TLDR: Deep Learning-Based Automated Privacy Policy Annotation with Key Policy Highlights

Abdulrahman Alabduljabbar
University of Central Florida
jabbar@knights.ucf.edu

Ahmed Abusnaina
University of Central Florida
ahmed.abusnaina@knights.ucf.edu

Ülkü Meteriz-Yıldıran
University of Central Florida
meteriz@knights.ucf.edu

David Mohaisen
University of Central Florida
mohaisen@ucf.edu

ABSTRACT

Privacy policies are the primary channel where service providers inform users about their data collection and use practices. However, privacy policies are often long and lack any specific structure. The average user struggles to understand their contents and usually skips them, regardless of their importance. Moreover, privacy policies may lack information on critical practices used by the service providers, such as data collection, use disclosure, tracking, and access. We tackle these challenges by introducing TLDR, a machine learning-based automated ensemble of privacy policy classifiers, for (i) categorizing the content into nine privacy policy categories with high performance and (ii) detecting missing information in the privacy policies. Towards addressing the length of the privacy policies, TLDR labels each paragraph in a policy by its content class, which enables users to focus on paragraphs of interest, such as paragraphs with information regarding data collection or tracking practices used by the service operators. TLDR reduces the average reading time by 39.14% by reducing the presented information to users. This process results in an increased understanding of the privacy policies by 18.84%. TLDR reduces the number of paragraphs and words required to be read by the user. This, in turn, reduces the required efforts to understand the service operator's practices.

CCS CONCEPTS

• **Security and privacy → Usability in security and privacy**.

KEYWORDS

Privacy, Machine Learning, Privacy Policy, Natural Language Processing.

ACM Reference Format:
Abdulrahman Alabduljabbar, Ahmed Abusnaina, Ülkü Meteriz-Yıldıran, and David Mohaisen. 2021. TLDR: Deep Learning-Based Automated Privacy Policy Annotation with Key Policy Highlights. In *Proceedings of the 20th Workshop on Privacy in the Electronic Society (WPES '21), November 15, 2021, Virtual Event, Republic of Korea.* ACM, New York, NY, USA, 16 pages. https://doi.org/10.1145/3463676.3485608

WPES '21, November 15, 2021, Virtual Event, Republic of Korea

ACM ISBN 978-1-4503-8527-5/21/11…$15.00
https://doi.org/10.1145/3463676.3485608

1 INTRODUCTION

The privacy policy statements are legal statements that inform Internet users about websites and businesses data collection and information usage practices. With a semi-universal enforcement of the General Data Protection Regulation (GDPR), privacy policies became more and more elaborate and technical about how Personally Identifiable Information (PII) is collected, stored, handled, and distributed [26]. Those policies are long and complex, and it is argued that the ordinary user may not thoroughly understand the context of the policy, nor the website's actual practices [30].

Although service providers are continuously improving the readability and comprehensiveness of their policies and the disclosure of their practices, privacy policies remain difficult to understand [7, 10, 19, 30, 34]. It has been estimated that it would take the average user 201 hours to read the privacy policies encountered per year [30]. Moreover, it is unclear whether these policies are even sufficient to address the security and privacy aspects of services usage. As such, privacy and data practices of the service providers may be hidden within long, vague, and ambiguous policies, which may not be clearly disclosed to users [37, 51, 52].

A key challenge in this space is the lack of a standard format in privacy policies, which leads to ambiguity. Especially, users would be overwhelmed by both the breadth (i.e., number of policies) and depth (i.e., individual policy complexity) of those policies. While several attempts have been made towards making privacy policies easier to read, by introducing the Privacy Preference Project (P3P) [16] in 2002, privacy policies still lack a standard format as of the moment of writing this work. Motivated by that, recent studies [4, 5, 14, 15, 21, 28, 51, 54, 56] have worked on annotating privacy policies, manually and automated, to summarize and present the critical security and privacy aspects included in the privacy policy. Using state-of-the-art natural language processing and deep learning techniques, recent works [4, 14, 15, 21, 28, 51, 56] effectively annotate the privacy policies contents, on both sentence-level and segment-level, with high performance.

In this work, we investigate several annotation techniques for a practical automation of policy annotation, and to reduce the time and efforts required for such a task. The goal of our annotation is to provide users with easy-to-interpret high-level annotations on whether various privacy policies they encounter in their daily life meet certain requirements with respect to a broad set of privacy

and security expectations. Our pipeline, called TLDR, employs advances in deep representation and machine learning. In particular, we built an ensemble of classifiers using six word representation techniques; word mapping [28], count vectorizer [42], TF-IDF [20], Doc2Vec [24], Universal Sentence Encoder (USE) [8], and WordPiece [53], and learning algorithms; Logistic Regression (LR) [48], Support Vector Machine (SVM) [13], Random Forest (RF) [22], Convolutional Neural Networks (CNN) [55], Deep Neural Networks (DNN) [40], and Bidirectional Encoder Representations from Transformers (BERT) [18], for automating privacy policy annotation.

TLDR operates at the paragraph (segment) level, and is trained on nine categories highlighting different uses typically found in the privacy policies. The ensemble outputs a binary decision for each category, positive and negative. A positive outcome indicates that a segment contains information on the privacy policy category, while a negative outcome indicates that a segment does not contain such information.

Through experiments on a widely used dataset, TLDR achieves high performance in categorizing privacy policy practices, with an average F_1 score of 91%, and can highlight important segments within a privacy policy. Through a user study, we show that TLDR reduces the reading time by 39.14%. Moreover, by eliminating unnecessary information in the policy statements, TLDR improves their understandability by 18.84%. TLDR also is shown effective in highlighting critical information of the privacy policy, and its extracted statements are shown to be preferred over the original policies in 67% of the times, as shown later in our user study.

Contribution. In this work, we deliver the following contributions:

(1) We propose TLDR, a pipeline that employs various deep privacy policy representation techniques and an automated ensemble of privacy policy classifiers, leveraging advances in machine learning and natural language processing (NLP) for policy representation and classification. TLDR achieves a state-of-the-art average F-1 score of 91%.
(2) Using TLDR, we analyze the privacy practices in Alexa top-10,000 websites, unveiling major issues in reporting user tracking and data security practices used in those popular websites. In particular, we analyze different privacy policy practices by websites of the same interest and topics.
(3) We develop a segment highlighting mechanism to reduce the number of segments that a user needs to read in order to cover certain privacy practices in a privacy policy. This, in turn, reduces the privacy policy average reading time by 39.14%, while preserving the critical aspects of the privacy practices.
(4) We conduct a user study to understand the effectiveness of TLDR in highlighting important segments within the privacy policy. In this user study, 67% participants expressed that they prefer to read the privacy policy with only the highlighted segments over the original policies, and 12% participants indicate "no preference" between the original and TLDR filtered privacy policies.

Organization. The background, including related work, is presented in section 2. The design of TLDR, including the used dataset and terminology, is discussed in section 3. In section 4, TLDR is evaluated across three dimensions: computationally-assessed quality of annotation and a contrast with state-of-the-art, a demonstration of TLDR's application in analyzing and understanding Alexa top-10,000 websites privacy policy practices, and a user study to show the effectiveness of TLDR in improving user experience, through better readibility of policies, reduction in their length, and improved time of reading. Concluding remarks are in section 5.

2 BACKGROUND

Privacy policies have emerged out of necessity at the dawn of the Internet. In this section, we provide a background on privacy policies, how they have evolved, and their significance.

Privacy Policy. Service providers are required to prepare a legal document, a privacy policy, revealing how they collect, use, disclose, and manage data from their clients in accordance with the various privacy laws [33, 35]. Particularly, the PII of users that is covered by the privacy law [1] is governed by the same law for their protection. The definition of PII is broad, and includes everything that can be used to infer an individual's identity, such as name, address, date of birth, marital status, contact information, medical history, travel itinerary, or even intentions to acquire goods and services (cookies, browser history, etc.) [1]. The protections provided to users and to their PII differ significantly from a country to another, as the governing laws change. However, those protections have been the subject of an intense recent public debate in light of the disclosure of various recent security breaches, data monetization practices, and user manipulation events (e.g., the 2016 US presidential election [23, 39], and the Cambridge Analytica scandal [9, 11, 31]).

2.1 Historical Background

The current privacy regulations for privacy policies date back to the infancy of the Internet. The initial steps for introducing privacy laws have been motivated by the fact that technological advances will significantly impact human rights, and society as a whole. As a result, one of the earliest actions was taken by the Council of Europe by recognizing the new threats introduced by advances in computing systems, in 1968 [44]. Subsequently, the Organisation for Economic Co-operation and Development (OECD), concerned by the growing trends in data leaving their jurisdiction by traveling out of the borders of member countries, has advocated for stricter laws requiring conventions for the protection of personal data taking into account the growing automatic processing capabilities. This OECD effort has resulted in Convention 108—the Convention for the Protection of Individuals with regard to Automatic Processing of Personal Data, which was introduced in 1981 [44].

To address the mounting concerns around the fairness, accuracy, and privacy of the private information collected about consumers concerning their credit, the Fair Credit Reporting Act (FCRA) was introduced in the United States in the late 1960s, and became a law in 1970. The law allowed consumers to review their credit files and correct errors, if they exist. In subsequent years, the United States Department of Health and Human Services drafted a policy named the Fair Information Practices (FIPs) in 1973 [44]. The committee after the FIPs also led structuring the Privacy Act in 1974 [44]. Today, FIPs outline various principles for the collection and use of private data. Those principles outline guidelines concerning the collection limitation, data quality, collection purpose, use limitation,

security safeguarding, the openness of the collection, individuals participation, and accountability.

General Data Protection Regulation. In 2018, the European Union's (EU) General Data Protection Regulation (GDPR), which was proposed in 2016, was introduced into a law [26, 44]. GDPR enforced businesses to reveal how consumers' data is handled and established requirements for using and sharing the collected data. Although GDPR is enforced in the EU, it still imposed fines for all misconduct whether the business is in the EU territory or not.

Such a broad enforcement of GDPR put other countries in motion as well, where major businesses have updated their privacy policies to address the GDPR law. For example, after the introduction of GDPR to mitigate any legal consequences, the word count of the privacy policy for Instagram has increased from 2,981 to 4,221, and from 2,881 to 5,617 in case of Wikipedia. However, this word count increase was not the only added complexity to the privacy policies: the language introduced to address aspects of GDPR have been "protectionist" of the businesses in the first place, resulting in an increased reading level, where high level of expertise is needed to read and comprehend the meaning of those privacy policies.

While the main purpose of GDPR and its enforcement is to simplify the process of data management for the consumers, recent studies demonstrated just the opposite: the significant increase in the complexity of policies as measured by the reading level and word count has made it more challenging for the ordinary users [43, 49].

2.2 Related Work

Privacy policies provide information about the data collection and processing practices followed by websites and service providers. While these websites are responsible for providing information about collection, storage, and management of users data, privacy policies may not be clear nor comprehensive. With limited time and expertise, the ordinary users may not be able to understand those policies even if they read them. Therefore, there is a need to analyze these policies to address various issues, including their readability and comprehensibility. To this end, the initial body of work on automatic privacy policy analysis makes the machine-readable policies a primary focus. The Platform for Internet Content Selection (PICS) [12] format is suggested for the web services for demonstrating their privacy policies. Similarly, The Platform for Privacy Preferences (P3P) [16] has been designed to provide a machine-readable language for explaining privacy policies to web users.

Instead of using machine-readable languages for privacy policies, natural language is more appropriate. Therefore, the research focus has shifted to extracting legal information from documents using NLP techniques to comprehend privacy policy aspects. Table 1 summarizes the efforts in this direction, compared with our work in terms of used techniques, annotation level, learning technique, dataset, and performance (measured by the F_1 score).

The prior work focusing on information extraction in privacy policies was initiated by a pilot study due to Ammar *et al.* [4] where they classified the information disclosure policies to law enforcement officials and the account closure policies. The authors show that natural language analysis is a feasible option. Subsequently, Constante *et al.* [14] performed a rule-based identification of users' data collected by web services and utilized NLP to evaluate the success of the identification. In their subsequent study, Constante *et al.* [15] also utilized machine learning approaches to determine whether a privacy policy gives sufficient details regarding certain privacy aspects of the associated web service or not.

Zimmeck *et al.* [56] present a browser extension that retrieves NLP-based analyses of policies from a privacy policy repository with the associated analysis. Similarly, Andow *et al.* and Zaeem *et al.* [5, 54] summarized the manually annotated privacy policies and found several contradictions in data collection and sharing practices. Harkous *et al.* [21] trained a privacy-centric language model with 130K privacy policies and proposed an automated framework for privacy policy analysis, achieving 88.4% accuracy in structured queries and 82% accuracy in the top-3 answers of free-form queries.

Recently, Wilson *et al.* [51] introduced a benchmark privacy policy dataset , OPP-115, annotated by competent law students annotators. OPP-115 has nine categories, where each paragraph is annotated based on those categories. The study also contributed an automatic classification on the OPP-115 dataset using Paragraph2Vec as the embedding mechanism, and three machine learning techniques for classification; Logistic Regression (LR), Support Vector Machine (SVM), and Hidden Markov Model (HMM). The study reported 0.66 micro F_1 score on average for the used classifiers.

Similar to [51], Liu *et al.* [28] performed an automatic classification on OPP-115, with new embedding, classifiers, and classification granularity. By extracting tf-idf features at the sentence and paragraph granularity, instead of Paragraph2Vec, and by using several machine learning methods; LR, SVM, and Convolutional Neural Networks (CNN), they evaluated the practicality of automatic classification. They achieved 0.66 and 0.78 micro F_1 scores for the sentence-based and paragraph-based classifications, respectively.

Liu *et al.* [27] explore unsupervised learning techniques on OPP-115, allowing them to analyze policies without an expert annotator. They utilized the Non-negative Matrix Factorization (NMF) method, which can construct a vocabulary for each category, with mappings between topic models and categories defined by the experts.

While the literature highlights the potential of the directions and techniques, it falls short by not delivering high accuracies on accepted benchmarks. In this work, we investigated the technical gap in the literature by employing numerous text representations and machine learning techniques from the previous studies. With an accurate ensemble, we retrieve the data collection and privacy practices of a website, and automatically select the paragraphs highlighting the topic of interest. Moreover, we performed a case study on Alexa top-10,000 websites privacy policies which further demonstrates the practicality of TLDR, and extended our work with a user study showing the benefits of our policy highlighting mechanism to the ordinary user in terms of information reduction, reading time, and policy understanding.

3 THE TLDR PIPELINE

Privacy policies are diverse with no standard format. This may, in many cases, result in vague information reporting, where information is embedded within multiple sentences or even paragraphs. Extracting information regarding the privacy policy and associated practices, in most cases, is not a straightforward task, and requires

Table 1: Summary of the related work. The table shows the best performing method of each study. All datasets are manually annotated, however, the OPP-115 dataset is annotated by expert law students, and therefore is used in this study.

Reference	Year	Representation	Annotation Level	Learning	Dataset	F_1 score
Ammar *et al.* [4]	2012	*n*-grams	Word	LR	57 policies	0.77
Constante *et al.* [14]	2012	POS tag	Word	-	12 policies	0.83
Constante *et al.* [15]	2012	Bag-of-Words	Document	Ridge	64 policies	0.90
Zimmect *et al.* [56]	2014	TF-IDF + bi-gram	Document	Naive Bayes	50 policies	0.90
Wilson *et al.* [51]	2016	Paragraph2Vec	Segment	SVM	OPP-115	0.66
Harkous *et al.* [21]	2018	Custom (fastText)	Segment	CNN	OPP-115	0.83
Liu *et al.* [28] (1)	2018	TF-IDF	Sentence	LR	OPP-115	0.66
Liu *et al.* [28] (2)	2018	TF-IDF	Segment	SVM	OPP-115	0.78
TLDR (1)	2021	WordPiece	Segment	BERT-Segment	OPP-115	0.91
TLDR (2)	2021	WordPiece	Segment	BERT-Document	OPP-115	0.91

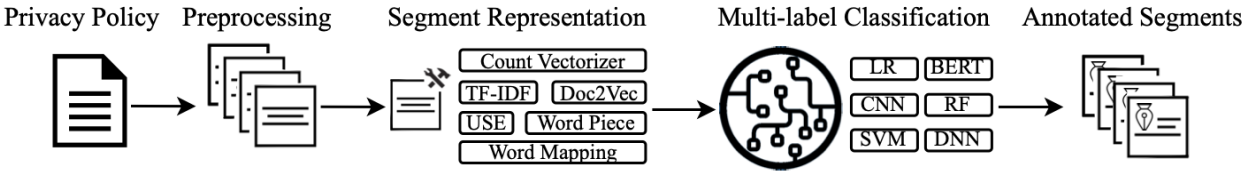


Figure 1: The data preprocessing and ensemble prediction pipeline. The processed segments are represented using different feature representation techniques, and then fed to the corresponding category classifier for multi-label classification.

various data and text representations depending on the dimensionality in which the patterns of the policy may exist. As such, we study utilizing various text representation and pattern extraction techniques for the classification of privacy policy practices.

In particular, we developed TLDR, an ensemble of automated machine and deep learning models to extract privacy and data collection practices from the policies. The TLDR pipeline is shown in Figure 1. In our pipeline, the segments are first preprocessed, then various text representation techniques are applied to extract deep representative features. Afterward, an ensemble of classifiers is used to predict the corresponding labels of each segment in a multi-label classification setting. Upon establishing a baseline for our learning model, we proceed to study the effectiveness of TLDR by conducting a user study for performance metrics, and its capabilities in policy highlighting. In the following, we describe the process of implementing the various aspects of our pipeline.

3.1 Ground Truth and Key Terminology

It is challenging to build a ground truth dataset for policy annotation, as that requires manual labor and domain expertise. As a baseline for privacy policy annotation, we used the Online Privacy Policies (OPP-115) dataset, proposed by Wilson *et al.* [51]. The dataset consists of privacy policies collected from 115 websites, manually annotated by *ten law school students*. Each policy is split into paragraphs, referred to as "segments". Each segment was labeled by *three annotators*, selecting phrases associated with each privacy policy practices category. Among the 115 policies, there are 3,792 segments, averaging 33 segments per policy. All annotators worked independently and needed 72 minutes to annotate each policy on average, associating sentences with one or more category.

Figure 2 shows the taxonomy used by Wilson *et al.* [51] for segment labeling. Privacy policies are categorized into high-level and low-level categories, including critical information regarding the privacy policy practices, such as "first-party data collection", "third party information sharing", and "user tracking" practices. We used the high-level categories (excluding "Others") for our automated annotator. A brief description of each category is in Table 2.

We note that the annotation within a sentence of a segment can be generalized to the entire segment. For instance, the segment is assigned a binary label (positive or negative) for the presence or absence of each privacy policy category. In this work, we used the segment-level labels produced by *the majority vote*: Once two annotators agree that a segment contains a privacy policy practice in a given category, we associate the segment with that category.

3.2 Privacy Policy Preprocessing

The privacy policies in OPP-115 are stored as Hypertext Markup Language (HTML) files, with *segments* contained in each privacy policy identified by the separator ("|||") defined by Wilson *et al.* [51]. Each segment consists of several *sentences*, and typically discusses one or more aspects of the privacy practices of the service provider. For each segment, the *stopwords* are removed using the Gensim [38], an open-source library used for processing extensive text collections. Stopwords are common words that do not add meaning to a sentence. Words that fall in this category may include prepositions and pronouns, and can be removed without sacrificing the meaning of the sentence.

The Natural Language Tool Kit (NLTK) [29] WordNet Lemmatizer is used for *lemmatization* and *stemming*. Word lemmatization is done by grouping the different forms of a word together so that they can be analyzed as a single term. Similarly, word-stemming is

Table 2: The description of each privacy policy high level category. The ensemble is trained on these categories, classifying each segment as positive and negative in context of each category.

Category	Description
First Party Collection/Use	The way and purpose of collecting and using user information by a service provider.
Third Party Sharing/Collection	How user information is collected and shared with third parties.
User Choice/Control	The ability of users to have the choice and control over their information.
User Access, Edit, & Deletion	The method and ability of users for accessing, editing, or deleting their information.
Data Retention	How long the stored user information is retained.
Data Security	The methods of securing and protecting user information form different types of breaches.
Policy Change	Describing whether and how a service provider will inform users about any changes to the privacy policy.
Do Not Track	Describing whether and how a service provider will honor online and advertising tracking.
Specific Audiences	Targeting a specific group of users (e.g., children, Europeans, or California residents).

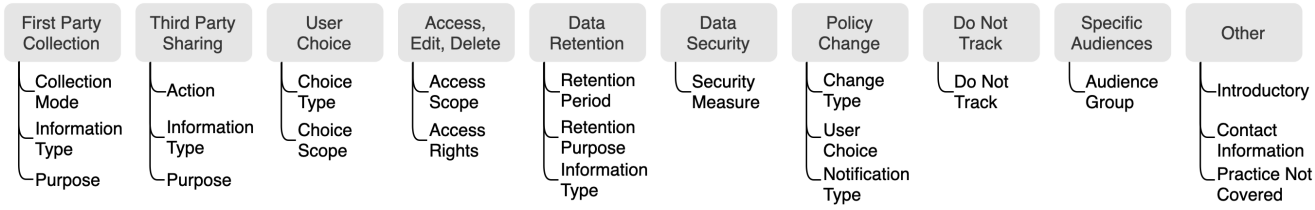


Figure 2: The taxonomy used by Wilson *et al.* [51] in categorizing the privacy policy practices and labeling each segment. We consider the high level nine categories in the process of building the ensemble classifier.

done by reducing the inflection in words to their root forms, such as mapping a group of words to the same stem, even though the stem itself may not be a valid word in the language. The segment preprocessing removes the generic words and words/sub-words that do not contribute to the meaning or context of the segment, making the learning process more efficient and accurate.

3.3 Segment Representation

To find a suitable highly-discriminative representation for each category, various text representation techniques are used in TLDR: word mapping [28], count vectorizer [42], TF-IDF [20], Doc2Vec [24], Universal Sentence Encoder (USE) [8], and WordPiece [53]. In the following, we provide a brief description of those techniques as used in TLDR.

Word Mapping. Originally proposed by Liu *et al.* [28], word mapping consists of a predefined set of terms, T, which are used to represent the segment. Terms that are most frequent within sentences as labeled by the manual annotators, and indicate the presence of a privacy practice category, are considered in T. Given a segment $s \in S$, the word mapping representation is defined as:

$$V_s = \begin{cases} 1 & t \in S, \forall t \in T \\ 0 & t \notin S, \forall t \in T \end{cases}, \tag{1}$$

where V_s is the vector representing the segment s. Table 3 shows the most frequent words per category, considered as the vocabulary baseline of this presentation. The presence or absence of each word will be represented by "1" or "0" in a vector of size $1 \times |T|$.

Count Vectorizer. Similar to the word mapping, this approach is used to convert a segment $s \in S$ to a vector V_s of terms counts. The terms are considered n-grams of the n number of sequential words, with a sliding window of one. The vocabulary of the count vectorizer contains all unique terms in the segments set S. The feature vector V_s is of size $1 \times |T|$, where V_{s_i} is calculated as follows:

$$V_{s_i} = \begin{cases} V_{s_i} + 1, & s_i = t \\ V_{s_i}, & s_i \neq t \end{cases}, \quad \forall s_i \in s \; \forall t \in T, \tag{2}$$

where s_i is the i^{th} term in the segment s. In simple terms, the generated vector V_s represents the count of each term t in the vocabulary T in the segment s.

Term Frequency–Inverse Document Frequency (TF-IDF). In the TF-IDF approach, a term t in a segment $s \in S$ of a vocabulary T is assigned a weight using the following representation model:

$$\text{TF-IDF}(t, s, S) = \text{TF}(t, s) + \text{IDF}(t, S), \tag{3}$$

where $\text{TF}(t, s)$ is the term frequency of term t in segment s and $\text{IDF}(t, S)$ is defined as follows:

$$\text{IDF}(t, S) = \log(|S|/DF(t, S)) + 1, \tag{4}$$

where $\text{DF}(t, S)$ is the number of segments that contain the term t. The core concept of the TF-IDF is to find the inverse document frequency of each n-gram term. Using TF-IDF, a widely used technique for text representation, we explore statistics of the occurrence of n-consecutive words or terms across the segments. The probability of sequence of words, using the chain rule, is calculated as follows:

$$p(w_1, \ldots, w_T) = p(w_1) \prod_{i=2}^{T} p(w_i | w_1, \ldots, w_{i-1}). \tag{5}$$

Given the arbitrary length of segments, (5) may become infeasible to compute. To this end, we employ a commonly used approximation, where only the previous n terms are considered in computing the representation. For example, for the conditional probability of a word w_i given n-words can be described as:

$$p(w_i | w_1, \ldots, w_{i-1}) = p(w_i | w_{i-n+1}, \ldots, w_{i-1}). \tag{6}$$

Table 3: A predefined set of the most frequent terms from the manual annotation process, and used in the word mapping approach as the vocabulary of interest in representing each segment.

Category	Terms
First Party Collection/Use	use, collect, demographic, address, survey, service, information, require, identify
Third Party Sharing/Collection	party, share, sell, disclose, company, advertiser, provide, partner, transfer, sell, report
User Choice/Control	opt, unsubscribe, disable, choose, consent, agree, withdraw, refuse, permit
User Access, Edit, & Deletion	delete, profile, correct, account, change, update, modify, affiliate, track
Data Retention	retain, store, delete, deletion, database, participate, maintain, ensure, reserve, hold
Data Security	secure, security, seal, safeguard, protect, ensure, confidentiality, authorization, protect, practices
Policy Change	change, change privacy, policy time, current, policy agreement, time, current agreement
Do Not Track	signal, track, track request, respond, browser, advertising, content, visit, cookie, service, respond
Specific Audiences	child, California resident, European, age, parent

The model in (6) is a probabilistic approximation, and can be realized computationally for n-grams by counting their relative occurrences in all segments using the maximum likelihood estimation:

$$\hat{p}(w_a) = \frac{c(w_a)}{N}, \hat{p}(w_b|w_a) = \frac{c(w_a, w_b)}{\sum_{w_b} c(w_a, w_b)} \approx \frac{c(w_a, w_b)}{c(w_a)}, \quad (7)$$

where N is the length of the corpus (number of terms) and $c(\cdot)$ is the count of each term in the corpus.

Doc2Vec. Originally proposed by Le *et al.* [24], Doc2Vec is a pre-trained model that creates a numeric representation of a document, regardless of the document length. This approach is an adaptation of the original Word2Vec approach [32], extended for documents (segments). In a nutshell, Doc2Vec finds the best numerical representation of a segment s according to the words in that segment, using Word2Vec text representations and the continuous bag-of-words as underlying techniques for words, and by merging the words numerical representations into a single feature vector V_s.

Universal Sentence Encoder (USE). USE [8] encodes text into high-dimensional vectors that can be used in several applications, e.g., text classification. The generated embeddings are discriminative and unique, and have been shown to be effective in conducting various NLP tasks, including text classification and semantic similarity. Technically, USE takes a variable-length English text as an input and outputs a 1×512 dimensional vector representing the text. As a pre-trained model, USE is originally trained on sentences, phrases, or short paragraphs, and optimized on "greater-than word length texts". In the literature, several datasets are curated and used for the training, including the Stanford Natural Language Inference (SNLI) corpus [6], and Wikipedia [50] English articles. In essence, USE is a deep learning transformer [45] trained with a deep averaging network (DAN) encoder on billions of articles.

WordPiece. We use WordPiece [53] to represent segments for BERT [18]. For this purpose only, the original segment is preprocessed with WordPiece. WordPiece creates a vocabulary of a fixed number of words, subwords, and characters. Such a variable granularity solves the out-of-vocabulary problem by splitting unrecognized words into subwords. When no subword matches in the predefined dictionary, the candidate word is split further into characters, and mapped to the corresponding embedding.

Preprocessing and feature representation are critical in TLDR implementation to unveil the hidden patterns within each segment without the need for human labor or manual annotation.

3.4 Learning Algorithms

TLDR trains an ensemble of learning algorithms for associating segments with their corresponding privacy policy categories. Doing so reveals the abstract content of the privacy policy without the need for reading such content. This allows us to provide an overview of the content of the privacy policy, and highlight any aspects that the policy is missing. We leverage six machine and deep learning algorithms for privacy policy detection, evaluating their effectiveness in detecting various segment-level categories. The following is a brief description of each learning algorithm.

Logistic Regression (LR). In a simple notation, LR is a statistical model that uses a logistic function to model a binary dependent variable, known as binary classification ("0" or "1"). Given an input training set (X, Y), LR learns to distinguish between positive ("1") and negative ("0") segments for each category by drawing a boundary line (assuming a linear relationship). In the higher domain, LR estimates the boundary between the positive and negative classes and optimizes the boundary by minimizing the following:

$$\text{Loss}(f(X), Y) = \begin{cases} -\log(f(X)), & Y = 1 \\ -\log(1 - f(X)), & Y \neq 1 \end{cases}, \quad (8)$$

where $f(X)$ is the prediction and Y is the ground truth label.

Support Vector Machine (SVM). The SVM algorithm operates by finding a hyperplane in an N-dimensional space, where N is the length of a feature vector that distinctly classifies the segments. Toward that, SVM finds a plane that has the maximum distance between segments of both classes (positive and negative). To do so, SVM calculates the loss of each segment, defined as follows:

$$\text{Loss}(X, Y, F(X)) = \begin{cases} 0, & y \times f(x) \geq 1 \\ 1 - y \times f(x), & y \times f(x) < 1 \end{cases}. \quad (9)$$

Random Forest (RF). Typically used with non-linear classification tasks, RF consists of N decision trees, each of which is trained on random features selected for individual trees. Such a technique allows for variance reduction in the output of the individual trees and mitigates the effect of noise on the training process. For RF with N decision trees, the final prediction is calculated by either a majority vote on the predictions of the decision trees, or by outputting the average prediction of all the trees, calculated as $f_{RF} = \frac{1}{N} \sum_{n=1}^{N} f_n(X_s')$, where f_n is the prediction of the n^{th} tree, X_s' is the vector representation of segment s for a randomly selected feature set $(X_\cdot' \subset X_\cdot)$.

Convolutional Neural Networks (CNN). The CNN model is a powerful deep learning algorithm, typically used in image classification and pattern recognition. The basic unit of the CNN model is a convolution layer, which consists of several filters convolving over the input to generate feature maps. Once a feature vector is fed into a convolutional layer, it becomes abstracted to a feature map, with the shape of (feature map height) × (feature map width) × (feature map depth), with two attributes: (1) convolutional kernels defined by a width and a height (hyper-parameters), (2) the depth of the convolution filter, which is equal to the depth of the segment vector representation the feature map. In general, CNN provides excellent results in extracting patterns in higher dimensionality when the pattern location is irrelevant (in the feature space).

Deep Neural Networks (DNN). DNN consists of multiple consecutive fully connected layers, extracting deep encoded patterns. For a single layer l, the model configures the layer parameters in the learning stage. Each layer is denoted by $h^{(l)} = a(W^{(l)} \times X + b^{(l)})$, where $a(.)$ is an activation function of layer l, $W^{(l)}$ is the weights of the features from layer $l-1$ to layer l, and $b^{(l)}$ is the bias of layer l.

Bidirectional Encoder Representations from Transformers (BERT). BERT [18] is a transformer-based language model that benefits from the attention mechanism provided by the transformer architecture [46]. In essence, BERT has two six-layers of encoders and decoders, each of which has the ability to learn contextual relations between words in a given context. To utilize the BERT model, each segment is preprocessed using WordPiece, where the word, subword, and character-level terms matching mechanisms are used. BERT fits a wide variety of NLP language tasks, including text classification, question answering, and named entity recognition.

Learning Algorithm Selection. The selection of the learning algorithm for each category is essential to achieve a highly accurate privacy policy annotator. For instance, learning algorithms vary in the level of pattern extraction, with CNN, DNN, SVM, and BERT extracting patterns in high dimensionality, whereas LR and RF are used for extracting statistical and low dimensionality patterns within the segments. Moreover, both machine and deep learning are used, with techniques such as convolutional filters (CNN), information attention (BERT), and random decision trees (RF).

Acknowledging that privacy policy categories are unique, TLDR leverages the best performing data representation for category classification. The wide range of explored representations and learning algorithms is due to the diversity of categories, where treating them indiscriminately results in a reduced performance.

4 EVALUATION AND DISCUSSION

The evaluation of TLDR is threefold. First, we assess the performance of TLDR using classification evaluation metrics on the OPP-115 for both baseline performance and to contrast TLDR's classification capabilities with the literature (4.1). Second, to highlight TLDR's applications in context, we conduct a case study by analyzing the privacy policies of the Alexa top-10,000 websites (4.2). Third, to highlight the usefulness of TLDR in terms of improving user experience, understandability of policies, and time spent on reading them, we conduct a user study for our evaluation (4.3).

4.1 Annotation Results of TLDR

4.1.1 Experimental Setup. **Training and Validation.** For this evaluation, we use the OPP-115 dataset. We split the OPP-115 dataset into *training* and *validation* sets. In TLDR, we follow two splitting techniques: *segment-based splitting*, where segments are randomly split into 80-20 training and validation sets regardless of the associated documents, and (2) *document-based splitting*, where 80% of the documents are used for training the ensemble while the remaining 20% of the documents are used for validation.

The segments are represented using five-word representation techniques: word mapping, count vectorizer, TF-IDF, Doc2Vec, and USE. WordPiece word representation is only used with BERT.

Hyper-Parameter Selection. We obtained the best parameters per word using brute force and grid search. Namely, the most frequent candidate words are vectorizer and TF-IDF are set in the range $[1,000, 5,000]$, with an increment of 1,000. We set the Doc2Vec encoding vector size to $1\times64, 128, 256, 512$, respectively, selecting the best performing vector size. The USE pre-trained model does not require any parameters, and outputs a vector of size 1×512 for each segment. Moreover, we used the default LR and SVM configuration in the learning stage, provided by the SKLearn libarary [36]. The configuration of the LR model includes using the L_2 distance as the penalty metric, 100 maximum iterations, and balanced class weights. Similarly, the SVM model is configured with *"rbf"* kernel, and balanced class weights. For RF, we set the number of decision trees to 100, with a majority vote prediction output.

We adopt the architecture by Harkous *et al.* [21] for our CNN model, and replace the convolutional layers with fully connected layers to build the DNN model. We configure the BERT model with a 512 maximum number of words, with the number of features in the range of [1,000, 5,000] with an increment of 1,000. The BERT model is trained with learning rates of $[5\times e^{-5}, 3\times e^{-5}, 2\times e^{-5}]$, and 10 training epochs. The best performing BERT model is obtained with 1,000 features and $2 \times e^{-5}$ learning rate.

Evaluation Metrics. The learning algorithms are evaluated using precision, recall, and F_1 score. The precision metric answers the question of "*How many segments labeled as positive are correct?*", and can be mathematically calculated as $P = TP/(TP + FP)$ where TP is the true positives, referring to the positive segements that were correctly classified, and FP is the false positives, referring to the negative segments that were incorrectly classified as positive by the learning model. The recall metric answers the question of "*How many positive segments were classified correctly?*", and is defined as $R = TP/(TP + FN)$ where FN is the false negatives, referring to the positive segments incorrectly classified as negative. The F_1 score is a measure that combines the precision and recall, thus called their harmonic mean, and is calculated as $F_1 = 2 \times (P \times R)/(P + R)$.

4.1.2 Annotation Results. Figure 3 shows the evaluation of each category using the six learning algorithms of TLDR and Table 4 shows the best performing architecture for each category used for analysis. The evaluation of the different configurations of the ensemble is shown in Tables 5 and 6 in the appendix. The BERT-based model is shown to outperform all other techniques in all categories, and therefore is used as a baseline model to build an ensemble of learning algorithms for automated privacy policy annotation.

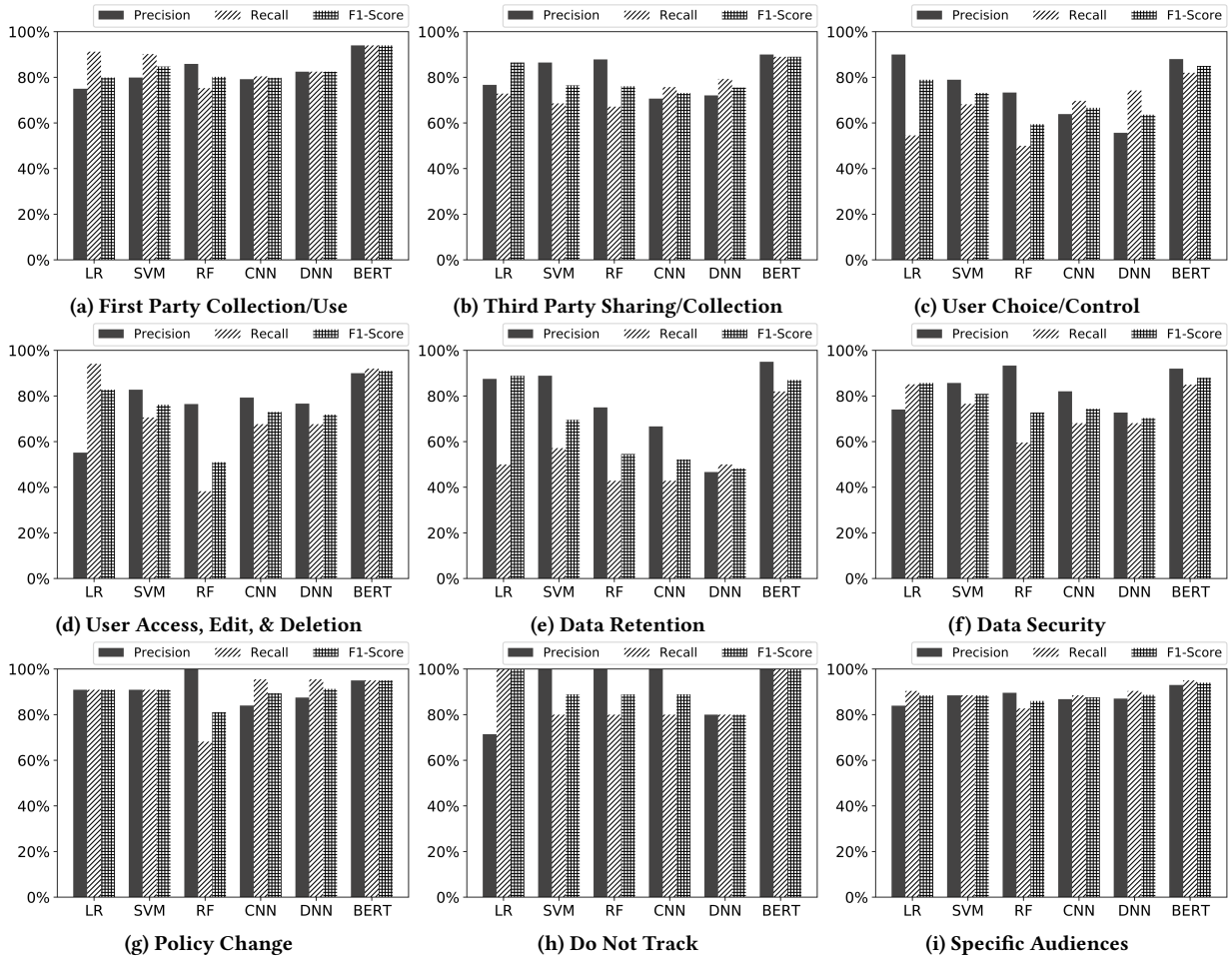


Figure 3: The performance of the learning algorithms on each privacy policy category on OPP-115 dataset. The best performing learning algorithm is then used in the ensemble classifier.

Table 4: The performance (F_1 score) of the ensemble classifier using best performing word representations and learning algorithms on OPP-115. TS: TLDR-Segment, TD: TLDR-Document, W: Wilson et al. [51], H: Harkous et al. [21], L: Liu et al. [28]. The full names of the categories are in Table 3.

Category	TS	TD	W	H	L
First party	0.92	0.94	0.75	0.79	0.81
Third party	0.90	0.89	0.7	0.79	0.79
User choice	0.88	0.85	0.61	0.74	0.70
User access	0.90	0.91	0.61	0.80	0.82
Data retention	0.78	0.87	0.16	0.71	0.43
Data security	0.88	0.88	0.67	0.85	0.80
Policy change	0.98	0.95	0.75	0.88	0.85
Do not track	1.00	1.00	1.00	0.95	1.00
Specific audiences	0.93	0.94	0.70	0.95	0.85
Overall	0.91	0.91	0.66	0.83	0.78

Similarly, we report the best performing evaluation results of Wilson *et al.* [51] and Liu *et al.* [28] on the OPP-115 dataset using the F_1 score. As shown, except for "Specific Audiences" and "Do Not Track" categories, TLDR outperforms its counterparts by a large margin, particularly for "Data Retention". The independent results for all models using the best performing representation are shown in Tables 7 and 8 in the appendix.

The main intuition of this work is to provide a highly accurate privacy annotation system that is capable of highlighting segments that include information of interest to users, as recent studies showed that reading privacy policy is a time-consuming task. High F_1 score indicates that the segments returned as positive by the ensemble are most likely to cover the information of the category of interest in a particular policy.

4.2 Case Study: Alexa Top-10,000 Websites

We used the implemented ensemble (TLDR) towards analyzing Alexa [3] top-10,000 websites privacy policies. The Alexa top-10,000 websites represent the most visited websites by users worldwide. Analyzing such websites would uncover the common practices of popular websites and their service providers (owners), targeting a large portion of Internet users. While analyzing more websites might seem preferable, given the cumbersome process of privacy policy crawling, and the limited potential insight from the tail of

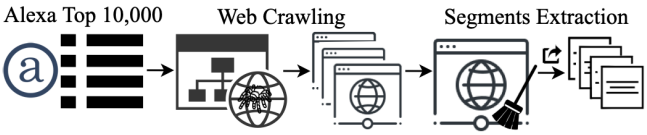


Figure 4: Our data collection and segment extraction pipeline, including crawling the website structure and searching for the privacy policy. Once found, paragraphs are extracted and preprocessed to extract the policy segments.

the distribution of the website with respect to their popularity, we limited ourselves to the top-10,000 websites, although extending the study beyond those websites is an open direction. In the following, a description of the dataset collection process and evaluation results.

4.2.1 Dataset Collection and Processing. **Privacy Policy Extraction.** We start by obtaining the privacy policies of the websites among the Alexa top-10,000 websites list. This is done using Selenium [41], an automated browser testing framework that provides extensions to emulate user interaction with browsers. The privacy policies were crawled from the Alexa top-10,000 websites in the period of November 4–8, 2020; *i.e.,* after the introduction of GDPR.

Among the top-10,000 websites, we successfully extracted the privacy policies of 5,598 websites. Using Selenium, we crawl all the website visible pages from the home directory. Then, the privacy policies are extracted by searching the webpages within a website for terms such as *privacy policy*, and *privacy*.

Once found, the associated HTML with the privacy policy is saved for processing. The remaining websites are either non-English or do not directly link their privacy policy in the website structure. Using the HTML paragraph tag (< p >), we extract all paragraphs using Beautiful Soup [25], a python library for parsing HTML and XML documents. The extracted paragraphs are considered as our potential segments. We removed all segments with less than ten words, as in most cases they are introductory sentences and do not contribute towards the privacy practices, nor contain privacy and data collection practices. The remaining segments are then associated with the extracted privacy policy for category analysis. The process of website crawling and cleaning is illustrated in Figure 4.
Validation. To evaluate the correctness of the extracted policies, we manually inspect 1,000 extracted policies, and verified that 95.8% of them are correctly extracted. As such, we proceed with the extracted policies under the assumption that they are correctly extracted.

Topic Extraction. We extracted the topics associated with the 5,598 Alexa top websites to understand the differences in the privacy policies reporting between different topics. We used Webshrinker [17], a machine learning-powered domain data and threat classifier, to obtain the categorization of the domains of the websites. To this end, we extracted 66 different categories and grouped them into 18 high-level categories. Note that a website may be associated with one or more categories; to address this, we consider topic with the highest confidence score returned by Webshrinker API as the label. Table 9 in the appendix shows the high-level topics and their sub-topic taxonomy used in this work.

Data Preprocessing & Representation. The extracted segments (345,920) are then preprocessed and represented in a way similar to OPP-115's segment preprocessing outlined in section 3.2.

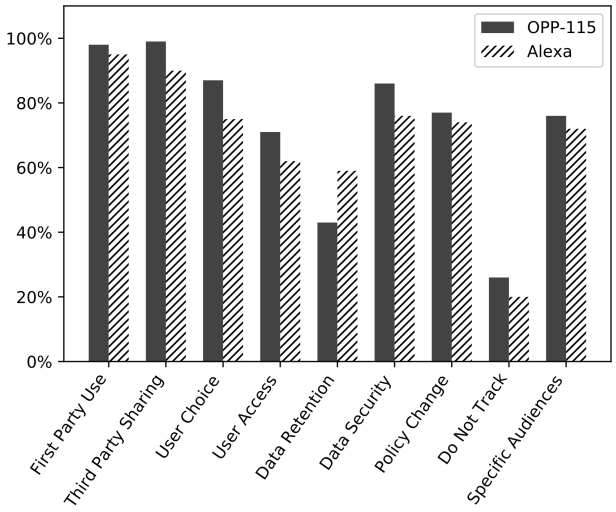


Figure 5: Percentage of websites with positive segments per category for Alexa top-10,000 websites and OPP-115 dataset.

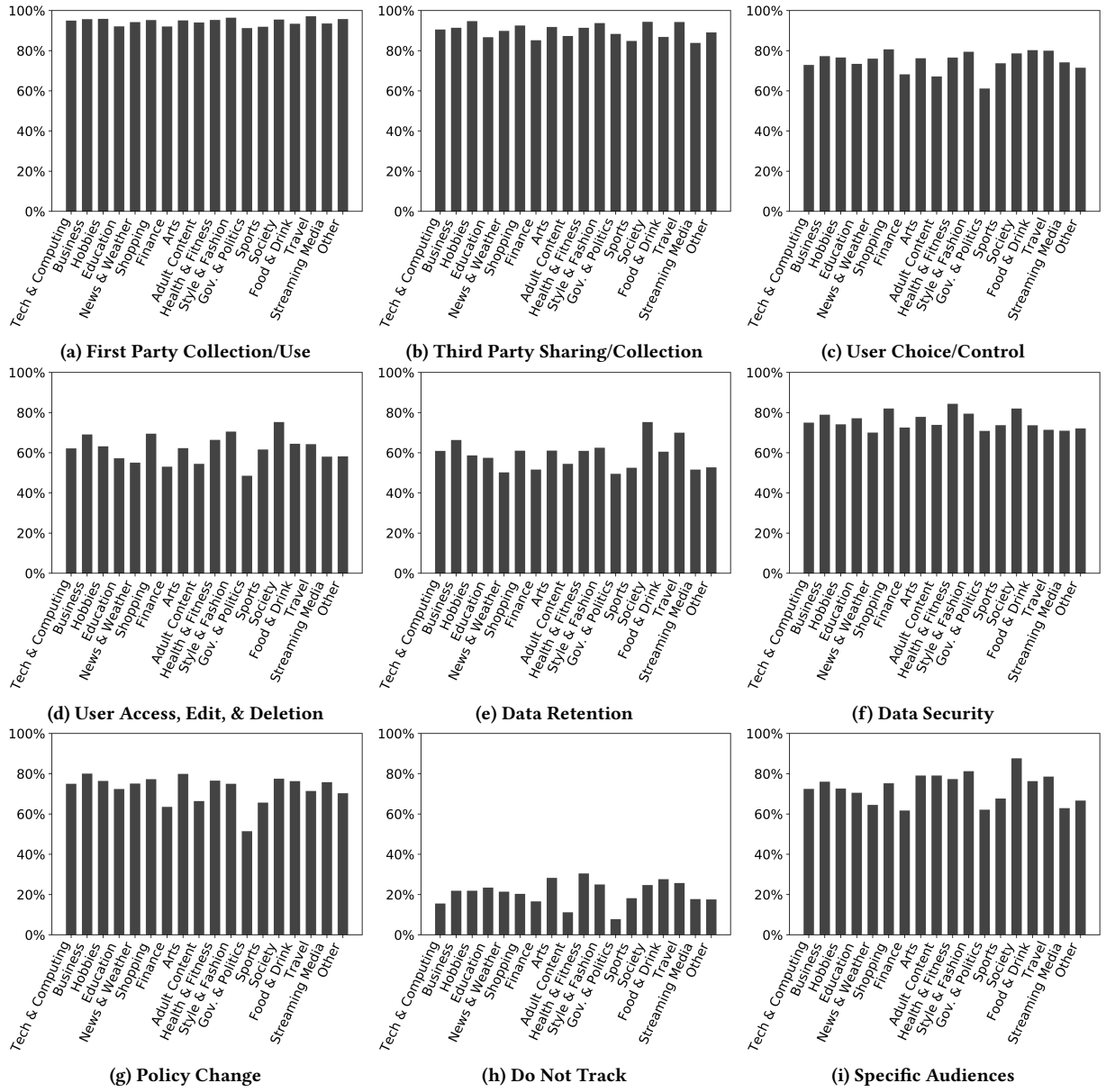
We removed the stop-words, then the words in the segment were lemmatized and stemmed. We limit the configurations of the hyperparameters to the best performing within the feature representations and learning algorithms referred to in Table 4. The data preprocessing and representation tasks are illustrated in Figure 1.

4.2.2 Evaluation & Discussion. We used TLDR to extract existing privacy practices within Alexa top-10,000 websites. Figure 5 shows the percentage of websites containing information regarding the policy categories for both OPP-115 and Alexa top-10,000 websites.

Overall, the "first-party use" and "third party sharing" categories are the most common within the privacy policies, with 95% of the websites containing first-party use information, and 90% of them including information regarding third-party sharing. On the other hand, the "do not track" category is the least common within the privacy policies, with only 20% of the websites reporting information associated with it. Given that the ensemble achieves an F_1 score of 100% on this category, the results are of high confidence.

Missing Information. By examining the ensemble results, we found that a large number of websites' privacy policies miss key information and attributes, by not covering important areas including data security and user tracking. This comes as a surprise, given that the extracted privacy policies are from the top-visited websites as of 2020, which are potentially the subject of great interest, and their policies are the subject of great scrutiny.

Next, we conduct a topic-based analysis on the privacy policies to unveil the differences between the reported privacy practices of different websites with the same interest. Figure 6 shows the behavior of the websites associated with the same topic per privacy category. In this experiment, we unveil that only 8% of the "Law, Government, & Politics" associated websites report user tracking policies, and only 51% report information about the change of policies of the service. This might be due to the possibility that "Law, Government, & Politics" associated websites use a different terminology within the privacy policy reporting compared to other domains, leading to lower detection rate (*i.e.,* accuracy). A point worth noting is that only 85% of the personal-finance websites state



(a) First Party Collection/Use (b) Third Party Sharing/Collection (c) User Choice/Control

(d) User Access, Edit, & Deletion (e) Data Retention (f) Data Security

(g) Policy Change (h) Do Not Track (i) Specific Audiences

Figure 6: The percentage of the Alexa top-10,000 websites with positive segments per privacy policy categories. The websites are divided according to the associated topic obtained from Webshrinker API.

information regarding "third party sharing/collection" practices. On the other hand, 30% of health and fitness websites include information regarding user tracking, and 84% of them discuss data security, the highest percentages among all topics.

Information Highlighting. The privacy policies are typically long, unclear, and hard to understand. We recall that reading privacy policies may require 201 hours per year [30]. Therefore, with the accurate performance, the proposed ensemble can be used to highlight the information of certain categories of interest, reducing the amount of information required to be read to understand the website privacy policy and expectations. Figure 7 shows the percentage of the segments and words highlighted by TLDR for a category of interest. The ensemble selects the segments that contain information regarding each category. On average, TLDR can reduce the number of segments required to be read by more than 45%, and up to 99% when only considering *"Do Not Track"* privacy policy-related information. Meanwhile, Table 10 in the appendix shows example segments classified as positive and negative per category, including the source website as of November 2020.

Malicious Websites Policy Reporting Behavior. Next, we move toward understanding the difference in the privacy practices reporting behavior of malicious websites. To do so, we first extract

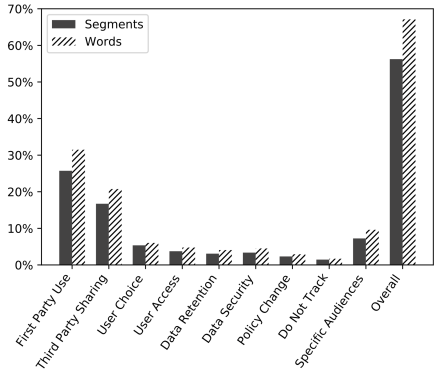


Figure 7: The percentage of segments and words from Alexa top-10,000 highlighted by TLDR and associated with each category. Overall: all categories.

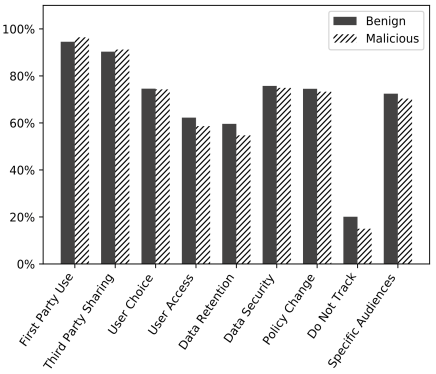


Figure 8: The percentage of websites with positive segments per category for both Alexa top-10,000 benign and malicious websites.

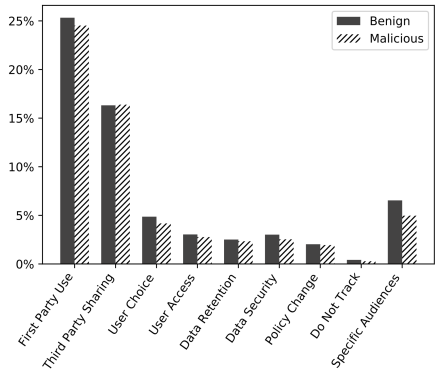


Figure 9: The percentage of segments obtained from both Alexa top-10,000 benign and malicious websites associated with each privacy policy category.

malicious websites among Alexa top-10,000 websites. Toward this, we used the VirusTotal API [47] to extract information regarding the malicious activities of the websites. Among the 5,598 websites with extracted privacy policies, there were 307 malicious websites and 5,291 benign websites, as labeled by VirusTotal. Figures 8 and 9 show the difference between benign and malicious websites in reporting their privacy and data collection practices. As shown, there is no significant difference in reporting behavior between benign and malicious websites. This may be contributed to the fact that malicious websites use generic privacy policy templates to satisfy search engines or host-specific requirements. We recall that studying the confidence of the reported practices is out-of-scope of this work, and therefore, we can not verify whether the stated practices are followed correctly.

4.3 User Study: Highlights Improve UX

Next, we conduct a user study to understand the effectiveness of TLDR in highlighting privacy policy information. Removing unnecessary policy information reduces the efforts required to understand the privacy practices of the service providers. However, omitting important privacy policy aspects can be critical by reducing the users awareness of the reported practices.

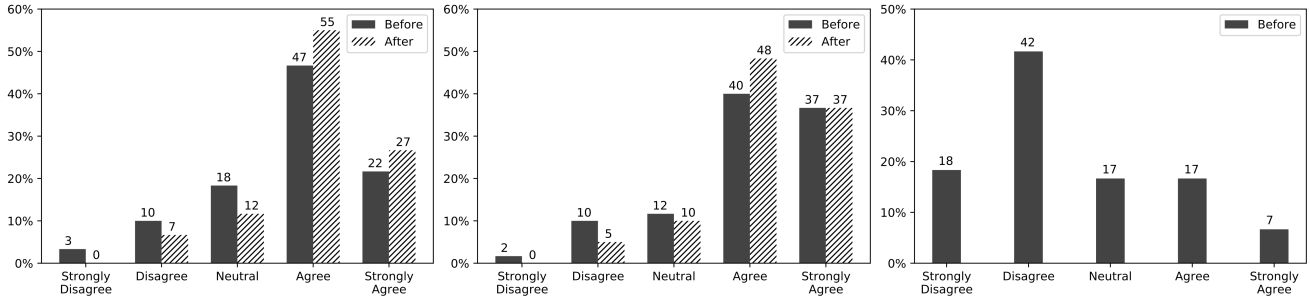
The user study consists of three unique privacy policies per participant and four surveys. An introductory survey is done before starting the study, and three surveys, each of which are done after reading a privacy policy before and after the filtering. The selected privacy policies are of short (800–1200 words), medium (1800–2200 words), and long (2800–3200 words) lengths. We note that the selection of the privacy policies is constraint by the time and efforts needed to read and understand each policy. For each privacy policy, two instances were initiated, the original and the TLDR filtered policies (*i.e.,* after removing all paragraphs with no associated privacy policy information). The participants then read the two instances of each policy in a random order, and are not aware of the filtering process nor the objective of this study. Each participant was given a unique set of privacy policies, with a total of 20 participants (*i.e.,* 60 overall unique privacy policies).

After reading both instances of the privacy policy, the participant is asked to fill a survey as follows:

- **Privacy policy token.** To ensure the randomness in the selection of the privacy policies, and the order of the two instances (*i.e.,* original and after filtering), we provide a token for each participant to keep a record of the order of the privacy policies for analysis purposes.
- **Reading time.** The participant is expected to keep a record of the reading time of the privacy policy. This is essential to understand how TLDR helps in reducing the average reading time, and thus the efforts needed to read the privacy policy.
- **Understanding the privacy policy.** For both instances of the privacy policy, the participant is asked whether it is understandable, readable, and covers the essential information needed to understand the service privacy practices. We recall that the participant is unaware of the filtering process, and there is no order in which we ask this question.
- **Missing privacy information.** The participant is asked whether each instance of the privacy policy includes critical information that is missing in the other instance. We only report the answers associated with "whether the original privacy policy includes critical practices unreported in the filtered instance". Note that the question is asked for each instance to ensure removing any bias caused by focusing on one instance over the other.

Results and Observations. Users' online privacy is crucial. In this study, 80% of the participants are strongly concerned regarding their online privacy and data. However, due to the length and ambiguity of privacy policies, 75% of the participants indicated that they do not read them. In particular, 80% of participants stated that privacy policies are hard to understand and comprehend.

Moreover, according to 82% of the participants, the privacy policies are understandable and suitable upon applying the TLDR filtering process as shown in Figure 10a. This is surprising, as only 69% of the participants indicated the original privacy policies are understandable, a percentage that is lower than the TLDR filtered policies (*i.e.,* 18.84% increase in comparison with the original privacy policy). This may be a result of removing unnecessary (legal) information that does not necessarily contribute to the privacy practices. The participants also reported a reduced reading time for TLDR filtered



(a) Suitable and understandable. **(b) Covers the essential information.** **(c) Includes missing critical privacy practices.**

Figure 10: The user study participants answers for privacy policies (1) before and (2) after the filtering. Figure 10c shows the answers of participants on whether the unfiltered privacy policy includes critical information missing in the filtered instance.

policies, *i.e.,* an average of 39.14% time reduction in comparison with the original policy.

As noted, removing important information may result in missing critical data collection and sharing practices. According to the participants' responses, as shown in Figure 10c, only 23% of the policies include critical information that is not reported in the TLDR filtered instance. Figure 10 shows the participants' answers for privacy policies before and after the TLDR filtering process regarding the understandability, coverage, and including critical practices.

Ethical Considerations. During the user study, we follow the best efforts privacy preserving practices. Participants were not asked to enter any private information that might reveal their identity, including their names and nationality. Further, we follow the best efforts in removing any bias that might appear during the study. Each participant was given a unique set of privacy policies in a different order than the other participants. We then used a token-based system to analyze the data. All the questions asked during the study focused on both instances of the privacy policy equally.

5 CONCLUDING REMARKS

In this work, we revisit the automated privacy policy annotation problem by exploring and improving the accuracy of annotation through various learning and representation techniques. We supplement our study by highlighting the policy segments that contain key information of interest. In particular, we propose TLDR, a pipeline employing deep representation techniques and an ensemble of machine and deep learning-based model to automatically and accurately categorize each segment (paragraph) in the privacy policy to its corresponding high-level content category, achieving an accuracy of more than 90% in various categories.

Through an analysis of the privacy policies of the Alexa top-10,000 websites, we further unveil some missing information typically expected in the privacy policy concerning various categories of high-level behaviors. We observed no difference in policy structure of benign websites, when compared to websites reported as malicious by VirusTotal. Finally, we conduct a user study to show the effectiveness of TLDR in highlighting the important aspects of the privacy policies. In particular, participants took 39.14% less time reading the TLDR filtered privacy policies, in comparison with the original privacy policies. Moreover, 67% of the participants prefer reading the TLDR-provided privacy policies. TLDR is an effective tool for analyzing the privacy policy practices by the service providers, and privacy policy highlighting tool to reduce the reading time and needed effort by the service users.

Acknowledgement. This work is supported by the National Research Foundation (NRF) under grant 2016K1A1A2912757 and a seed grant from CyberFlorida. A short version of this work appeared in ACM CCS 2021 as a poster [2].

REFERENCES

[1] U.S. General Service Administration. 2018. Rules and Policies - Protecting PII - Privacy Act. https://www.gsa.gov/reference/gsa-privacy-program/rules-and-policies-protecting-pii-privacy-act
[2] Abdulrahman Alabduljabbar, Ahmed Abusnaina, Ulku Meteriz, and David Mohaisen. 2021. Automated Deep Automated Privacy Policy Annotation with Information Highlighting. In *The 2021 ACM SIGSAC Conference on Computer and Communications Security (CCS '21)*.
[3] Amazon. 2020. Alexa top websites. https://www.alexa.com/topsites
[4] Waleed Ammar, Shomir Wilson, Norman Sadeh, and Noah A Smith. 2012. Automatic categorization of privacy policies: A pilot study. *School of Computer Science, Language Technology Institute, Technical Report CMU-LTI-12-019* (2012).
[5] Benjamin Andow, Samin Yaseer Mahmud, Wenyu Wang, Justin Whitaker, William Enck, Bradley Reaves, Kapil Singh, and Tao Xie. 2019. PolicyLint: Investigating Internal Privacy Policy Contradictions on Google Play. In *28th USENIX Security Symposium, USENIX*. 585–602.
[6] Samuel R Bowman, Gabor Angeli, Christopher Potts, and Christopher D Manning. 2015. A large annotated corpus for learning natural language inference. *arXiv preprint arXiv:1508.05326* (2015).
[7] Fred H. Cate. 2010. The Limits of Notice and Choice. *IEEE Secur. Priv.* 8, 2 (2010), 59–62.
[8] Daniel Cer, Yinfei Yang, Sheng-yi Kong, Nan Hua, Nicole Limtiaco, Rhomni St. John, Noah Constant, Mario Guajardo-Cespedes, Steve Yuan, Chris Tar, Yun-Hsuan Sung, Brian Strope, and Ray Kurzweil. 2018. Universal Sentence Encoder. *CoRR* abs/1803.11175 (2018).
[9] Rosalie Chan. 2019. https://www.businessinsider.com/cambridge-analytica-whistleblower-christopher-wylie-facebook-data-2019-10
[10] Federal Trade Commission. 2011. Protecting Consumer Privacy in an Era of Rapid Change - A Proposed Framework for Businesses and Policymakers (Preliminary FTC Staff Report). *J. Priv. Confidentiality* 3, 1 (2011).
[11] Nicholas Confessore. 2018. Cambridge Analytica and Facebook: The Scandal and the Fallout So Far. https://www.nytimes.com/2018/04/04/us/politics/cambridge-analytica-scandal-fallout.html
[12] World Wide Web Consortium et al. 2003. Platform for Internet content selection (PICS). *http://www. w3c. org/PICS/* (2003).
[13] Corinna Cortes and Vladimir Vapnik. 1995. Support-vector networks. *Machine learning* 20, 3 (1995), 273–297.
[14] Elisa Costante, Jerry den Hartog, and Milan Petkovic. 2012. What Websites Know About You. In *Data Privacy Management and Autonomous Spontaneous Security, 7th International Workshop, DPM (Lecture Notes in Computer Science, Vol. 7731)*. 146–159.
[15] Elisa Costante, Yuanhao Sun, Milan Petkovic, and Jerry den Hartog. 2012. A machine learning solution to assess privacy policy completeness: (short paper). In *Proceedings of the 11th annual ACM Workshop on Privacy in the Electronic Society, WPES*. 91–96.

[16] Lorrie Faith Cranor. 2002. *Web privacy with P3P - the platform for privacy preferences.* O'Reilly.
[17] Developers. 2020. Webshrinker. https://www.webshrinker.com/.
[18] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT*. 4171–4186.
[19] Joshua Gluck, Florian Schaub, Amy Friedman, Hana Habib, Norman M. Sadeh, Lorrie Faith Cranor, and Yuvraj Agarwal. 2016. How Short Is Too Short? Implications of Length and Framing on the Effectiveness of Privacy Notices. In *Twelfth Symposium on Usable Privacy and Security, SOUPS*. USENIX Association, 321–340.
[20] David J Hand and Niall M Adams. 2014. Data Mining. *Wiley StatsRef: Statistics Reference Online* (2014), 1–7.
[21] Hamza Harkous, Kassem Fawaz, Rémi Lebret, Florian Schaub, Kang G Shin, and Karl Aberer. 2018. Polisis: Automated analysis and presentation of privacy policies using deep learning. In *27th USENIX Security Symposium (USENIX Security 18)*. 531–548.
[22] Tin Kam Ho. 1995. Random decision forests. In *Proceedings of 3rd international conference on document analysis and recognition*, Vol. 1. 278–282.
[23] Graham Kates. 2017. Facebook, for the first time, acknowledges election manipulation. https://www.cbsnews.com/news/facebook-for-the-first-time-acknowledges-election-manipulation/
[24] Quoc V. Le and Tomas Mikolov. 2014. Distributed Representations of Sentences and Documents. In *Proceedings of the 31th International Conference on Machine Learning, ICML (JMLR Workshop and Conference Proceedings, Vol. 32)*. 1188–1196.
[25] Leonard. 2020. Beautiful Soup. https://www.crummy.com/software/BeautifulSoup/
[26] T. Linden, Hamza Harkous, and Kassem Fawaz. 2020. The Privacy Policy Landscape After the GDPR. *Proceedings on Privacy Enhancing Technologies* 2020 (2020), 47 – 64.
[27] Frederick Liu, Shomir Wilson, F. Schaub, and N. Sadeh. 2016. Analyzing Vocabulary Intersections of Expert Annotations and Topic Models for Data Practices in Privacy Policies. In *AAAI Fall Symposia*.
[28] Frederick Liu, Shomir Wilson, Peter Story, Sebastian Zimmeck, and Norman Sadeh. 2018. Towards automatic classification of privacy policy text. *School of Computer Science Carnegie Mellon University* (2018).
[29] Edward Loper and Steven Bird. 2002. NLTK: the natural language toolkit. *arXiv preprint cs/0205028* (2002).
[30] Aleecia M McDonald and Lorrie Faith Cranor. 2008. The cost of reading privacy policies. *Isjlp* 4 (2008), 543.
[31] Sam Meredith. 2018. Facebook-Cambridge Analytica: A timeline of the data hijacking scandal. https://www.cnbc.com/2018/04/10/facebook-cambridge-analytica-a-timeline-of-the-data-hijacking-scandal.html
[32] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient Estimation of Word Representations in Vector Space. In *1st International Conference on Learning Representations, ICLR*.
[33] Stephen P. Mulligan, Wilson C. Freeman, and Chris D. Linebaugh. 2019. https://fas.org/sgp/crs/misc/R45631.pdf
[34] President's Council of Advisors on Science and Technology. May 2014. Big data and privacy: A technological perspective. Report to the President, Executive Office of the President. *PCAST Big Data and Privacy* (May 2014).
[35] Sara P. 2020. Privacy Policies are Mandatory by Law. https://www.termsfeed.com/blog/privacy-policy-mandatory-law/
[36] Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, et al. 2011. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research* 12 (2011), 2825–2830.
[37] Ashwini Rao, Florian Schaub, Norman M. Sadeh, Alessandro Acquisti, and Ruogu Kang. 2016. Expecting the Unexpected: Understanding Mismatched Privacy Expectations Online. In *Twelfth Symposium on Usable Privacy and Security, SOUPS*. USENIX Association, 77–96.
[38] Radim Rehrek and Petr Sojka. 2011. Gensim—statistical semantics in python. *Retrieved from genism. org* (2011).
[39] CNN Editorial Research. 2020. 2016 Presidential Campaign Hacking Fast Facts. https://edition.cnn.com/2016/12/26/us/2016-presidential-campaign-hacking-fast-facts/index.html
[40] Jürgen Schmidhuber. 2015. Deep learning in neural networks: An overview. *Neural networks* 61 (2015), 85–117.
[41] Selenium. 2020. SeleniumHQ Browser Automation. https://www.selenium.dev/
[42] SKLearn. 2020. SKLearn: Count Vectorizer. shorturl.at/dsCH0
[43] Rob Sobers. 2020. How Privacy Policies Have Changed Since GDPR. https://www.varonis.com/blog/gdpr-privacy-policy/
[44] Daniel Solove. 2006. A Brief History of Information Privacy Law. *GW Law Faculty Publications & Other Works* (07 2006).
[45] Jakob Uszkoreit. 2017. Transformer: A novel neural network architecture for language understanding. *Google AI Blog* 31 (2017).
[46] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is All you Need. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems*. 5998–6008.
[47] VirusTotal. 2020. VirusTotal. https://www.virustotal.com/
[48] Strother H Walker and David B Duncan. 1967. Estimation of the probability of an event as a function of several independent variables. *Biometrika* 54, 1-2 (1967), 167–179.
[49] Charlie Warzel and Ash Ngu. 2019. Google's 4,000-Word Privacy Policy Is a Secret History of the Internet. https://www.nytimes.com/interactive/2019/07/10/opinion/google-privacy-policy.html
[50] Wikipedia. 2020. Wikipedia. https://wikipedia.org
[51] Shomir Wilson, Florian Schaub, 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 (Volume 1: Long Papers)*.
[52] Shomir Wilson, Florian Schaub, Rohan Ramanath, Norman Sadeh, Fei Liu, Noah A Smith, and Frederick Liu. 2016. Crowdsourcing Annotations for Websites' Privacy Policies: Can It Really Work?. In *Proceedings of the 25th International Conference on World Wide Web, WWW*. ACM, 133–143.
[53] Yonghui Wu, Mike Schuster, Zhifeng Chen, Quoc V Le, Mohammad Norouzi, Wolfgang Macherey, Maxim Krikun, Yuan Cao, Qin Gao, Klaus Macherey, et al. 2016. Google's Neural Machine Translation System: Bridging the Gap between Human and Machine Translation.
[54] Razieh Nokhbeh Zaeem, Rachel L. German, and K. Suzanne Barber. 2018. PrivacyCheck: Automatic Summarization of Privacy Policies Using Data Mining. *ACM Trans. Internet Techn.* 18, 4 (2018), 53:1–53:18.
[55] Wei Zhang et al. 1988. Shift-invariant pattern recognition neural network and its optical architecture. In *Proceedings of annual conference of the Japan Society of Applied Physics*.
[56] Sebastian Zimmeck and Steven M. Bellovin. 2014. Privee: An Architecture for Automatically Analyzing Web Privacy Policies. In *Proceedings of the 23rd USENIX Security Symposium*. 1–16.

A APPENDIX

Ensemble Variations Evaluation. We report the performance of the implemented ensembleo variations of the learning algorithms and word representations on each privacy policy category for OPP-115 dataset in Tables 5 and 6 for the document-based data splitting and segment-based data splitting, respectively. We also report the best performing representation variance for each models are shown in Tables 7 and 8.

Webshrinker Topic Taxonomy. The topic-based privacy policy reporting analysis requires the extraction of the domain topic of each website. To do so, we obtained the topic and field of interest of each website in Alexa top-10,000 list using Webshrinker API [17]. The obtained topics are fine-grained, with overall 66 topics. We manually grouped them into 18 main topics as shown in Table 9. The number in the parentheses indicates the total number of websites under each topic. Notice that "Technology & Computing" consists of 17 topics, and a total of 1,576 websites.

TLDR Segment Annotation. The TLDR ensemble labels each segment as positive or negative per privacy policy category in a multi-label fashion. The positive label indicates the presence of information regarding the category in the segment, while negative label indicate the absence of such information. Table 10 shows selected segments from popular websites in Alexa top-10,000 list, and their corresponding labels. Segments that are labeled as positive are the segments of interest, which the user should read to obtain information about privacy policy practices in regard to certain category.

Table 5: Document-based Evaluation: The performance of the learning algorithms with all used word representations on each privacy policy category on OPP-115 dataset.

Model	Word Representation	Category-1			Category-2			Category-3			Category-4			Category-5			Category-6			Category-7			Category-8			Category-9		
		P	R	F1	P	R	F1	P	R	F1	P	R	F1	P	R	F1	P	R	F1	P	R	F1	P	R	F1	P	R	F1
LR	Word Mapping	0.67	0.68	0.67	0.65	0.79	0.71	0.85	0.53	0.65	0.72	0.38	0.50	0.67	0.43	0.52	0.81	0.55	0.66	1.00	0.73	0.84	0.80	0.80	0.80	0.80	0.75	0.77
	Count Vectorizer	0.76	0.81	0.78	0.76	0.74	0.75	0.72	0.59	0.65	0.82	0.53	0.64	0.88	0.50	0.64	0.67	0.77	0.71	0.91	0.91	0.91	1.00	0.60	0.75	0.82	0.88	0.85
	TF-IDF	0.72	0.88	0.79	0.77	0.73	0.75	0.57	0.77	0.66	0.62	0.76	0.68	0.86	0.43	0.57	0.74	0.85	0.79	0.84	0.95	0.89	1.00	0.60	0.75	0.84	0.90	0.87
	Doc2Vec	0.64	0.81	0.72	0.64	0.76	0.69	0.69	0.55	0.61	0.58	0.53	0.55	0.80	0.29	0.42	0.85	0.60	0.70	0.85	0.77	0.81	0.71	1.00	0.83	0.76	0.62	0.68
	USE	0.75	0.91	0.82	0.64	0.82	0.72	0.90	0.55	0.68	0.55	0.94	0.70	0.28	0.71	0.40	0.60	0.89	0.72	1.00	0.59	0.74	0.60	0.60	0.60	0.74	0.94	0.83
RF	Word Mapping	0.78	0.69	0.73	0.78	0.65	0.71	0.73	0.50	0.59	0.76	0.38	0.51	0.75	0.43	0.55	0.81	0.53	0.64	1.00	0.68	0.81	1.00	0.80	0.89	0.82	0.87	0.84
	Count Vectorizer	0.86	0.70	0.77	0.88	0.67	0.76	0.92	0.35	0.51	1.00	0.26	0.42	1.00	0.29	0.44	0.93	0.60	0.73	1.00	0.68	0.81	1.00	0.40	0.57	0.90	0.83	0.86
	TF-IDF	0.86	0.75	0.80	0.89	0.65	0.75	0.92	0.35	0.51	1.00	0.29	0.45	1.00	0.29	0.44	0.96	0.57	0.72	1.00	0.68	0.81	1.00	0.40	0.57	0.90	0.83	0.86
	Doc2Vec	0.87	0.39	0.54	0.90	0.26	0.40	1.00	0.03	0.06	0.00	0.00	0.00	1.00	0.07	0.13	1.00	0.09	0.16	1.00	0.23	0.37	0.00	0.00	0.00	1.00	0.13	0.24
	USE	0.89	0.54	0.67	0.86	0.36	0.51	0.92	0.33	0.49	1.00	0.09	0.16	0.00	0.00	0.00	1.00	0.32	0.48	1.00	0.55	0.71	0.00	0.00	0.00	0.92	0.63	0.75
SVM	Word Mapping	0.69	0.71	0.70	0.69	0.78	0.73	0.95	0.55	0.69	0.88	0.44	0.59	0.26	0.64	0.38	0.58	0.83	0.68	1.00	0.73	0.84	1.00	0.60	0.75	0.71	0.88	0.79
	Count Vectorizer	0.81	0.82	0.82	0.73	0.79	0.76	0.72	0.71	0.72	0.79	0.65	0.71	0.89	0.57	0.70	0.86	0.77	0.81	0.91	0.91	0.91	1.00	0.60	0.75	0.87	0.79	0.83
	TF-IDF	0.80	0.84	0.82	0.86	0.69	0.76	0.79	0.68	0.73	0.90	0.56	0.69	1.00	0.50	0.67	0.84	0.77	0.80	0.95	0.86	0.90	1.00	0.60	0.75	0.91	0.81	0.86
	Doc2Vec	0.72	0.82	0.76	0.68	0.71	0.70	0.58	0.68	0.63	0.75	0.53	0.62	0.73	0.57	0.64	0.78	0.68	0.73	0.83	0.91	0.87	1.00	0.80	0.89	0.97	0.58	0.72
	USE	0.80	0.90	0.85	0.71	0.81	0.76	0.89	0.61	0.72	0.83	0.71	0.76	0.75	0.43	0.55	0.83	0.70	0.76	0.95	0.82	0.88	1.00	0.60	0.75	0.88	0.88	0.88
CNN	Word Mapping	0.67	0.70	0.68	0.62	0.65	0.64	0.54	0.52	0.53	0.59	0.47	0.52	0.33	0.43	0.38	0.71	0.57	0.64	0.28	0.77	0.41	0.75	0.60	0.67	0.59	0.90	0.71
	Count Vectorizer	0.79	0.80	0.80	0.67	0.76	0.71	0.64	0.70	0.67	0.79	0.68	0.73	0.67	0.43	0.52	0.72	0.70	0.71	0.84	0.95	0.89	0.75	0.60	0.67	0.87	0.88	0.88
	TF-IDF	0.78	0.79	0.78	0.71	0.76	0.73	0.59	0.62	0.60	0.79	0.65	0.71	0.50	0.50	0.50	0.82	0.68	0.74	0.79	0.86	0.83	1.00	0.60	0.75	0.78	0.87	0.82
	Doc2Vec	0.65	0.71	0.67	0.63	0.66	0.65	0.67	0.44	0.53	0.78	0.41	0.54	1.00	0.21	0.35	0.79	0.55	0.65	0.90	0.82	0.86	1.00	0.80	0.89	0.80	0.71	0.76
	USE	0.31	1.00	0.47	0.22	1.00	0.37	0.00	0.00	0.00	0.48	0.85	0.61	0.00	0.00	0.00	0.82	0.66	0.73	0.86	0.82	0.84	0.01	1.00	0.02	0.07	1.00	0.14
DNN	Word Mapping	0.69	0.68	0.69	0.67	0.81	0.74	0.39	0.71	0.51	0.41	0.74	0.53	0.36	0.64	0.46	0.49	0.77	0.60	0.15	0.95	0.26	0.50	0.60	0.55	0.55	0.92	0.69
	Count Vectorizer	0.78	0.85	0.81	0.77	0.69	0.73	0.58	0.62	0.60	0.77	0.68	0.72	0.56	0.36	0.43	0.73	0.68	0.70	0.88	0.95	0.91	1.00	0.60	0.75	0.86	0.81	0.83
	TF-IDF	0.75	0.86	0.80	0.71	0.74	0.72	0.61	0.62	0.62	0.66	0.68	0.67	0.29	0.50	0.37	0.69	0.66	0.67	0.88	0.95	0.91	1.00	0.60	0.75	0.87	0.79	0.83
	Doc2Vec	0.70	0.77	0.73	0.66	0.70	0.68	0.57	0.64	0.60	0.49	0.56	0.52	0.47	0.50	0.48	0.70	0.68	0.69	0.56	0.82	0.67	0.80	0.80	0.80	0.68	0.77	0.72
	USE	0.82	0.82	0.82	0.72	0.79	0.76	0.56	0.74	0.64	0.69	0.74	0.71	0.33	0.29	0.31	0.59	0.81	0.68	0.62	0.95	0.75	1.00	0.60	0.75	0.87	0.90	0.89
BERT	WordPiece	0.94	0.94	0.94	0.90	0.89	0.89	0.88	0.82	0.85	0.90	0.92	0.91	0.95	0.82	0.87	0.92	0.85	0.88	0.95	0.95	0.95	1.00	1.00	1.00	0.93	0.95	0.94

Table 6: Segment-based Evaluation: The performance of the learning algorithms with all used word representations on each privacy policy category on OPP-115 dataset.

Model	Word Representation	Category-1			Category-2			Category-3			Category-4			Category-5			Category-6			Category-7			Category-8			Category-9		
		P	R	F1	P	R	F1	P	R	F1	P	R	F1	P	R	F1	P	R	F1	P	R	F1	P	R	F1	P	R	F1
LR	Word Mapping	0.67	0.64	0.66	0.67	0.74	0.70	0.85	0.55	0.67	0.79	0.38	0.51	0.64	0.47	0.54	0.79	0.54	0.64	1.00	0.74	0.85	0.83	0.83	0.83	0.81	0.72	0.76
	Count Vectorizer	0.75	0.76	0.76	0.79	0.67	0.72	0.61	0.72	0.66	0.73	0.55	0.63	0.89	0.53	0.67	0.71	0.78	0.74	0.81	0.96	0.88	1.00	0.67	0.80	0.83	0.88	0.85
	TF-IDF	0.72	0.88	0.79	0.77	0.73	0.75	0.57	0.77	0.66	0.62	0.76	0.68	0.86	0.43	0.57	0.74	0.85	0.79	0.84	0.95	0.89	1.00	0.60	0.75	0.84	0.90	0.87
	Doc2Vec	0.65	0.76	0.70	0.78	0.60	0.67	0.68	0.55	0.61	0.61	0.48	0.54	0.63	0.33	0.43	0.76	0.63	0.69	0.82	0.78	0.80	1.00	0.67	0.80	0.76	0.65	0.70
	USE	0.76	0.89	0.82	0.67	0.78	0.72	0.90	0.54	0.67	0.46	0.97	0.62	0.20	0.60	0.30	0.92	0.56	0.70	1.00	0.61	0.76	0.57	0.67	0.62	0.74	0.95	0.83
RF	Word Mapping	0.77	0.63	0.69	0.77	0.60	0.68	0.76	0.49	0.60	0.85	0.38	0.52	0.78	0.47	0.58	0.82	0.56	0.67	1.00	0.70	0.82	1.00	0.83	0.91	0.84	0.85	0.84
	Count Vectorizer	0.85	0.64	0.73	0.87	0.61	0.72	0.85	0.31	0.45	1.00	0.21	0.34	1.00	0.27	0.42	0.93	0.61	0.74	1.00	0.65	0.79	1.00	0.33	0.50	0.92	0.80	0.86
	TF-IDF	0.86	0.67	0.75	0.90	0.59	0.71	0.89	0.34	0.49	1.00	0.28	0.43	1.00	0.27	0.42	0.93	0.63	0.75	1.00	0.70	0.82	1.00	0.33	0.50	0.89	0.83	0.86
	Doc2Vec	0.78	0.39	0.52	0.88	0.24	0.38	1.00	0.03	0.05	0.00	0.00	0.00	0.00	0.00	0.00	1.00	0.07	0.14	1.00	0.17	0.30	0.00	0.00	0.00	1.00	0.13	0.24
	USE	0.87	0.54	0.67	0.90	0.32	0.48	0.92	0.32	0.48	1.00	0.07	0.13	0.00	0.00	0.00	1.00	0.34	0.51	1.00	0.57	0.72	0.00	0.00	0.00	0.93	0.67	0.78
SVM	Word Mapping	0.68	0.66	0.67	0.69	0.72	0.70	0.95	0.54	0.68	0.88	0.48	0.62	0.27	0.60	0.38	0.56	0.83	0.67	1.00	0.74	0.85	1.00	0.67	0.80	0.72	0.87	0.79
	Count Vectorizer	0.81	0.74	0.77	0.78	0.72	0.75	0.71	0.72	0.71	0.74	0.69	0.71	0.89	0.53	0.67	0.82	0.76	0.78	0.88	0.91	0.89	1.00	0.67	0.80	0.89	0.78	0.83
	TF-IDF	0.78	0.80	0.79	0.78	0.73	0.76	0.78	0.69	0.73	0.76	0.66	0.70	0.80	0.53	0.64	0.88	0.71	0.78	0.95	0.83	0.88	1.00	0.67	0.80	0.93	0.83	0.88
	Doc2Vec	0.73	0.77	0.75	0.68	0.70	0.69	0.58	0.68	0.62	0.73	0.55	0.63	0.67	0.53	0.59	0.92	0.59	0.72	1.00	0.70	0.82	1.00	0.83	0.91	0.75	0.72	0.74
	USE	0.81	0.85	0.83	0.74	0.76	0.75	0.90	0.62	0.73	0.90	0.62	0.73	1.00	0.47	0.64	0.79	0.76	0.78	0.95	0.83	0.88	1.00	0.67	0.80	0.90	0.90	0.90
CNN	Word Mapping	0.67	0.66	0.66	0.69	0.72	0.71	0.44	0.59	0.51	0.36	0.62	0.46	0.36	0.33	0.34	0.47	0.66	0.55	0.30	0.74	0.43	0.80	0.67	0.73	0.68	0.83	0.75
	Count Vectorizer	0.75	0.79	0.77	0.72	0.69	0.70	0.69	0.62	0.65	0.72	0.72	0.72	0.29	0.53	0.37	0.64	0.71	0.67	0.81	0.91	0.86	0.67	0.67	0.67	0.89	0.85	0.87
	TF-IDF	0.81	0.75	0.78	0.67	0.70	0.68	0.63	0.63	0.63	0.75	0.72	0.74	0.75	0.40	0.52	0.68	0.66	0.67	0.84	0.91	0.88	1.00	0.67	0.80	0.86	0.80	0.83
	Doc2Vec	0.71	0.73	0.72	0.66	0.61	0.63	0.57	0.51	0.54	0.71	0.52	0.60	0.83	0.33	0.48	0.81	0.54	0.65	0.90	0.78	0.84	0.83	0.83	0.83	0.84	0.68	0.75
	USE	0.72	0.86	0.78	0.00	0.00	0.00	0.09	1.00	0.17	0.04	1.00	0.07	0.02	1.00	0.04	0.78	0.78	0.78	0.03	1.00	0.06	0.01	1.00	0.02	0.00	0.00	0.00
DNN	Word Mapping	0.69	0.71	0.70	0.74	0.68	0.71	0.43	0.79	0.56	0.38	0.83	0.52	0.18	0.67	0.28	0.53	0.68	0.60	0.42	0.96	0.58	0.04	0.83	0.07	0.78	0.83	0.81
	Count Vectorizer	0.80	0.77	0.79	0.78	0.69	0.73	0.68	0.55	0.61	0.83	0.69	0.75	0.33	0.40	0.36	0.60	0.76	0.67	0.87	0.87	0.87	1.00	0.67	0.80	0.88	0.83	0.85
	TF-IDF	0.78	0.78	0.78	0.79	0.70	0.74	0.57	0.62	0.59	0.68	0.72	0.70	0.45	0.33	0.38	0.70	0.63	0.67	0.84	0.91	0.88	1.00	0.67	0.80	0.87	0.78	0.82
	Doc2Vec	0.72	0.68	0.70	0.70	0.74	0.72	0.50	0.68	0.57	0.59	0.59	0.59	0.42	0.53	0.47	0.56	0.76	0.65	0.76	0.83	0.79	1.00	0.83	0.91	0.65	0.80	0.72
	USE	0.81	0.80	0.81	0.72	0.74	0.73	0.61	0.72	0.66	0.58	0.79	0.67	0.29	0.47	0.36	0.60	0.78	0.68	0.83	0.87	0.85	1.00	0.67	0.80	0.84	0.88	0.86
BERT	WordPiece	0.92	0.92	0.92	0.90	0.89	0.90	0.90	0.87	0.88	0.86	0.94	0.90	0.99	0.70	0.78	0.88	0.88	0.88	1.00	0.96	0.98	1.00	1.00	1.00	0.91	0.96	0.93

Table 7: Segment-based Evaluation: The best performing configuration of the learning algorithms on each privacy policy category on OPP-115 dataset.

Categories	LR			RF			SVM			CNN			DNN			BERT		
	P	R	F1	P	R	F1	P	R	F1	P	R	F1	P	R	F1	P	R	F1
First Party Collection/Use	0.76	0.89	0.82	0.86	0.67	0.75	0.81	0.85	0.83	0.72	0.86	0.78	0.81	0.80	0.81	0.92	0.92	**0.92**
Third Party Sharing/Collection	0.77	0.73	0.75	0.87	0.61	0.72	0.78	0.73	0.76	0.69	0.72	0.71	0.79	0.70	0.74	0.90	0.89	**0.90**
User Choice/Control	0.90	0.54	0.67	0.76	0.49	0.60	0.90	0.62	0.73	0.69	0.62	0.65	0.61	0.72	0.66	0.90	0.87	**0.88**
User Access, Edit and Deletion	0.62	0.76	0.68	0.85	0.38	0.52	0.90	0.62	0.73	0.75	0.72	0.74	0.83	0.69	0.75	0.86	0.94	**0.90**
Data Retention	0.89	0.53	0.67	0.78	0.47	0.58	0.89	0.53	0.67	0.75	0.40	0.52	0.42	0.53	0.47	0.99	0.70	**0.78**
Data Security	0.74	0.85	0.79	0.93	0.63	0.75	0.82	0.76	0.78	0.78	0.78	0.78	0.60	0.78	0.68	0.88	0.88	**0.88**
Policy Change	0.84	0.95	0.89	1.00	0.70	0.82	0.88	0.91	0.89	0.84	0.91	0.88	0.84	0.91	0.88	1.00	0.96	**0.98**
Do Not Track	0.83	0.83	0.83	1.00	0.83	0.91	1.00	0.83	0.91	0.83	0.83	0.83	1.00	0.83	0.91	1.00	1.00	**1.00**
Specific Audiences	0.84	0.90	0.87	0.89	0.83	0.86	0.90	0.90	0.90	0.89	0.85	0.87	0.84	0.88	0.86	0.91	0.96	**0.93**
Overall	0.80	0.78	0.78	0.88	0.62	0.72	0.87	0.75	0.80	0.77	0.75	0.75	0.75	0.76	0.75	0.93	0.90	**0.91**

Table 8: Document-based Evaluation: The best performing configuration of the learning algorithms on each privacy policy category on OPP-115 dataset.

Categories	LR			RF			SVM			CNN			DNN			BERT		
	P	R	F1	P	R	F1	P	R	F1	P	R	F1	P	R	F1	P	R	F1
First Party Collection/Use	0.75	0.91	0.82	0.86	0.75	0.80	0.80	0.90	0.85	0.79	0.80	0.80	0.82	0.82	0.82	0.94	0.94	**0.94**
Third Party Sharing/Collection	0.77	0.73	0.75	0.88	0.67	0.76	0.86	0.69	0.76	0.71	0.76	0.73	0.72	0.79	0.76	0.90	0.89	**0.89**
User Choice/Control	0.90	0.55	0.68	0.73	0.50	0.59	0.79	0.68	0.73	0.64	0.70	0.67	0.56	0.74	0.64	0.88	0.82	**0.85**
User Access, Edit and Deletion	0.55	0.94	0.70	0.76	0.38	0.51	0.83	0.71	0.76	0.79	0.68	0.73	0.77	0.68	0.72	0.90	0.92	**0.91**
Data Retention	0.88	0.50	0.64	0.75	0.43	0.55	0.89	0.57	0.70	0.67	0.43	0.52	0.47	0.50	0.48	0.95	0.82	**0.87**
Data Security	0.74	0.85	0.79	0.93	0.60	0.73	0.86	0.77	0.81	0.82	0.68	0.74	0.73	0.68	0.70	0.92	0.85	**0.88**
Policy Change	0.91	0.91	0.91	1.00	0.68	0.81	0.91	0.91	0.91	0.84	0.95	0.89	0.88	0.95	0.91	0.95	0.95	**0.95**
Do Not Track	0.71	1.00	0.83	1.00	0.80	0.89	1.00	0.80	0.89	1.00	0.80	0.89	0.80	0.80	0.80	1.00	1.00	**1.00**
Specific Audiences	0.84	0.90	0.87	0.90	0.83	0.86	0.88	0.88	0.88	0.87	0.88	0.88	0.87	0.90	0.89	0.93	0.95	**0.94**
Overall	0.78	0.81	0.78	0.87	0.63	0.72	0.87	0.77	0.81	0.79	0.74	0.76	0.73	0.76	0.75	0.93	0.90	**0.91**

Table 9: The Webshrinker taxonomy. Webshrinker returned 66 different topics, which we categorized into 18 main topics and used in our analysis.

i	Main Topic	Subtopics	i	Main Topic	Subtopics
1	Technology & Computing (1576)	Technology & Computing (1446)	34	News & Weather Information (434)	
2		Web Search (69)	35	Education (482)	Education (423)
3		Unmoderated UGC/Message Boards (26)	36		Science (33)
4		Internet Technology (14)	37		Family & Parenting (13)
5		Content Server (5)	38		K-6 Education (7)
6		Web Design/HTML (4)	39		Books & Literature (5)
7		Computer Peripherals (2)	40		Language Translation (1)
8		Computer Certification (1)	41	Shopping (295)	Shopping (288)
9		Computer Reviews (1)	42		Marketing (3)
10		Cell Phones (1)	43		Couponing (3)
11		Desktop Publishing (1)	44		Beauty (1)
12		Unix/Linux (1)	45	Food (76)	Food & Drink (75)
13		Databases (1)	46		Italian Cuisine (1)
14		VPNs/Proxies & Filter Avoidance (1)	47	Hobbies & Interests(571)	Hobbies & Interests (469)
15		Net Conferencing (1)	48		Home & Garden (46)
16		Antivirus Software (1)	49		Automotive (41)
17		Computer Networking (1)	50		Pets (15)
18	Business (723)	Business (587)	51	Travel (69)	
19		Careers (88)	52	Adult Content (134)	
20		Real Estate (35)	53	AHealth & Fitness (128)	
21		Buying/Selling Cars (4)	54	Style & Fashion (112)	
22		Investing (2)	55	Law, Government, & Politics (103)	Law & Government (101)
23		Auto Parts (1)	56		Politics (1)
24		Stocks (1)	57		Immigration (1)
25		Job Search (1)	58	Sports (99)	Sports (98)
26		Hotels (1)	59		Basketball (1)
27		Buying/Selling Homes (1)	60	Society (89)	
28		Career Advice (1)	61	Streaming Media (62)	
29		Business Software (1)	62	Other (124)	Illegal Content (40)
30	Personal Finance (277)		63		Uncategorized (30)
31	Arts & Entertainment(244)	Arts & Entertainment (236)	64		Under Construction (29)
32		Television & Video (7)	65		Religion & Spirituality (24)
33		Desktop Video (1)	66		Weapons (1)

Table 10: Different Alexa top-10,000 websites extracted positive and negative segments per privacy policy categories. The positive segments represent the paragraphs of interest for the user (to read).

Category	Website	Label	Example
First Party Collection/Use	twitter.com	0	what you tweet becomes instantly public worldwide on twitter, and can appear on other media like websites, newspapers, or television.
		1	whenever you use twitter, twitter gets a general sense of your location from things like the network you're using to access the internet. this helps twitter to provide you with relevant content.
Third Party Sharing/Collection	tumblr.com	0	we use information about how you interact with tumblr in order to personalize it for you, to keep both you and our community safe from harm, and to improve tumblr for everyone who uses it.
		1	in order to provide you with all of this tumblr content for free, tumblr selectively runs advertisements. we, and our parent company, automattic, work with online advertising companies to provide you with advertising that is as relevant and useful as possible. to make our ads more relevant and useful, we make educated guesses about your interests based on your activity on our sites and services. the ads we show you often reflect these interests and educated guesses.
User Choice/Control	microsoft.com	0	if you sign into a service offered by a third party with your microsoft account, you will share with that third party the account data required by that service.
		1	you have a variety of tools to control the data collected by cookies, web beacons, and similar technologies. for example, you can use controls in your internet browser to limit how the websites you visit are able to use cookies and to withdraw your consent by clearing or blocking cookies.
User Access, Edit, & Deletion	google.com	0	when you're signed in and interact with some google services, like leaving comments on a youtube video or reviewing an app in play, your name and photo appear next to your activity. we may also display this information in ads depending on your shared endorsements setting.
		1	when you delete data, we follow a deletion process to make sure that your data is safely and completely removed from our servers or retained only in anonymized form. we try to ensure that our services protect information from accidental or malicious deletion. because of this, there may be delays between when you delete something and when copies are deleted from our active and backup systems.
Data Retention	github.com	0	please note, github may share your username, usage information, and device information with the owner(s) of the organization you are a member of, to the extent that your user personal information is provided only to investigate or respond to a security incident that affects or compromises the security of that particular organization.
		1	if you would like to cancel your account or delete your user personal information, you may do so in your user profile. we retain and use your information as necessary to comply with our legal obligations, resolve disputes, and enforce our agreements, but barring legal requirements, we will delete your full profile (within reason) within 90 days of your request. you may contact github support or github premium support to request the erasure of the data we process on the basis of consent within 30 days.
Data Security	bankofamerica.com	0	by providing your mobile number you are consenting to receive a text message. text message fees may apply from your carrier. text messages may be transmitted automatically.
		1	we've teamed up with ibm to offer trusteer rapport — online fraud protection software available for bank of america customers. trusteer rapport delivers extra security while you're signed in to our site.
Policy Change	yahoo.com	0	personal data collected by the services is shared with oath, a verizon affiliate that is home to media, technology, and communication brands. oath's privacy policy governs its use of that information.
		1	the company may update this privacy policy. we will notify you about materially significant changes in the way we treat previously collected personal data by sending a notice to the primary email address specified in your account or by placing a prominent notice on our site.
Do Not Track	wikipedia.org	0	we use a variety of commonly-used technologies, like cookies, to understand how you use the wikimedia sites, make our services safer and easier to use, and to help create a better and more customizable experience for you.
		1	because we protect all users in this manner, we do not change our behavior in response to a web browser's "do not track" signal.
Specific Audiences	twitch.tv	0	the twitch services may link to third-party websites or services. the privacy practices of those third parties are not governed by this privacy notice. we encourage you to review the privacy policies of these third-party websites and services to understand their practices.
		1	protecting the privacy of young children is especially important. for that reason, twitch does not knowingly collect or maintain personal information (as defined by the united states children's online privacy protection act) from persons under 13 years-of-age. if twitch learns that personal information of persons under 13 has been collected on or through the twitch services, twitch will take appropriate steps to delete this information.