. **Personal Information:** Includes personally identifiable information (PII), such as names, addresses, social security numbers, and biometric data.
2. **Financial Information:** Covers data related to financial transactions, account details, payment methods, credit card numbers, and other sensitive financial data.
3. **Computer Information:** Pertains to data linked to electronic devices, including device identifiers, Internet Protocol (IP) addresses, and other unique digital identifiers.
4. **Cookies and Tracking Elements:** Encompasses data associated with cookies and tracking technologies used to monitor user behavior and preferences.
5. **User Online Activities:** Reflects user interactions with online services and websites, including activities tracked and recorded by these platforms.
6. **Unspecified:** Applies to data instances that do not clearly fall under any of the above categories or are not explicitly described in the text.

To systematically identify and categorize data instances within privacy policies, we adopt a two-tiered classification approach as follows:

1. **Existence Checking Phase:** This initial phase assesses whether a specific data type is present in a given text segment. Each segment is categorized as either specified or unspecified based on the presence of identifiable data types. If data existence is confirmed, the segment progresses to the next phase.
2. **Data Type Classification Phase:** In this phase, segments containing identifiable data are analyzed and assigned to one of the predefined categories. This step ensures a detailed and accurate align-

ment of data instances with their appropriate classifications.

By employing this structured approach, we ensure precision and consistency in identifying and categorizing data instances within privacy policies. The two-tiered methodology enhances the accuracy and efficiency of our analysis, providing a comprehensive understanding of the data practices disclosed in these documents.

### Experimental Results.

Our study employed the *SetFit* Few-Shot Learning approach to tackle the challenge of classifying data segments within privacy policies. Leveraging the OPP-115 dataset, which categorizes data types as either specified or unspecified, we established a solid foundation for our analysis. This dataset enabled us to extract the data type attribute assigned to each segment and utilize it as the segment's label, facilitating a systematic classification process.

In the initial phase, referred to as *Existence Checking*, we fine-tuned *SetFit* to determine whether each segment represented a specified or unspecified data type. This phase laid the groundwork for further classification by identifying relevant segments requiring detailed analysis. The performance of the model during this phase was evaluated using key metrics, including *precision*, *recall*, and *F1-score*, with the results presented in Table 6.

Building upon the outcomes of the first phase, the subsequent phase, *Data Type Classification*, focused on assigning each identified segment to one of the predefined data type categories. These categories included personal information, financial records, computer-related data, and tracking elements. By further fine-tuning *SetFit*, we enhanced its ability to accurately classify segments into these specific categories. The performance metrics for this phase, including *precision*, *recall*, and *F1-score*, are summarized in Table 7.

For our experiments, the fine-tuning process utilized the following hyperparameters::

- Training Data: 600 samples
- Testing Data: 200 samples
- Sentence Transformer: paraphrase-MiniLM-L3v
- Number of Epochs: 10

The results, presented in Table 7, highlight the model's performance in detecting existing data within text segments, achieving an overall F1-score of 82%. Furthermore, the model demonstrated exceptional proficiency in classifying specific data types. It achieved F1-scores of 96% for personal information, 94% for financial information, and 98% for computer



information. However, its performance in identifying user online activities and cookies/tracking data was comparatively lower, with F1-scores of 62% and 69%, respectively.

It is important to note that prior studies have not focused on specifying the types of data handled or shared within privacy policies, nor have they explored automated methods for identifying such data types. By undertaking this task, our research makes a substantial contribution to the field by introducing an automated approach to categorize data types within privacy policies. This contribution marks a significant step forward in understanding and addressing the complexities of privacy policy analysis in an increasingly data-driven world.

Table 6: Performance of the Data Existence Phase.

Classes	P	R	$F_1$
Specified Data Type	80%	84%	82%
Unspecified Data Type	88%	84%	86%

Table 7: Performance of the Data Type Classification Phase.

Classes	P	R	$F_1$
Personal Information	94%	98%	96%
Financial Information	93%	95%	94%
Computer Information	95%	100%	98%
Cookies and Tracking Elements	75%	64%	69%
User Online Activities	80%	50%	62%

### 5.3 Purpose Type Identification

This task aims to classify the purpose types mentioned within privacy policies into specific categories, each representing a distinct purpose. To achieve this, we focus on key categories inspired by those defined in OPP-115 (Wilson et al., 2016), which include: advertising, basic services and features, additional services and features, analytical and research, personalization and customization, and legal requirements.

To address this task, we employed Few-Shot Learning approach for multi-class classification. Our objective was to identify the purpose of data practices within segments of privacy policies and categorize them into the following distinct classes:

1. Advertising: Encompasses data practices related to promotional activities, market research, and consumer engagement strategies designed to enhance brand visibility and customer outreach.
2. Basic Services and Features: Includes data practices integral to the delivery of core functionalities and primary offerings within the organization.
3. Additional Services and Features: Covers data practices pertaining to supplementary services and features that enhance the user experience and complement the organization's core offerings.

4. Analytical and Research: Refers to data practices focused on data analysis, research initiatives, and insights generation to better understand market trends, consumer behavior, and organizational performance metrics.
5. Personalization and Customization: Involves data practices aimed at tailoring services, products, and user experiences to individual preferences and requirements, thereby enhancing personalization capabilities.
6. Legal Requirements: Represents data practices undertaken to comply with legal obligations related to the collection, use, or sharing of personal information online.
7. Unspecified: Applies when no specific purpose is mentioned within the segment.

We utilized the Few-Shot Learning approach, specifically the SetFit model, to develop a classification approach capable of accurately assigning purpose types to their respective categories. The process involved a two-phase classification strategy as follows:

**Existence Checking Phase:** In this initial phase, we determined whether a specific purpose was mentioned within the segment. Segments were classified as either containing a specified purpose or falling under the "unspecified" category. If a specified purpose was identified, the segment proceeded to the next phase. This step ensured that only relevant segments were included in the purpose type classification. For the Existence Checking Phase, the SetFit model was fine-tuned to classify segments as either specified or unspecified. Evaluation metrics, including precision, recall, and F1-score, are presented in Table 8.

**Purpose Type Classification Phase:** In this phase, segments identified as containing a specified purpose were assigned to one of the predefined categories. The classification aimed to discern the nature of the purpose and accurately map it to its corresponding category. In the Purpose Type Classification Phase, we further fine-tuned the SetFit model to classify segments into the predefined purpose categories (advertising, basic services and features, additional services and features, analytical and research, personalization and customization, and legal requirements). The performance of this classification is assessed using *precision*, *recall*, and *F1-score*, as detailed in Table 9.

The model was fine-tuned with the following hyperparameters:

- Training Data: 600 samples
- Test Data: 200 samples
- Sentence Transformer: paraphrase-MiniLM-L3v
- Number of Epochs: 10

### Experimental Results.

Based on the results of our experiment, we observed that our model demonstrated capability in discerning existing purpose in the text segment with a 90% F-1 score as shown in Table 8. Furthermore, it displayed proficiency in identifying advertising, basic service, additional services, and legal requirements, achieving an F1-score of 88%, 80%, 95% and 81%, respectively, as shown in Table 9. However, when it came to identifying analytical research and personalization purposes the model achieved a lower F1 score of 72% and 64% respectively.

Table 8: Performance of the Purpose Existence Phase.

Classes	P	R	$F_1$
Specified Purpose	91%	90%	90%
Unspecified Purpose	91%	92%	92%

Table 9: Performance of the Purpose Type Classification Phase.

Classes	P	R	$F_1$
Advertising	87%	89%	88%
Basic services and feature	86%	75%	80%
Additional services and feature	100%	91%	95%
Analytical and research	68%	76%	72%
Personalization and customization	73%	57%	64%
Legal requirement	81%	81%	81%

It is noteworthy that there has been no prior research dedicated to specifying the purposes for collecting or sharing data within privacy policies, and there has been no study on the automated identification of such purpose types. Thus, this task constitutes another notable contribution to the automated privacy policy analysis.

## 6 PRIVACY POLICY ANALYSIS

This section provides an in-depth analysis of privacy policies, emphasizing their effectiveness in achieving transparency, a cornerstone of trust and regulatory compliance in data practices. Transparency, as introduced earlier, is the principle that ensures users can understand and evaluate how their personal data is collected, shared, and used. Privacy policies serve as critical tools for communicating these practices, and their transparency directly impacts user trust and compliance with regulations such as GDPR and CCPA (Adjerid et al., 2013; Pan and Zinkhan, 2006).

Transparent privacy policies provide clarity, consistency, and comprehensiveness in the disclosure of data practices. In this study, transparency is evaluated through the lens of consistent disclosure, which

measures the extent to which privacy policies clearly and accurately represent data collection, sharing, and usage practices. Using the *Privacy2Practice* framework, this analysis examines whether privacy policies meet the criteria for transparency by aligning their disclosures with regulatory requirements.

To assess transparency, privacy policies from two distinct groups of companies were analyzed.

- **Group 1:** Two large technology companies were selected, including Google <sup>1</sup> and Microsoft <sup>2</sup>.
- **Group 2:** Two small organizations were selected for this study and are referred to as Company\_X and Company\_Y to preserve their confidentiality.

### Evaluation of Group 1's Privacy Policy Transparency.

*First Party Transparency:* Both Google and Microsoft demonstrate clear responsibility for data collection, with their privacy policies explicitly stating that the companies themselves handle this task.

*Third Party Transparency:* Google restricts data sharing to its own services and applications, providing users with clarity on third-party involvement. Microsoft similarly limits data sharing to its ecosystem, such as media-related services.

*Data Type Transparency:* The data types collected by both companies, including personal information, cookies, and online activity, are identified transparently. Neither company collects financial information and this exclusion is clearly communicated.

*Purpose Transparency:* Both companies provide explicit justifications for data collection. Google focuses on improving user experiences and enabling research and analytics, while Microsoft emphasizes service delivery and additional features.

Thus, Group 1 sets a high standard for transparency by ensuring clarity of roles, with a clear delineation of first-party and third-party responsibilities, which helps users understand who is collecting and handling their data. Additionally, their comprehensive disclosures provide detailed identification of the types of data collected, enabling users to grasp the scope and nature of the information being handled. Furthermore, Group 1 offers transparent explanations of the purposes behind their data practices, ensuring users are aware of why their data is being collected and how it is utilized. By aligning their disclosures with transparency principles, Group 1 not only demonstrates compliance with regulatory standards but also fosters trust among its users.

### Evaluation of Group 2's Privacy Policy Transparency.

<sup>1</sup><https://policies.google.com/privacy?hl=en-US>

<sup>2</sup><https://privacy.microsoft.com/en-us/privacystatement>

*First Party Transparency:* Company\_X and Company\_Y both declare their responsibility for data collection.

*Third Party Transparency:* Company\_X explicitly names third parties like MailChimp, providing some level of transparency. In contrast, Company\_Y fails to disclose third-party details, leaving users uncertain about data sharing practices.

*Data Type Transparency:* Company\_X identifies personal information, cookies, and data of online activity. Company\_Y collects similar data, but lacks clarity in explaining the inclusion of sensitive financial information.

*Purpose Transparency:* While Company\_X offers some information on the purposes of data collection, such as research and service enhancement, company\_Y provides limited details, creating ambiguity in its data practices.

Thus, Group 2 reveals significant transparency gaps that undermine the clarity and trustworthiness of their privacy practices. One major issue is the ambiguity surrounding third-party disclosures, as the policies provide limited information about the entities with whom data is shared. Additionally, there is inadequate clarity regarding the purposes of data collection and sharing, with insufficient explanations for why user data is being handled in specific ways. These shortcomings hinder users' ability to fully understand the data practices of Group 2 companies, ultimately eroding trust and raising concerns about transparency and compliance.

Thus, this analysis highlights the pivotal role that transparency plays in privacy policies, serving as a foundation for building trust and ensuring regulatory compliance. Group 1 exemplifies best practices by embracing transparency principles through clearly defined roles for first and third parties, detailed disclosures of data types, and purpose-driven explanations for data collection and sharing. In contrast, Group 2 illustrates the challenges faced in achieving comparable levels of transparency, often lacking clarity and consistency. These findings emphasize the critical need for automated frameworks such as *Privacy2Practice* to evaluate and improve the transparency of privacy policies, ensuring that users are fully informed and their data are adequately protected.

## 7 CONCLUSION

Privacy policies are critical in defining how organizations handle user data, including its collection, storage, sharing, and protection. However, these policies

are often expressed in dense natural language, making them difficult for users to comprehend and evaluate for transparency and compliance with privacy regulations.

To address these challenges, this paper introduced *Privacy2Practice*, an automated framework designed to analyze and extract key data practices embedded in privacy policy documents. Using advanced NLP techniques and a Few-Shot Learning approach, the framework focuses on three primary dimensions of data practices: (1) entities involved in data collection and sharing, (2) data types handled, and (3) purposes for data collection and sharing. Using four tasks, including entity type identification, data type identification, purpose type identification, and privacy analysis, *Privacy2Practice* ensures a systematic and structured approach to analyzing privacy policies.

The strength of the proposed approach lies in its use of a Few-Shot Learning algorithm that excels in text classification tasks with minimal training data. For entity type identification, the model achieved F1 scores of 97% and 93% for first-party and third-party entities, respectively, demonstrating exceptional precision. Similarly, the framework accurately categorized the types of data (F1 score: 82%) and determined the purposes of data collection and sharing (F1 score: 90%). This level of precision highlights the ability of the framework to effectively address the complexity of privacy policy analysis, providing actionable information to organizations and end users alike.

Our future work will focus on enhancing the framework by detecting and resolving contradictions between policy claims, actual operational practices, and software behavior. Additionally, we aim to integrate automated compliance mechanisms within application behaviors to ensure continuous alignment with privacy regulations.

## ACKNOWLEDGEMENTS

This work was supported in part by funding from NSF under Award Numbers DMS 2123761, CNS 1822118, CNS 2335687, NIST, ARL, Statnett, AML, NewPush, and Cyber Risk Research.

Also, this research is supported by a grant (No. CRPG-00-0000) under the Cybersecurity Research and Innovation Pioneers Initiative, provided by the National Cybersecurity Authority (NCA) in the Kingdom of Saudi Arabia.

## REFERENCES

- Adjerid, I., Acquisti, A., Brandimarte, L., and Loewenstein, G. (2013). Sleights of privacy: Framing, disclosures, and the limits of transparency. In *Proceedings of the ninth symposium on usable privacy and security*, pages 1–11.
- Alfawzan, N., Christen, M., Spitale, G., and Biller-Andorno, N. (2022). Privacy, data sharing, and data security policies of women’s mhealth apps: Scoping review and content analysis. *JMIR mhealth uhealth* 10, 5 (may 2022), e33735.
- Amaral, O., Abualhaija, S., and Briand, L. (2023). MI-based compliance verification of data processing agreements against GDPR. In *IEEE 31st International Requirements Engineering Conference (RE)*, pages 53–64, Germany. IEEE.
- Andow, B., Mahmud, S. Y., Wang, W., Whitaker, J., Enck, W., Reaves, B., Singh, K., and Xie, T. (2019). {PolicyLint}: investigating internal privacy policy contradictions on google play. In *28th USENIX security symposium*, pages 585–602, Santa Clara, CA, USA. USENIX Association.
- Andow, B., Mahmud, S. Y., Whitaker, J., Enck, W., Reaves, B., Singh, K., and Egelman, S. (2020). Actions speak louder than words: {Entity-Sensitive} privacy policy and data flow analysis with {PoliCheck}. In *29th USENIX Security Symposium*, pages 985–1002, Online. USENIX Association.
- Annas, G. J. (2003). Hipaa regulations: a new era of medical-record privacy? *New England Journal of Medicine*, 348:1486.
- Antón, A. I., Earp, J. B., He, Q., Stufflebeam, W., Bolchini, D., and Jensen, C. (2004). Financial privacy policies and the need for standardization. *IEEE Security & privacy*, 2(2):36–45.
- Bui, D., Yao, Y., Shin, K. G., Choi, J.-M., and Shin, J. (2021). Consistency analysis of data-usage purposes in mobile apps. In *Proceedings of ACM SIGSAC Conference on Computer and Communications Security*, pages 2824–2843, Republic of Korea. ACM.
- Cohen, J. E. (2008). Privacy, visibility, transparency, and exposure. *The University of Chicago Law Review*, 75(1):181–201.
- de la Torre, L. (2018). A guide to the california consumer privacy act of 2018. Available at SSRN 3275571.
- Elluri, L., Joshi, K. P., and Kotal, A. (2020). Measuring semantic similarity across EU GDPR regulation and cloud privacy policies. In *2020 IEEE International Conference on Big Data (Big Data)*, pages 3963–3978. IEEE.
- Erkkilä, T. (2020). Transparency in public administration. In *Oxford research encyclopedia of politics*. Oxford.
- Hamdani, R. E., Mustapha, M., Amariles, D. R., Troussel, A., Meeüs, S., and Krasnashchok, K. (2021). A combined rule-based and machine learning approach for automated GDPR compliance checking. In *Proceedings of the Eighteenth International Conference on Artificial Intelligence and Law*, pages 40–49, São Paulo, Brazil.
- Harkous Hamza, K. F., Lebre, R., Schaub, F., Shin, K. G., and Aberer, K. (2018). Polisis: Automated analysis and presentation of privacy policies using deep learning. In *27th USENIX Security Symposium*, pages 531–548, Baltimore, MD, USA. USENIX Association.
- Honnibal, M. and Montani, I. (2017). SpaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing. Accessed: 2023.
- Jensen, C. and Potts, C. (2004). Privacy policies as decision-making tools: an evaluation of online privacy notices. In *Proceedings of the SIGCHI conference on Human Factors in Computing Systems*, pages 471–478, Vienna, Austria.
- Mousavi Nejad, N., Scerri, S., and Lehmann, J. (2018). Knight: Mapping privacy policies to GDPR. In *Knowledge Engineering and Knowledge Management: 21st International Conference, EKAW 2018, Nancy, France, November 12-16, 2018, Proceedings 21*, pages 258–272. Springer.
- Neupane, S., Tazi, F., Paudel, U., Baez, F. V., Adamjee, M., De Carli, L., Das, S., and Ray, I. (2022). On the data privacy, security, and risk postures of iot mobile companion apps. In *IFIP Annual Conference on Data and Applications Security and Privacy*, pages 162–182. Springer.
- Nguyen, T. T., Backes, M., Marnau, N., and Stock, B. (2021). Share first, ask later (or never?) studying violations of {GDPR’s} explicit consent in android apps. In *30th USENIX Security Symposium (USENIX Security 21)*, pages 3667–3684.
- Pan, Y. and Zinkhan, G. M. (2006). Exploring the impact of online privacy disclosures on consumer trust. *Journal of retailing*, 82(4):331–338.
- Parnami, A. and Lee, M. (2022). Learning from few examples: A summary of approaches to few-shot learning. *arXiv preprint arXiv:2203.04291*.
- Sunkle, S., Kholkar, D., and Kulkarni, V. (2015). Toward better mapping between regulations and operations of enterprises using vocabularies and semantic similarity. *Complex Systems Informatics and Modeling Quarterly*, 5:39–60.
- Tankard, C. (2016). What the GDPR means for businesses. *Network Security*, 2016(6):5–8.
- Tunstall, L., Reimers, N., Jo, U. E. S., Bates, L., Korat, D., Wasserblat, M., and Pereg, O. (2022). Efficient few-shot learning without prompts. *arXiv preprint arXiv:2209.11055*.
- Wilson, S., Schaub, F., Dara, A. A., Liu, F., Cherivirala, S., Leon, P. G., Andersen, M. S., Zimmeck, S., Sathiyendra, K. M., Russell, N. C., et al. (2016). The creation and analysis of a website privacy policy corpus. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*, pages 1330–1340, Berlin, Germany.
- Wong, R. Y., Chong, A., and Aspegren, R. C. (2023). Privacy legislation as business risks: How GDPR and CCPA are represented in technology companies’ investment risk disclosures. *Proceedings of the ACM on Human-Computer Interaction*, 7(CSCW1):1–26.
- Zaeem, R. N. and Barber, K. S. (2020). The effect of the GDPR on privacy policies: Recent progress and future promise. *ACM Transactions on Management Information Systems (TMIS)*, 12(1):1–20.