Evolution of Composition, Readability, and Structure of Privacy Policies over Two Decades

Andrick Adhikari
University of Denver
Denver, Colorado, USA
andrick.adhikari@du.edu

Sanchari Das
University of Denver
Denver, Colorado, USA
sanchari.das@du.edu

Rinku Dewri
University of Denver
Denver, Colorado, USA
rinku.dewri@du.edu

ABSTRACT

Privacy policies outline data collection and sharing practices followed by an organization, together with choice and control measures available to users to manage the process. However, users have often needed help reading and understanding such documents, regardless of their being written in a natural language. The fundamental problems with privacy policies persist despite advancements in privacy design, frameworks, and regulations. To identify the causes of privacy policies being persistently challenging to comprehend, it is vital to investigate historical policy patterns and understand the evolution of privacy policies concerning information packaging and presentation. To this aid, we create a sentence-level classifier to conduct a large-scale longitudinal analysis on different privacy policies from 130,604 organizations, totaling approximately one million policies from 1997 to 2019. We annotate 10,717 sentences from 115 policies in the OPP-115 corpus to implement the classifier and then use those annotations to train the XLNet and BERT classifiers. Results from our analysis reveal that specific data practice categories experience more frequent policy changes than others, making it challenging to track relevant information over time. In addition, we discover that every category has distinct composition, readability, and structural issues, which exacerbate when categories frequently co-occur in a document. Based on our observations, we provide recommendations for policy articulation and revision to make privacy policy documents conform to better coherence and structure.

KEYWORDS

Sentence Classification, Longitudinal Analysis, Neural Networks, Privacy Policies, Information Organization, Change Detection.

1 INTRODUCTION

Privacy policies are legal documents that communicate practices relating to consumer data collection, management, use, and sharing. They serve as the primary means to inform users about the data collected from them and the controls that are provided to manage this process while being compliant with associated rules and standards [19]. However, despite being written in a natural language, there are several obstacles with privacy policies that prevent the general public from effectively utilizing them to make informed privacy-related decisions. Along these lines, the readability and clarity of privacy policies are significant concerns and are often introduced by how policies are written and structured. The average length of privacy policies is over 2500 words, and they are typically difficult to read and comprehend [43]. This makes users less likely to try to read or understand what is written in the policies.

In order to improve the utilization of privacy policies, recent advancements in automation, machine learning, and deep learning have led to the creation of tools that let users access the information contained in a privacy policy without requiring additional help from policy writers [55]. Classification of privacy policy texts is the most popular approach in this context, which enables users to selectively obtain a high-level overview of a policy in terms of pre-defined category labels [4, 29, 46, 47, 55, 66]. The labels encompass presumed data practices of consumer importance, such as data collection, sharing, choice and control, regulatory conformance, security, and data retention, among others. However, while the automated classification of policies helps with readability and comprehension, the inherent issues with privacy policies persist. Prior studies have highlighted the challenges associated with readability [23, 24, 34, 44, 65], ambiguity [40, 53, 54], and accessibility [28, 32, 34] of information. The difficulties associated with information presentation have also changed over time, with frequent policy revisions required to reflect changes in data practices. Additionally, each category of information in a policy uses a different style of language and has its unique set of issues. Depending on how the categories are organized in a policy, these categorical descriptions, when combined, create the challenges and issues that privacy policies display.

Categorization of policy texts is yet to be used to investigate the particular problems related to each category, how these categories interact with one another, and how they have changed over time. Therefore, it is essential to look into historical policy patterns to identify the causes of privacy policies being persistently challenging to grasp. Thus, this study explores the following questions to examine how privacy policies have evolved over the last two decades.

- How has categorical information evolved in privacy policies over time?
- What is the general pattern for information coherence and organization within each category?
- How does different types of information in a policy interact with one another to communicate data practice information?
- What particular aspects of information gain prominence as privacy issues become more concerning?

We aim to address these questions by leveraging a sentence-level classifier that can individually categorize the sentences in a privacy policy. First, the classifier is applied to over one million policies,

Proceedings on Privacy Enhancing Technologies 2023(3), 138–153

https://doi.org/10.56553/popets-2023-0074

spread over 130, 604 organizations, dating from 1997 to 2019. After that, we analyze each policy's categorical information to understand how privacy policy semantics and categorical structures have changed over time. We then identify the general trends in information organization and coherence within each category and whether they have made policies more accessible. Through this process, we contribute to expanding our understanding of the challenges in the domain of usable privacy policies, summarized in the following items.

(1) This work contributes to the advancement in knowledge of the semantic and categorical evolution of privacy policies, as determined by current trends in policy articulation and amendment. Our extensive longitudinal investigation from a categorical information standpoint reveals aspects of privacy policy trends.
(2) We identify practices in place that hinder effective notice and choice through analysis of the trend in readability, coherence, entropy, and inter-category relation over time, as introduced by policy revisions.
(3) Our results reveal category-specific problems and the inter-category relationships that compound categorical problems, thereby contributing to the complexity of privacy policies. We also assess the results within the purview of the General Data Protection Regulation (GDPR), which demonstrates that information completeness does not imply effective policy communication.
(4) We draw attention to the classification of privacy policies at the sentence level when examining privacy policies and exemplify how such a granular analysis can reveal crucial factors that influence the usability of a privacy policy.

Note that we analyze the largest existing privacy policy corpus with an automated approach to draw comprehensive elemental evolution patterns and generate a much closer representation of policy trends than that from a selectively chosen smaller corpus, which may not include all aspects of privacy policy articulations in existence over the years and risk inaccuracy in representations due to selective study. Preparing a corpus for qualitative study to obtain any statistically significant evolution patterns would mean manually annotating a significant portion of policies across different websites and multiple years. We believe quantitative analysis with a classifier is better suitable for a task of such magnitude.

The remainder of the paper is organized as follows. Section 2 presents related work in the field. The methodology and evaluation metrics we use in the analysis are described in Section 3. Section 4 presents the observed tendencies in privacy policies, followed by a discussion in Section 5. Finally, we conclude in Section 7 after discussing limitations and future work in Section 6.

2 RELATED WORK

2.1 Privacy Policy Challenges

Privacy policies face accessibility, readability, comprehension, and ambiguity challenges from various stakeholder perspectives. Evaluation of privacy policies using empirical readability metrics reveals that the majority of the population is unable to comprehend privacy policies, which calls for at least a college-level reading proficiency [23, 24, 34, 44]. The number of different ways a policy can be interpreted makes privacy policies ambiguous. In order to give organizations more flexibility, policymakers frequently use ambiguity or vagueness [54]. Occasionally, even legal and policy professionals need help to agree on a clear-cut policy composition [53, 54]. Ambiguity is further introduced by the need to have complete information in these policies [13, 40]. In addition, many users need help retrieving specific information from a policy [28], and sometimes locate where the policy is mentioned in a website [30, 32, 34]. Most of these prior works point to critical challenges of privacy policies; however, they focused on analyzing a single snapshot of the policies. Our longitudinal analysis explores if the challenges have been persistent over time.

2.2 Privacy Regulations and Recommendations

Several agencies have proposed methods and regulations to increase the transparency of data access practices in an organization [15]. For instance, the National Telecommunications and Information Administration offers guidelines for establishing policies for mobile applications [48], while the European Article 29 Working Party offers recommendations for IoT devices [7]. On the other hand, the General Data Protection Regulation (GDPR), which applies to data processing platforms in Europe, requires greater transparency in privacy policies [64]. In addition, the Federal Trade Commission (FTC) recommends adopting clear and straightforward privacy laws [16]. Other U.S. national security regulations, including the Video Privacy Protection Act (VPPA), the GLBA and HIPAA Privacy Rules governing finance and healthcare, respectively, the COPPA rule governing children's information, and others, all place restrictions on how privacy policies are written [50].

Even with these privacy regulations, data protection principles sometimes need to be clarified, and manual conformance verification can be error-prone [15, 58, 62]. Additionally, the design of regulations overlooks the cognitive frame of personal intrusion to comprehend privacy issues [45, 50]. Solutions have been proposed to address such issues. However, these methods frequently experience a lack of acceptance, as is the case with P3P machine-readable privacy policies [2, 8, 17, 18] and their extensions, alternative privacy formats [14, 21, 26, 61], and graphical practice icons [20, 27, 35, 51]. Thus, it is imperative to explore how the policies are designed and what we can learn from them from a bird's eye perspective for the various stakeholders.

2.3 NLP for Privacy Policies

Natural language processing (NLP) has been adopted as the preferred approach to extract pertinent privacy information from a policy document. NLP solutions work directly on policies currently present in most organizations without requiring additional cooperation from the organization. In the NLP application field for privacy policies, a variety of subjects are addressed, including information extraction [6, 10–12, 31, 59], content summarization [69], automated question-answering [56], and information alignment [41, 52]. Among the NLP-based solutions, the classification of privacy policy texts is the most researched sub-domain.

Classification in Privacy Policies. Since the feasibility of text classification in privacy policies was established by Ammar et al. [4],

we have seen many studies aimed at enhancing policy classification models through the training and testing of various machine learning models [42, 66, 67]. Following the development of neural network and deep learning models, we have seen an increase in the use of these models in the privacy domain as well, to further advance the capabilities of segment (paragraph) categorization tools [29, 46, 47]. Although segment classifiers are used for most categorization in the privacy policy domain, there is no established method for segmentation. Additionally, segment classifications cannot adequately reflect the more precise information available at the sentence level [1]. However, few studies have looked at sentence classification. For the categorization of both segments and sentences, Liu et al. tested SVM, LR, and CNN models [42]. Other researchers have used sentence categorization to perform specific tasks, including determining whether a sentence describes a user choice instance [9, 36, 57, 60]. Although NLP-based methods like categorization aim to make information in current policies accessible, they may also aid in deconstructing privacy policies which collectively pose challenges to users. Thus, we analyze how privacy policy material has changed over time with respect to composition, readability, and structural changes.

3 METHOD

A privacy policy comprises categorical information, with each category conveying a specific type of information. We begin our study by implementing a sentence-level classifier that can identify the data practice category contained in each policy sentence. Compared to a paragraph-level classifier, a sentence-level classifier can provide a better overview of the information organization in a policy, especially when a paragraph can contain a mix of information [1].

3.1 Sentence Classification

3.1.1 Training Data. We use sentences from policies in the OPP-115 corpus to implement a sentence-level classifier. OPP-115 is a corpus of 115 website policies and has 12 high-level data practice categories annotated by legal experts [66]. The annotation scheme and tools were created after carefully considering labeling techniques for policy segments that crowd workers could use to produce in-depth policy annotations [67]. The high-level categories are "First-Party Collection/Use (FPCU)," "Third-Party Sharing/Collection (TPSC)," "User Choice/Control (UCC)," "User Access, Edit, and Deletion (UAED)," "Data Retention (DR)," "Data Security (DS)," "Policy Change (PC)," "Do Not Track (DNT)," "International and Specific Audiences (ISA)," and "Other." "Introductory/Generic (IG)," "Practice Not Covered (PNC)," and "Privacy Contact Information (PCI)" are the three subcategories that make up the "Other" category. "Practice Not Covered" refers to ambiguous descriptions that cannot be confidently tagged with any other category. Please refer to Appendix A for brief descriptions of these categories.

Since the OPP-115 corpus has only segment (paragraph) annotations and attributes annotations for partial sentences in a segment, we manually annotated the 10,717 sentences in OPP-115 with the 12 established high-level categories for segments. Table B1 lists the frequency of sentences in each category. The sentences were annotated by a trained qualitative researcher and any discrepancies were resolved by another annotator. For annotating the sentences, we did exclusive coding with single label categorization. Privacy policies are consumer facing and our goal is to analyze the effectiveness of these policies for users, thus our coding is not done from a legal perspective.

3.1.2 Learning Models: BERT and XLNet. We selected BERT and XLNet for training and evaluation to decide on our final sentence-level classifier. In recent studies, BERT and XLNet have both surpassed previously bench-marked CNN-based models such as Polisis [29] in policy classification [1, 46, 47]. Additionally, pre-trained models for BERT and XLNet can be fine-tuned with the downstream task, such as training a custom word embedding. BERT is a transformers-based deep learning model that can accurately capture the contextual relationships between words and subwords [22, 63]. An encoder reads the text input for the transformers, and then the task output is predicted by a decoder. BERT is by nature bidirectional in terms of encapsulating contexts since the entire string of words is read at once, capturing both the left and right context of each word. XLNet is also a transformer-based model that gathers context from both forward and backward directions but also considers all permutations of a sequence of tokens [68].

3.1.3 Training and Evaluation. We train BERT and XLNet classifiers using the FastBert package[1]. The training computer contained an Intel(R) Xeon(R) E5-1620v4 3.50GHz CPU, 8GB of RAM, and an NVIDIA RTX A5000 GPU. Transformer-based deep learning requires sufficient CUDA cores to train the models and to produce predictions. We used the 8,192 CUDA cores and 256 Tensor cores of the NVIDIA RTX A5000 GPU for the training. We modified the FastBert learner to utilize mixed precision training, which significantly boosts computational efficiency by using half-precision (16-bit) for most tasks and single precision (32-bit and 64-bit) in critical parts of the network. We followed the approach adopted by Mustapha et al. while configuring BERT and XLNet for training, with batch sizes of 8, a default learning rate of 10^{-3}, and a total of 5 training epochs [47]. A single training epoch in XLNet took 5.8 minutes in our hardware, whereas a single training epoch in BERT took 3.2 minutes.

We utilized a standard 10×9 nested cross-validation approach with a 9:1 train/test split for method selection. A 10×9 nested cross-validation utilizes 10 different train/test splits of the data set. Each training set is then used to perform a 9-fold cross-validation, and the best model (determined by a loss metric) gets evaluated on the test set. This gives us 10 estimates of the method's performance, which we average to determine a final value. The average precision (Pr), recall (Re), and F1-score (F1) of the two methods are shown in Table 1. Given that we trained 90 models for each method, the nested cross-validation for XLNet took ≈44 hours and for BERT ≈24 hours. While BERT has better precision than XLNet in a few categories, including "Data Retention," "Data Security," and "Do Not Track," the micro- and macro-averages demonstrate that XLNet marginally outperforms the BERT classifier. As a result, we selected XLNet as the method of choice for our sentence classifier. Finally, we trained a new instance of XLNet using all the annotated data and subsequently used this model for sentence-level classification.

[1] https://github.com/utterworks/fast-bert

Table 1: BERT and XLNet nested cross-validation performance scores. Pr: Precision, Re: Recall, F1: F1-score.

Category	BERT			XLNet		
	Pr	Re	F1	Pr	Re	F1
PC	0.87	0.89	0.88	0.88	0.93	0.90
TPSC	0.80	0.80	0.80	0.80	0.82	0.81
UCC	0.78	0.76	0.77	0.78	0.76	0.77
DR	0.72	0.63	0.67	0.66	0.70	0.68
UAED	0.70	0.68	0.69	0.71	0.74	0.72
PNC	0.51	0.50	0.51	0.54	0.53	0.53
ISA	0.88	0.88	0.88	0.88	0.90	0.89
DS	0.87	0.85	0.86	0.85	0.87	0.86
PCI	0.78	0.75	0.76	0.79	0.78	0.79
DNT	0.93	0.85	0.88	0.89	0.86	0.88
IG	0.66	0.60	0.63	0.69	0.60	0.65
FPCU	0.81	0.85	0.83	0.83	0.85	0.84
micro avg	**0.78**	**0.78**	**0.78**	**0.79**	**0.79**	**0.79**
macro avg	**0.78**	**0.75**	**0.76**	**0.78**	**0.78**	**0.78**

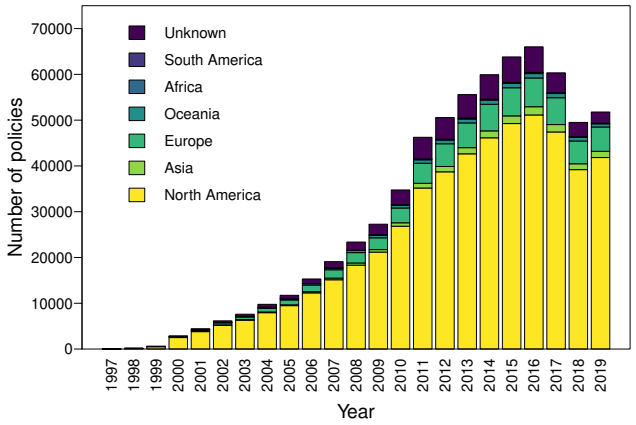

Figure 1: Number of policies per year in the PPCrawl corpus. Domains that did not return a result in ICANN lookup are placed in the *Unknown* geographic category.

We observe that categories such as "Introductory/Generic" and "Data Retention" have much fewer instances in the corpus. When used with nested cross-validation, the examples become more sparse in the training data. This can lead to poor performance of a model for low-frequency categories. As a result, observations relating to these categories may have a larger margin of error.

Our evaluation reveals that the classifier has high precision and recall for most categories. Table 1 shows a detailed breakdown of the classifier's performance with respect to different categories in the test data. For some categories such as "Privacy Contact Information", the classifier shows relatively lower performance in comparison to other categories. The lower performance in classification may be attributed to the fact that when linguistic characteristics pertaining to a specific category overlap with other categories, a classifier has difficulties resolving label ambiguities and creates misclassifications. Nonetheless, since the final XLNet model is trained on the entire corpus, we expect its performance to be "slightly" better.

3.2 Privacy Policy Evaluation

With the implemented XLNet sentence classifier, we investigate the evolution of content in privacy policies regarding categorical composition, readability, and structural changes over revisions.

3.2.1 Data for Analysis. In our study, we analyze the Princeton Privacy Crawl (PPCrawl) corpus, which was compiled by Amos et al. using a crawler that locates, downloads, and extracts historical privacy policies from the Internet Archive's Wayback Machine [5]. PPCrawl[2] is a repository of 1,071,488 English language privacy policies from 130, 604 different websites, organized by policy date and website Alexa rating. PPCrawl contains policies from 1997 to 2019, although the collection lacks a copy of any company's policy version for each year. Due to this, PPCrawl's number of policies varies between the years; the number of policies from each continent for each year is shown in Figure 1.

We first segregate the sentences of the 1, 071, 488 policies using the sentence tokenizer in NLTK and then use our XLNet classifier to categorize the sentences of each policy. Due to the data's magnitude, processing and categorizing the entire corpus took about 22 days. We performed a small verification exercise to observe the quality of the XLNet predictions on this unseen data. For this, we verified the correctness of the predictions for each sentence (total 1, 858 sentences) in the latest available policies of the top-10 Alexa-ranked websites in PPCrawl, as well as on 2, 000 randomly sampled sentences. We observed a macro-average precision and recall of 0.89 and 0.92 respectively in the former, and 0.93 and 0.95 respectively in the latter. Categories such as "Introductory/Generic" and "Data Retention" have F1-scores above 0.90 in both instances. Table B2 and B3 lists the detailed metrics from these exercises.

We treat unavailable policies for an organization in a given year as missing data and ignore those instances when computing statistics for the year.

3.2.2 Composition. We calculate the proportion of each category in each policy in order to study the categorical composition of privacy policies over time. Additionally, we calculated the length of each sentence and established the typical sentence length for each policy category. To better understand ignored categories and their effects on a policy's construction, we also examine the fraction of policies in a year with missing categories.

3.2.3 Semantic Change with WMD. Semantic change describes how a text's meaning changes over time. Privacy policies may introduce a semantic change in the event of a revision. We detect the semantic change and quantify the extent of the change using Word Mover's Distance (WMD). WMD measures the disparity between two text documents as the smallest distance in a vector space between embedded words in one document and embedded words in the other [37]. We embed a policy's text using Polisis's [29] privacy-specific word embedding rather than XLNet's contextualized word

[2]https://github.com/citp/privacy-policy-historical

embedding. Depending on the context of a word's appearance, contextualized word embedding could have a different vector for the same word. Since it is probable that in policy reform, a simple restructuring of language without any genuine change in meaning can occur, this will lead to a different embedding for the same information and create a positive WMD (indicating false semantic change). Thus, we employ a static word embedding specific to privacy policies, which will always have the same embedding for the same information.

In the analysis, we determine the semantic differences between the current policy of an organization and the next newer version of that organization's policy that is available in PPCrawl. These differences are computed between texts belonging to the same category to assess the changes at a categorical level.

3.2.4 Flesch Reading Ease. We rate each policy's readability for each category using the Flesch reading ease score. A Flesch reading ease score of more than 90 is considered "very easy" to read, while numbers below 30 imply "very confusing" texts (see Table B4 for intermediate levels). Sentence and word counts of a text are considered when computing this score. The Flesch reading ease score has been used in the past for readability assessment of entire privacy policies [23, 24, 34, 44]. However, it is possible that the wording used in some specific categories reduces the readability score of the entire document. In order to ascertain the readability trend for each category, we compute the score separately for each category.

3.2.5 Altieri's Spatial Entropy. A key component of information discourse is the categorical organization of data; better organization of categories in a policy leads to better information accessibility. Therefore, less information uncertainty exists when similar categorical information is collated in a policy, i.e., when it is organized categorically. Traditionally, information uncertainty is computed using Shannon's entropy. Shannon's entropy is formally defined as the expected value of an information function that measures the amount of information about each category. However, the spatial locations of information are also significant when studying the degree of uncertainty in category placement. Shannon's entropy cannot recognize the significance of space when information uncertainty must be evaluated over a spatial region. In order to assess the uncertainty in the way that categories are organized in a policy, we use Altieri's spatial entropy [3]. Spatial entropy measures the distribution of categories across a document. For example, if other categories are interleaved between "User Choice/Control" practices, then all user choice and control descriptions are scattered between other descriptions, and will ultimately increase the entropy of the category (indicating disorganized or unaligned practice descriptions). Altieri's spatial entropy combines residual entropy and mutual information. While residual entropy measures the amount of information in one variable after the effect of another variable is taken into account, mutual information measures the information that two variables share. It is outside the scope of this paper to go into details about how the two values are calculated, but interested readers can refer to the original work [3]. We used the sentence number to specify where a specific piece of categorical information is located in a document. We computed the metric for each category using the SpatialEntropy library[3].

3.2.6 Self-Attention Based Coherence. Coherence in information is an essential aspect of information discourse. Making broad connections between various textual components is necessary for reading. A key evaluating factor is the consistency of a policy's various components. Unrelated sections would be found in a poorly-written policy, whereas relevant sections with closely related terminology would be found in a well-written policy. For example, a sentence such as *"We collect location information from the user"*, followed by *"The location information is used to recommend useful services in the locality"* is less coherent then *"We collect location information from the user to recommend useful services in the locality"*. Coherence in used language thus determines the overall connection of information, as opposed to entropy which determines the organization of content in a policy.

We employ Li et al.'s self-attention-based entity coherence evaluation metric to track long-distance relationships between words and produce a coherence score [39]. A vector with values between 0 and 1 related to a word's location is used to express position encoding after first obtaining the word embedding from Stanford's free source 50-dimensional GloVe embedding [49]. The input matrix to self-attention is then formed using word embedding and position embedding to capture the associations between each pair of words throughout a policy. Finally, the connection between word pairs is created as a series of word vectors, and input into an LSTM (Long Short-Term Memory) neural network. A fully connected layer with a nonlinear activation function calculates the final coherence score.

We make use of the implementation of this technique in the LingFeat package [38]. However, instead of focusing on one category at a time, we compute the coherence score utilizing a complete policy. This is because only taking into account text from one category at a time leaves out content from other categories and misrepresents the coherence of a section of text.

3.2.7 Co-occurrence Matrix. Our analysis also examines how one category relies on another to convey privacy-focused information. We examine each paragraph (segment) of each PPCrawl policy to understand the relationships between various categories. Typically, privacy policy segments consist of one or more categories. The co-occurrence counts of categories are calculated using the category of each sentence in a paragraph. For instance, we increase the count for the "First-Party Collection/Use" and "Introductory/Generic" category pair if statements from both categories are in the same paragraph. This gives us a category co-occurrence matrix for each of the policies. The evolution of the relationship between categories is then examined using these matrices.

4 RESULTS

We computed the evaluation metrics discussed in Section 3.2 on the PPCrawl corpus, and present here some trends and observations based on those metrics[4]. We present more discussion about these observations in Section 5.

[3]https://github.com/Mr-Milk/SpatialEntropy
[4]The data from this project is made available at https://github.com/crisp-du/ppevo

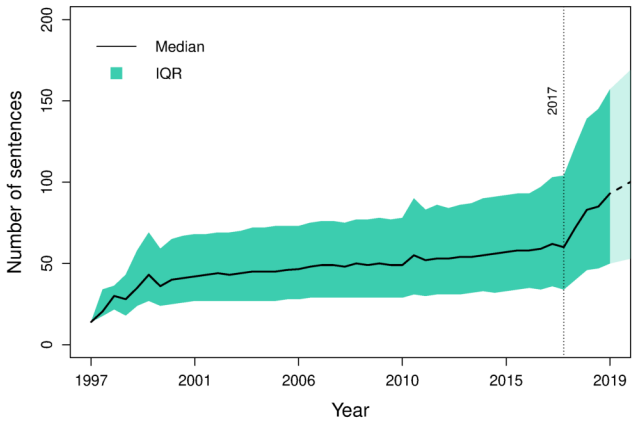

Figure 2: Number of sentences in a privacy policy over the years. IQR is interquartile range.

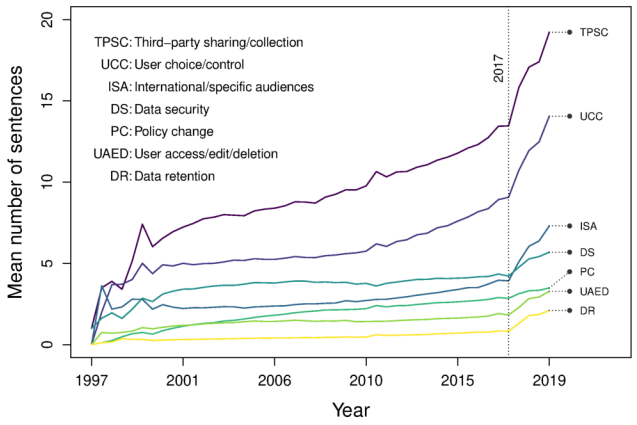

Figure 3: Categorical composition of a privacy policy over the years.

4.1 Structure of Privacy Policies

4.1.1 Policy Length. Privacy policies have become arduously long, making it challenging for readers to comprehend the provided information. The median and interquartile range of the number of sentences in policies for each year are shown in Figure 2. The number of sentences in policies has risen steadily over the years and has significantly increased since 2017, with more than 157 sentences for 25% of the policies in 2019. The increase in slopes follows the same trend for the 25^{th} and 75^{th} percentile values, showing that this tendency is not limited to specific instances but rather represents a general pattern. Adding new statements to an existing policy is more common than changing its current content when a policy is revised. Since revisions are reasonably regular, policies' lengths have gradually grown over time. When revising a policy, it is important to be careful not to leave previous practices behind as this may lead to more outdated, perplexing, or rhetorical content.

4.1.2 Categorical Composition. Our sentence-level categorization reveals that privacy policies are primarily composed of "First-Party Collection/Use" and "Introductory/Generic" information, with an average of 24% and 28% of a policy respectively (standard deviation of 3% and 8%) from 1997 to 2019. "Introductory/Generic" statements give consumers context to help them understand the supplied information. However, when overused, they lengthen the policy while making it more difficult for readers to find pertinent information. For example, an "Introductory/Generic" statement such as *"Through an open design, compelling editorial features, and analytics-based recommendations, for example, Myspace fosters a creative community of people who connect around mutual affinity and inspiration for shaping, sharing, and discovering what is next."* promotes the company's platform but provides no useful privacy-related information. We observe a decline in "Introductory/Generic" text from constituting 78% of a policy on an average in 1997 to approximately 22% in 2017, which is favorable towards having concise privacy practice descriptions. However, the relative proportion of such text remains significant enough to dilute relevant information.

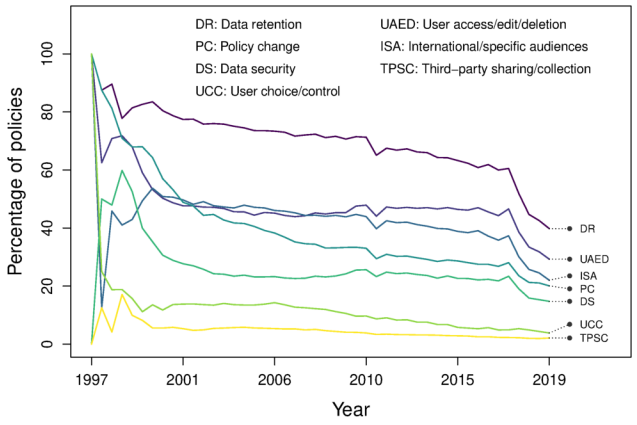

Figure 4: Percentage of policies with missing privacy categories over the years.

The average number of sentences per year that fall under different categories are shown in Figure 3. The plot illustrates a favorable trend by displaying an increase in crucial categories such as "Third-Party Sharing/Collection" and "User Choice/Control." However, with an average of two "Data Retention" sentences per policy in 2019, it remains the most under-addressed topic. The preservation of user information is a practice that should be communicated in a privacy policy; on the contrary, many websites seldom disclose their retention policy to users. The general shift in trend post-2017 may be attributable to the enforcement of the California Consumer Privacy Act (CCPA) and General Data Protection Regulation (GDPR) in 2018, as more sentences were included to address the privacy rights and consumer protection for residents of California, the United States and Europe.

Figure 4 depicts the percentage of policies in a year that does not cover a specific privacy category. "Data retention" is the category with the lowest representation; however, the percentage of websites that do not communicate their data retention practices (about 40%

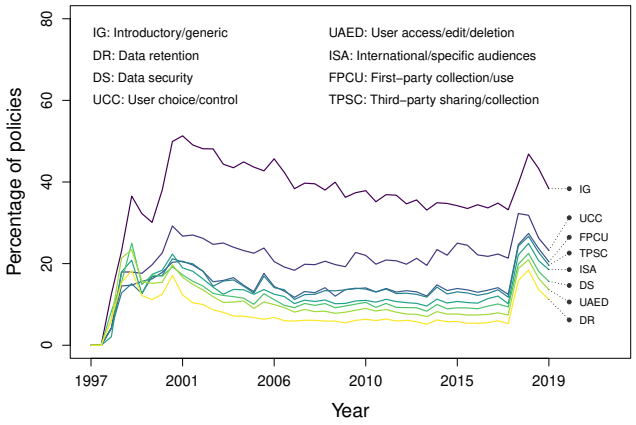

Figure 5: Percentage of policies that had a semantic change in a year.

in 2019) is steadily declining. The second most under-represented area, "User Access, Edit, and Deletion," further highlights the lack of control users have over their data once it has been collected. In over 50% of the policies from 2000 to 2017, this category went neglected and suffered from ongoing carelessness. With the implementation of the GDPR's right-to-access obligation, we see that after 2017, more than 60% of websites address this category, increasing to more than 70% in 2019. The plot also reveals that ≈97% of websites cover "User Choice/Control" practices in 2019, illustrating that most websites communicate some form of control choices.

4.1.3 Semantic Change. Semantic change refers to changes in the meaning of words in a sentence. In privacy policies, such changes can emerge during a policy revision. Any diff-based technique may be used to identify changes in sentence additions or deletions, as seen in [5]. To further understand the prevalence of semantic change, using the WMD metric, we examine the proportion of policies that alter their categorical content each year (Figure 5). Note that semantic change for a policy in a given year is computed with respect to a latest available prior version. Statistics on these semantic change values are computed by normalizing over the total number of available policies in the given year. In other words, we compute the proportion of policies in a year by dividing with a denominator value given by the number of available policies in the year. We also look at the distribution of this metric over the years (Figure 6).

The WMD values are relative and have no standard unit for reference; zero denotes no change, and relatively higher values denote more significant changes than earlier versions. Since PPCrawl does not always have a policy for a specific organization every year, to calculate the semantic change in a policy revision, we contrast the policy with the most recent version available before the said policy. Figure 5 shows that "User Choice/Control" is the category that experiences the second most frequent semantic change. In "User Choice/Control," semantic modifications are made to 26% of the policies on average; the most significant percentage was seen in 2018 when over 30% of the policies underwent semantic changes. Figure 6 illustrates the magnitude of these changes, which is relatively minimal except in 2018 and 2019. As a result, users interested in regulating access to their data should carefully revisit policies after revision since the opt-in or opt-out options (e.g., links) presented to them may change regularly with minor adjustments.

Figure 5 also shows a high correlation (0.99) between alterations in "First-Party Collection/Use" and "Third-Party Sharing/Collection" statements. However, in contrast to first-party practices, third-party practices see relatively more semantic shifts (Figure 6). The interdependence between the two activities raises the risk of confusing data collection and sharing (with third parties) due to the lack of differentiation and the possibility of ambiguity between the two types of practices.

The least commonly changed category is "Data Retention"; however, compared to other category modifications, the semantic shift in data retention has a larger magnitude (as indicated by the IQR for "Data Retention" in Figure 6). This suggests that businesses may abruptly adjust their retention policy in a significant way. It is worth noting that data retention statements such as *"Once it is no longer necessary for us to retain your personal information, we will dispose of it securely according to our data retention and deletion policies"* (eBay 2018 policy), may introduce ambiguity as to what factors determine that personal information is no longer necessary. The second least commonly changed category is "User Access, Edit, and Deletion," and the magnitude of the change is second only to "Data Retention." "Introductory/Generic" statements are changed frequently yet have the most negligible magnitude of change. This implies that a category will likely see significant changes if not frequently modified. An average of 11% of policies reporting changes to their data security methods did so with minor semantic alterations, except for adjustments brought on by the GDPR in 2018. Only 10% of the policies saw average updates to data security methods every year from 2002 to 2018. Statements relating to "International and Specific Audiences" have consistently seen changes over the years, despite regulations governing them being introduced infrequently. It indicates a continual effort across multiple organizations to correctly parse regulatory text and satisfy the stated requirements.

4.2 Comprehensibility of Privacy Policies

Elements such as readability, information coherence, and organization influence how approachable a privacy policy is to a user, whether it is easy to understand and access, and ultimately whether users can keep track of a policy's points in context with related information required to comprehend the described practices.

4.2.1 Readability. The readability of a privacy policy is among its most essential aspects. Figure 7 depicts the category-wise proportion of policies in a year that fall into different reading levels. For most policies, "First-Party Collection/Use" and "Third-Party Sharing/Collection" have a difficult readability level. However, since 2000, the percentage of confusing policies has steadily risen for both categories. "Data Security" practices have the highest percentage of confusing texts throughout the years, which can be due to the use of technical language in their descriptions. In contrast, the categories of "User Access, Edit, and Deletion" and "Do Not Track" indicate progress over the years, with a drop in the proportion

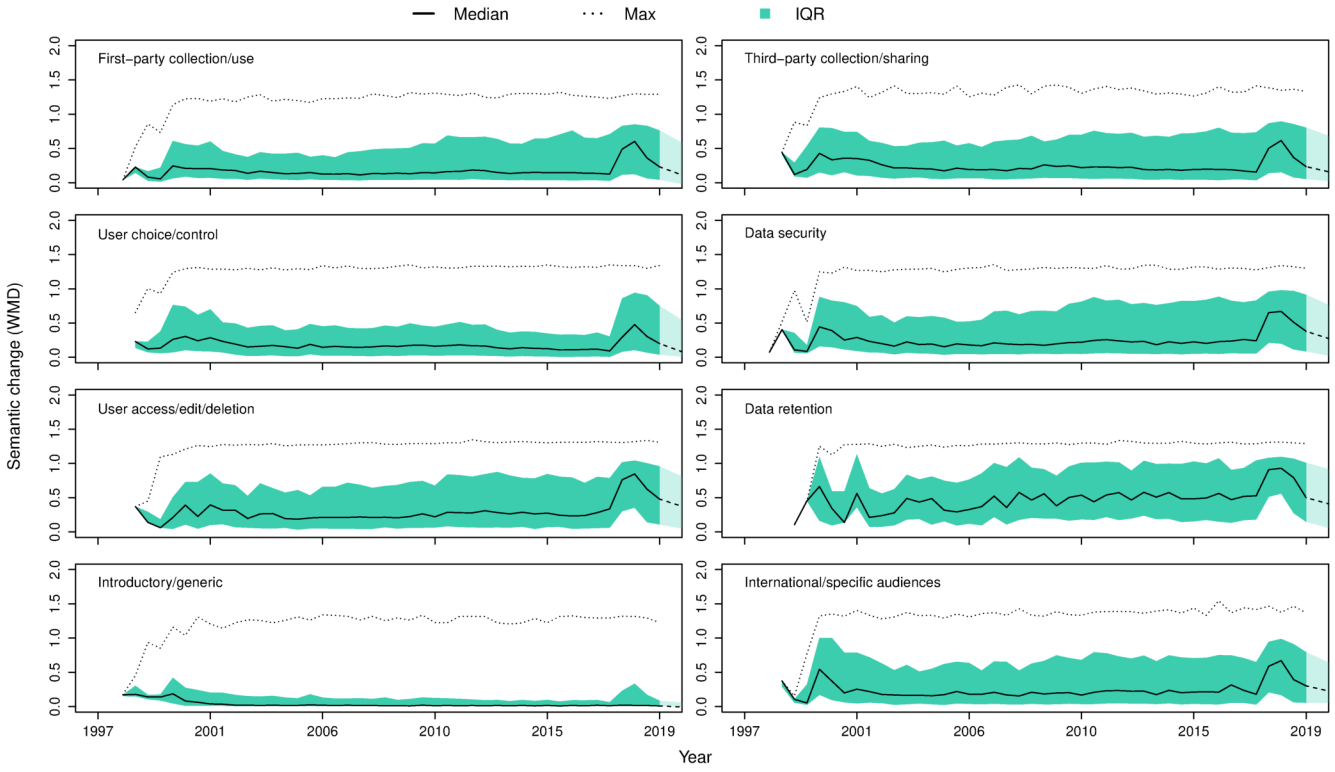

Figure 6: Magnitude of semantic change (measured with Word Mover's Distance) in privacy policies over the years across different categories. IQR is interquartile range.

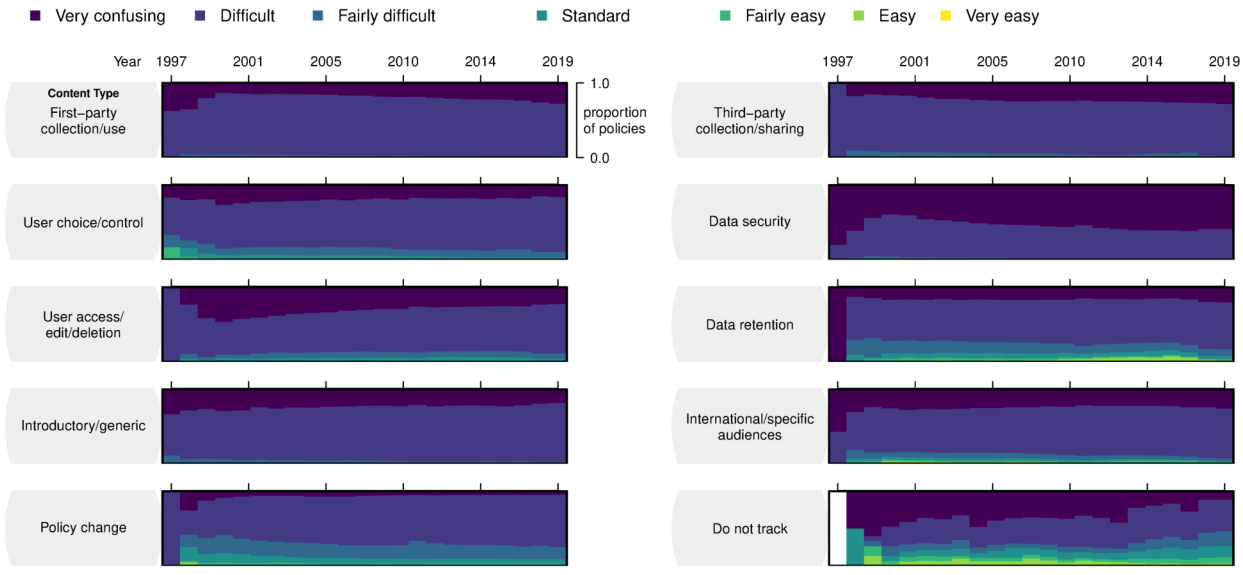

Figure 7: Category-wise distribution of readability (based on Flesch readability ease score) in privacy policies over the years.

of very confusing material and an increase in the proportion of standard, reasonably difficult, and fairly easy texts.

Given the nature of introductory and generic statements, it would be expected that they are most easy to read. Unfortunately, most policies continue to have introductory or generic statements that have a problematic readability level. Overall, the state of readability in privacy policies has remained (more or less) steady over the years and has room for significant improvements.

4.2.2 Coherence. Global connections between various textual elements are crucial for reading. Therefore, an essential evaluation element is the coherence of a policy's many components. A good document should have relevant sections with closely related terms. We employ self-attention to capture meaningful long-distance associations between words and estimate a coherence score for a policy [39]. There is no absolute reference for coherence, but relatively higher coherence values imply that the sentences read more smoothly. Figure 8 shows that each year, 90% of the policies have scores close to zero, suggesting that the text fails to coherently communicate the relationships between the many entities that make up a policy. The outliers make it evident that achieving higher coherence is possible, albeit only achieved by a minimal number of policies each year. For example, dofactory.com's policy (available from 2001 to 2004 in PPCrawl) has the highest coherence score (around 14). It is set up such that phrases from the same category are placed together and are not mixed in with descriptions of practices from other categories, with each sentence complete by itself. Netflix's policy from the same year (coherence score of zero) lacked organization, statements from many categories were mixed, and it relied on arbitrary descriptions placed in ambiguous text positions. The coherence between the material weakens substantially if the policy is not adequately aligned categorically.

4.2.3 Spatial Entropy. Altieri's spatial entropy calculates the degree of uncertainty in a category's location. A well-organized policy with instances of the same categories close together will have a low entropy, but an unorganized policy would experience more significant uncertainty. For example, the spatial entropy of the earlier mentioned 2003 Netflix policy is 2.24, while that of the dofactory.com policy is 1.

An illustration of the evolution in the uncertainty of information organization is depicted for each category in Figure 9. Over the years, entropy values have increased for "First-Party Collection/Use," "Third-Party Sharing/Collection," and "User Choice/Control." The IQR has also shifted towards higher values and has narrowed. This implies that, with time, these categories have become increasingly more disorganized. The 25^{th} percentile entropy values for "First-Party Collection/Use" statements are often higher than the 75^{th} percentile values for most categories in each year. The main goal of a privacy policy is to be clear regarding the reasons behind data collection and usage. However, given the relatively high entropy values, this information appears to be the most disorganized and the trend is worsening with time. Additionally, "Third-Party Sharing/Collection" and "User Choice/Control" were previously easily accessible in some policies (a near zero entropy for at least 25% of the policies) but have become just as difficult to access as first-party collection and usage descriptions.

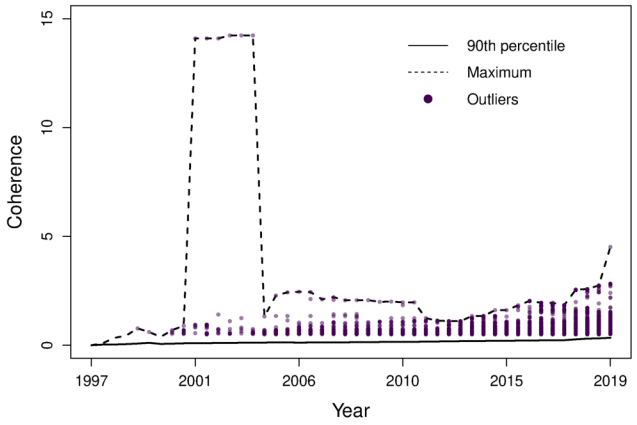

Figure 8: 90^{th}-percentile and maximum of coherence scores each year. Outliers shown are scores greater than $\max(0.5, 3^{rd}\text{−quantile} + IQR \times 1.5)$.

Until 2018, the median entropy levels for the categories"Data Security" and "International and Specific Audience" were consistently low. Both categories, after that, went through relative disorganization. "User Access, Edit, and Deletion" and "Data Retention" had even lower entropy values up to 2018, with the IQR lying near zero. Following 2018, the level of uncertainty increased for both categories. The implementation of GDPR in 2018 may have increased openness in data relevant to these categories, but the disorganized nature of the information's arrangement made it less accessible. Information on "Policy Change" and "Do Not Track" is still, comparatively, the most structured. Given that these categories contain few sentences and most websites do not mention them, they are expected to occupy a specific location in a policy.

4.2.4 Co-occurrence of Categories. In order to convey information, policy categories frequently co-occur with other categories in a policy. Figure 10 shows the evolution of categorical co-occurrences as a heat map among categories with passing years. Each matrix belongs to a particular category and indicates the degree to which the given category appears with the other categories in a given year. The degree represents the proportion of available policy instances in a year of the said category (given under a heatmap) that appear alongside another category in the same paragraph.

We observe that "Introductory/Generic," "First-Party Collection/Use," and "Third-Party Sharing/Collection" statements frequently co-occur with "User Access, Edit, and Deletion," "User Choice/Control," and "Data Retention" throughout. Even if the frequency of these occurrences has diminished over time, their existence is still noteworthy. The co-occurrence of user choice and control statements with first-party practice descriptions and third-party-specific information is significant, which is reasonable. Nevertheless, dependency on introductory or generic statements indicated incompleteness of control descriptions and required leveraging generic statements to convey complete information.

Choice and control information seems to be included in "First-Party Collection/Use" statements more often than "Third-Party

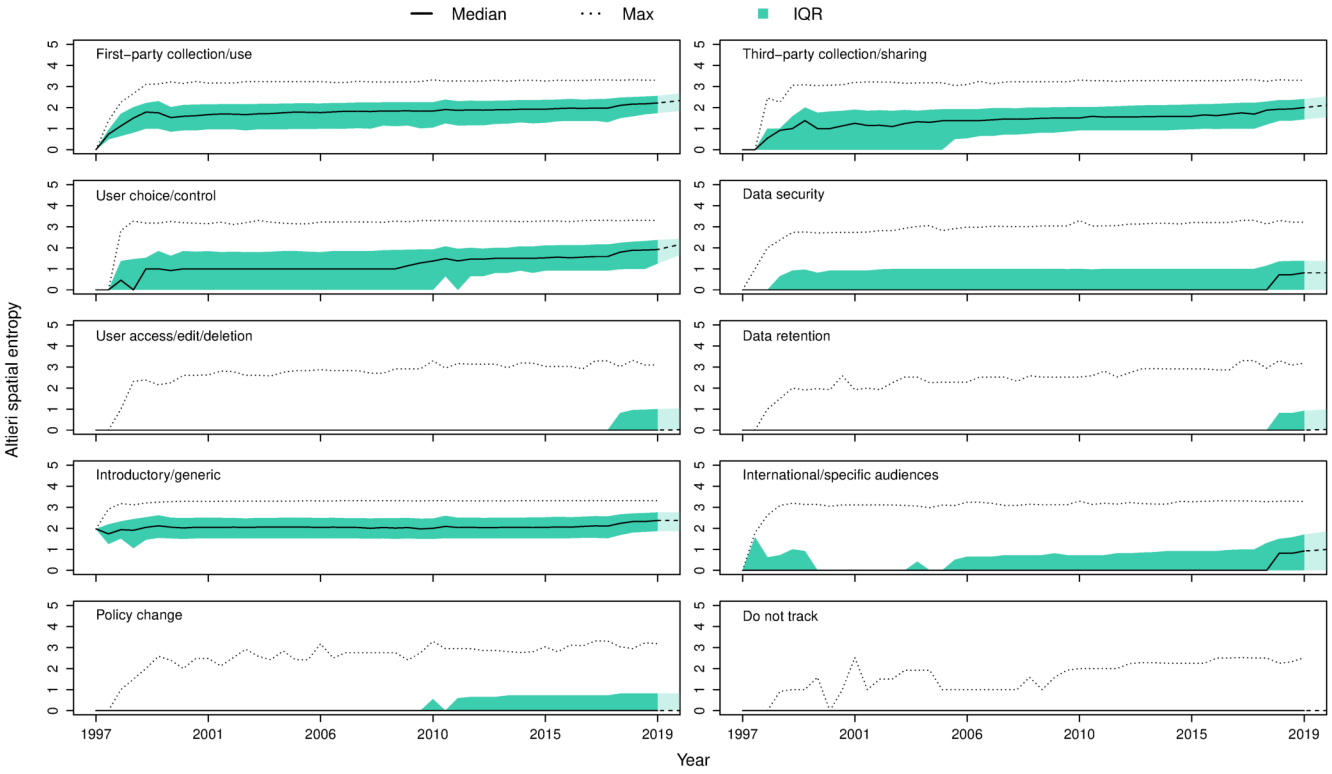

Figure 9: Altieri spatial entropy in privacy policies over the years across different categories. IQR is interquartile range.

Sharing/Collection" statements. This, along with the fact that "Third-Party Sharing/Collection" statements frequently appear with "First-Party Collection/Use" statements, makes it seem like first-party and third-party opt-in or out choices are not distinguishable. It would be expected that the co-occurrence of "User Choice/Control" are almost comparable with both first- and third-party practices if the choice and control for both practices are equally represented. Lack of control options for third-party practices may also be a factor. We note that dependency of choice and control statements on "First-Party Collection/Use" have decreased by ≈5% in recent years.

About 15% of the time, "User Access, Edit, and Deletion" statements co-occur with "First-Party Collection/Use" and "Introductory/Generic" statements. The number has gradually dropped over the years. However, there has recently been a minor increase in the dependence of "User Access, Edit, and Deletion" statements on "Introductory/Generic" statements. Co-occurrence with "User Choice/Control" has also grown in recent years. "User Choice/Control" defines control over data collection, and "User Access, Edit, and Deletion" describes control over collected data. High co-occurrence of the two categories can lead to confusion on which aspect (collection vs. collected) of the data is referred to.

The co-occurrence with "First-Party Collection/Use" of "Data Retention" descriptions has decreased over time. However, the fact that retention descriptions often occur with almost every other category in a policy suggests that retention regulations are usually described while detailing another practice.

Overall, it is worth mentioning that each category, to some extent, co-occurs with "First-Party Collection/Use," "Third-Party Sharing/Collection," and "Introductory/Generic" statements. Other categories have fewer statements in general. Coupling them with 'First-Party Collection/Use," "Third-Party Sharing/Collection," and "Introductory/Generic" statements in a single paragraph will make the information more difficult to find. Even ambiguous policy statements, marked by the "Practice Not Covered" category, appear most frequently with these three categories. This indicates that ambiguous practices frequently plague "First-Party Collection/Use" and "Third-Party Sharing/Collection" statements.

5 DISCUSSION AND IMPLICATIONS

Our results showed how policy snapshots changed over time as policies added more categorical information and also became longer. We also discovered how each category's readability, coherence, and organization have changed over time and how the relationships between various categories have evolved. In this section, we go over the implications of the observed results and expand on potential strategies that could help privacy policies to be presented better.

5.1 Streamline Revision

Revising policies is a continuous process. As a continual process, policy revisions have the ability to enhance a policy's quality over time by addressing some of the current issues during the revision.

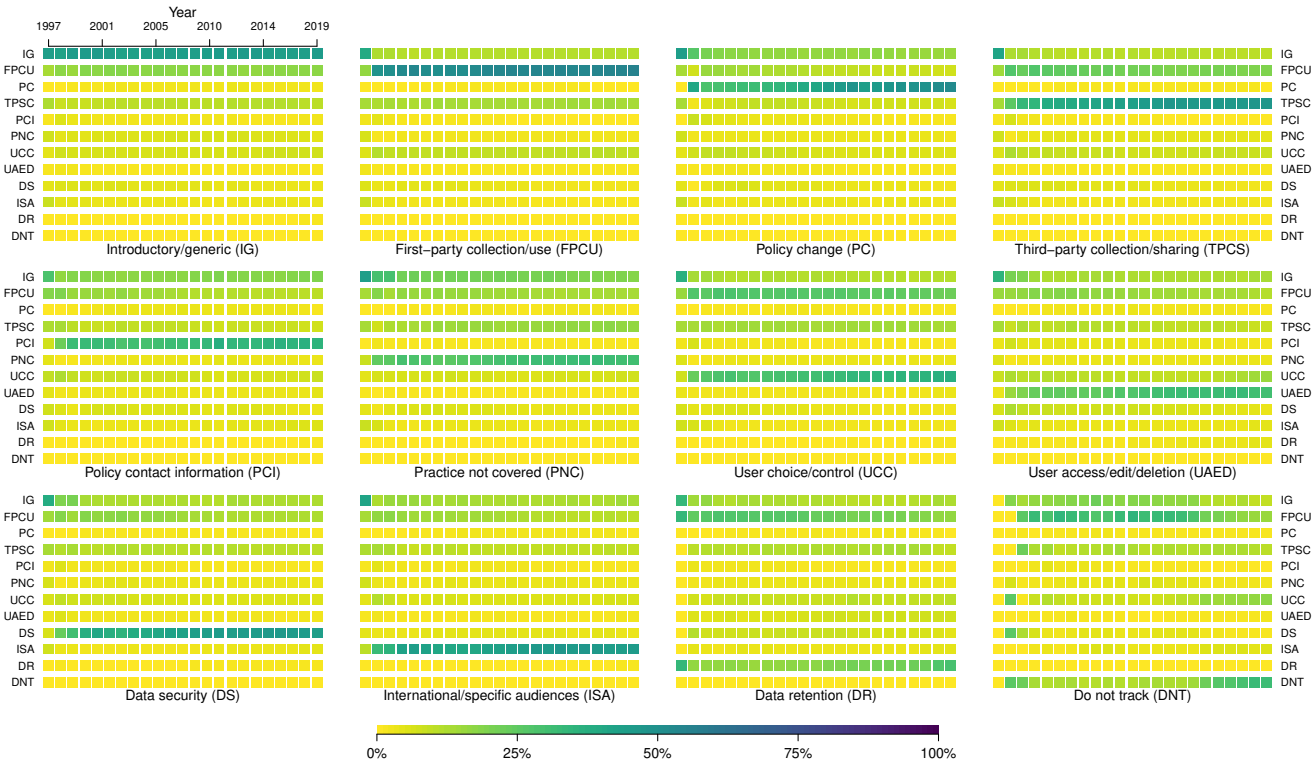

Figure 10: Percentage of co-occurrence of sentences among categories in privacy policies over the years. Each plot shows the co-occurrence of all categories with one single category over time.

5.1.1 Controlled Modification of Policy Text. We discovered that whenever a service provider expands its data practices, a corresponding description is also included in the policy. As features are developed, data practices change, sometimes leading to the abolition of older features and the accompanying data practices [55]. Therefore, it is preferable to modify the current policy's language as little as possible to include the new practices and eliminate the outdated ones. Our investigation demonstrates that adding a new practice to a revision is indeed doable by only changing an existing sentence. For example, consider the two sentences from Yahoo's 2002 privacy policy: *"For some financial products and services, we may also ask for your address, Social Security number, and information about your assets."* and *"We collect information about your transactions with us and with some of our business partners, including information about your use of financial products and services that we offer."*, which could have been modified into a single sentence such as, *"We may also gather your address, Social Security number, and details about your assets with us and some of our business partners for certain financial products and services, along with transaction information and service usage."*. This might have averted the introduction of a new sentence in the policy revision and improved readability by removing the need for readers to piece together information from two statements in different places. We see two benefits of being disciplined in how a policy is modified. Firstly, the policy length will increase by a minimum, limiting the length of time it takes to read the policy. Secondly, altering the current material will help remove outdated policy practices, which could not be eliminated by merely introducing new descriptions. As a result, outdated information is not accidentally shared.

5.1.2 Coherent Information. While some policies show high levels of content coherence, the vast majority have deficient levels of coherence among the many statements written to support practices. Therefore, it is essential to rewrite policies to integrate relevant materials to create coherent explanations of practices. For instance, the two sentences from Facebook's 2017 policy, *"We collect information from or about the computers, phones, or other devices where you install or access our Services, depending on the permissions you have granted"* and *"We may associate the information we collect from your different devices, which helps us provide consistent Services across your devices"*, are combined into a single sentence in Facebook's 2019 policy, *"We collect information from and about the computers, phones, connected TVs and other web-connected devices you use that integrate with our Products, and we combine this information across different devices you use."*. The material was made more coherent by simply integrating the two linked statements.

5.2 Categorical Policy Issues

Examining the whole policy is a common way to determine if privacy policies are usable. However, our research findings indicate

that each privacy category has a unique set of issues, resulting in poor notice and decision-making when combined.

5.2.1 User Control and Choice Consistency. Our findings demonstrate that the most commonly modified policy feature is the user's control and choices. It is reasonable for such a change to happen in terms of altered first- or third-party behavior. However, we see regular, modest changes in "User Control/Choice," which are separate from "First-Party Collection/Use" and "Third-Party Sharing/Collection." This suggests that while choice and control descriptions depend on the two categories, they do not dictate a change in the user's choice and control.

The need for more independence between control and choice from the first- and third-party practices suggests that these are not regarded as integral parts of the end-to-end process and are treated as secondary goals, resulting in policy revisions. The implementation of privacy compliance should be open, transparent, and planned across all employed processes [33]. Control choice revisions may be minimized by "User Choice/Control" policies that are more consistently implemented and founded on approved data collection, usage, and sharing methods. One method for maintaining such consistency is to have a fixed link to a separate page that lists user choice, control, and policy, instead of frequently changing opt-in/out weblinks in the policy text [34]. In addition, having a static page that lists all the control links will avoid the need to consult a policy to find the most recent link.

5.2.2 Minimizing Introductory and Generic Statements. Generic and introductory statements are frequently mixed with other categories, which affect the categorical organization of the policies. Additionally, the coherence among related descriptions can suffer if generic statements are inserted between related statements. Generic statements are meant to make the policies easier to use, but they may instead obfuscate the vital information provided by other categories. According to our analysis, "Introductory/Generic" statements change the most frequently, with minor changes in magnitude but rising entropy values over time. Therefore, it is necessary to reduce the use of introductory and generic statements to improve privacy policies. These claims only exist to facilitate the effective communication of other categories. However, if the other category statements are made to stand alone as complete statements, the need for "Introductory/Generic" statements can be reduced to a minimum.

5.2.3 Dissociation of Categories. Our analysis demonstrates that privacy categories frequently have strong relationships with one another and frequently depend on one another to describe practice information. At the same time, this method of articulation aims to set a particular category in a context. However, this method risks adding a category's problems to the description of a different category. For instance, adding a long first-party practice description as a context for the choice or control may make the already challenging "User Choice/Control" statements even more difficult to find. Therefore, as a whole, accessibility becomes increasingly tricky.

Furthermore, combining other categories also introduces ambiguity in the description, complicating policy usability. For example, the categories in a pair, such as "User Choice/Control" and "User Access, Edit, and Deletion," or "First-Party Collection/Use" and "Third-Party Sharing/Collection," represent distinct concepts yet ambiguous in the description due to high correlation. Furthermore, it can be challenging to distinguish between two concepts when descriptions of the two concepts in a policy are highly co-occurring.

5.3 GDPR Impact

The percentage of policies missing information specific to a given category significantly decreased across all categories in the wake of the GDPR implementation in 2018. Nevertheless, even after 2018, the categories of "User Access, Edit, and Deletion" and "Data Retention," which directly align with the GDPR's "right of access," "right to rectification," "right to erasure," and "right to be informed about the retained data policy," continue to be the most neglected among all the categories. This implies that while GDPR has improved several policies, a significant number of policies still require attention to disclose information fully.

Despite GDPR's positive impact on the openness of privacy policy information disclosure, information organization for each category unfortunately decreased. Both "Data Retention" and "User Access, Edit, and Deletion" categories observed a rise in entropy after 2019, a sign of increased disorder. Prior to this, the categories had a definite placement in a policy, despite the practice descriptions having less transparency. The readability of the categories also observed a decline.

"Data Retention" and "User, Access, Edit, and Deletion" also co-occur alongside "User Choice Control/Choice," "Data Security," "First-Party Collection/Use," "Third-Party Sharing/Collection," and "Introductory/Generic" sentences. This indicates that "Data Retention" and "User, Access, Edit, and Deletion" lack a specific role concerning privacy practices and are often described in a non-standardized context that introduces ambiguity in a policy. "User Choice/Control" sentences also observed a similar trend. People desire fine-grained control when disclosing their information [25]. Although the number of policies without "User Choice/Control," which was already relatively low before GDPR, did not change significantly due to GDRP, the organization and readability of these choice descriptions were adversely affected. In addition to being the most frequently changed category, post-GDPR website policies made these descriptions even more challenging to comprehend.

The structure of each policy category has generally declined with the implementation of GDPR, making policy communications more disorganized. In addition, while the length of privacy policies rose dramatically as a result of GDPR, the readability and consistency of the material have remained the same. Consequently, the sole beneficial effect of GDPR was to offer users more information; nonetheless, the user may still need help finding this information.

6 LIMITATIONS AND FUTURE WORK

We presented results from our analysis of privacy policies from 1997 to 2019 in the PPCrawl corpus, spanning over 20 years. However, the lack of policies post-2019 is a limitation of this work. The availability of post-2019 policies will provide a better overview of how organizations continue to address regulations such as GDPR and whether efforts are underway to make privacy policies more

approachable. Additionally, in practice, policies might vary considerably depending on the nature of the business. For instance, privacy policies communicating practices of a social media organization are articulated differently than privacy policies referring to banking or financial domains. We have not considered domain-specific analysis for this work. A business domain-specific selective examination of policies may also highlight characteristics that set different firms apart and reveal the problems and inclinations they are likely to face. From a method's perspective, correctly identifying policy statements about low-frequency categories is challenging. While deep-learning methods such as XLNet have demonstrated potential in identifying most categories, alternative approaches such as ensemble modeling coupled with cost-sensitive learning may be required to tackle issues with low availability of examples in specific categories.

7 CONCLUSION

Privacy policies are the primary means of distributing information on privacy practices and notifications to consumers. This study provides the results of a large-scale, longitudinal, category-based analysis of privacy policies spanning more than 20 years. We provide a holistic overview of the problems with privacy policies at a categorical level and track their evolution using a sentence-level classifier (XLNet). The implemented classifier aided in analyzing the composition, semantics, and structure of privacy policies over time. While specific categories see more frequent changes than others, we saw an overall rise in the informational completeness of privacy policies, positively reinforcing the transparency of these policies. However, the frequent changes make it challenging to trace the modifications implemented over time.

Additionally, we found that each category has its own unique set of readability and structural issues, and these category-specific problems are enhanced further with inter-category dependencies. Finally, it is concerning to note that, even though the problems in these categories are getting worse, a policy's textual content has a continuously low degree of coherence, with little to no evidence of an effort to improve comprehensibility. We offer some suggestions to improve the state of each category, such as the dissociation of categories and minimization of generic sentences, intending to keep privacy policies more approachable. Additionally, this study's findings can help develop better policies by adopting category-specific articulation practices and adhering to practices that do not incrementally make policies more challenging as they undergo various revisions.

ACKNOWLEDGMENTS

We thank Hamza Harkous and the rest of the Polisis team for inspiring this project, and making available the corpus used to train the privacy-specific word embedding models. We thank the authors of the OPP-115 and PPCrawl studies, for making the data sets accessible. We also thank the anonymous reviewers and the shepherd for their helpful comments.

This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

REFERENCES

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

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

[3] Linda Altieri, Daniela Cocchi, and Giulia Roli. 2018. A new approach to spatial entropy measures. *Environmental and Ecological Statistics* 25, 1 (2018), 95–110.

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

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

[6] 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 *Proceedings of the 28th USENIX Security Symposium*. USENIX Association, Santa Clara, United States, 585–602.

[7] Article 29 Working Party. 2014. *Opinion 8/2014 on the recent developments on the Internet of Things*. Technical Report 14/EN WP 223. European Data Protection Board.

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

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

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

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

[12] Duc Bui, Kang G Shin, Jong-Min Choi, and Junbum Shin. 2021. Automated extraction and presentation of data practices in privacy policies. *Proceedings on Privacy Enhancing Technologies* 2021, 2 (2021), 88–110.

[13] Jean Camp and Carlos Osorio. 2002. *Privacy-enhancing technologies for Internet commerce*. Technical Report. Harvard University, John F. Kennedy School of Government.

[14] Center for Information Policy Leadership. 2007. Ten steps to develop a multilayered privacy notice. https://www.informationpolicycentre.com/uploads/5/7/1/0/57104281/ten_steps_to_develop_a_multilayered_privacy_notice__white_paper_march_2007_.pdf. , 16 pages. accessed 2023-03-01.

[15] Shi-Cho Cha, Tzu-Yang Hsu, Yang Xiang, and Kuo-Hui Yeh. 2018. Privacy enhancing technologies in the Internet of Things: Perspectives and challenges. *IEEE Internet of Things Journal* 6, 2 (2018), 2159–2187.

[16] Federal Trade Commission et al. 2012. Protecting consumer privacy in an era of rapid change: Recommendations for businesses and policymakers. FTC Report. , 112 pages.

[17] Lorrie Cranor. 2002. A P3P preference exchange language 1.0 (APPEL1.0). https://www.w3.org/TR/P3P-preferences/. accessed: 2023-03-01.

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

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

[20] CyLab Usable Privacy and Security Laboratory. 2019. Privacy Bird. http://www.privacybird.org/. accessed: 2023-03-01.

[21] Giuseppe D'Acquisto, Josep Domingo-Ferrer, Panayiotis Kikiras, Vicenç Torra, Yves-Alexandre de Montjoye, and Athena Bourka. 2015. Privacy by design in big data: An overview of privacy enhancing technologies in the era of big data analytics. *arXiv preprint arXiv:1512.06000* 1.0 (2015), 1–80.

[22] 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*. Association for Computational Linguistics, Minneapolis, Minnesota, 4171–4186.

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

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

[25] Carlos Bermejo Fernandez. 2021. *Privacy and privacy enhancing technologies for post-GDPR ubiquitous computing*. Ph. D. Dissertation. Hong Kong University of Science and Technology (Hong Kong).

[26] Joshua Gluck, Florian Schaub, Amy Friedman, Hana Habib, Norman 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 *Proceedings of the 12th Symposium on Usable Privacy and Security*. USENIX Association, Denver, United States, 321–340.

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

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

[29] 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 *Proceedings of the 27th USENIX Conference on Security Symposium*. USENIX Association, Baltimore, United States, 531–548.

[30] Henry Hosseini, Martin Degeling, Christine Utz, and Thomas Hupperich. 2021. Unifying privacy policy detection. *Proceedings on Privacy Enhancing Technologies* 4 (2021), 480–499.

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

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

[33] ISO/IEC. 2011. *Information technology—Security techniques—Privacy framework*. International standard ISO/IEC 29100:2011(E). International Organization for Standardization, Geneva, Switzerland.

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

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

[36] Vinayshekhar Bannihatti Kumar, Abhilasha Ravichander, Peter Story, and Norman Sadeh. 2019. Quantifying the effect of in-domain distributed word representations: A study of privacy policies. In *Proceedings of the AAAI Spring Symposium on Privacy-Enhancing Artificial Intelligence and Language Technologies*. AAAI Digital Library, Palo Alto, 7.

[37] Matt Kusner, Yu Sun, Nicholas Kolkin, and Kilian Weinberger. 2015. From word embeddings to document distances. In *Proceedings of the 32nd International Conference on Machine Learning*. Proceedings of Machine Learning Research, Lille France, 957–966.

[38] Bruce W. Lee, Yoo Sung Jang, and Jason Lee. 2021. Pushing on text readability assessment: A transformer meets handcrafted linguistic features. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Punta Cana, Dominican Republic, 10669–10686.

[39] Xia Li, Minping Chen, Jianyun Nie, Zhenxing Liu, Ziheng Feng, and Yingdan Cai. 2018. Coherence-based automated essay scoring using self-attention. In *Proceeding of the 2017 Chinese Computational Linguistics and Natural Language Processing Based on Naturally Annotated Big Data*. Springer, Nanjing, China, 386–397.

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

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

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

[43] Aleecia M McDonald and Lorrie Faith Cranor. 2008. The cost of reading privacy policies. *A Journal of Law and Policy for the Information Society* 4, 3 (2008), 543.

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

[45] George R Milne, Mary J Culnan, and Henry Greene. 2006. A longitudinal assessment of online privacy notice readability. *Journal of Public Policy and Marketing* 25, 2 (2006), 238–249.

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

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

[48] National Telecommunications and Information Administration. 2013. Short form notice code of conduct to promote transparency in mobile apps practices. https://www.ntia.doc.gov/files/ntia/publications/july_25_code_draft.pdf. accessed: 2023-03-01.

[49] Jeffrey Pennington, Richard Socher, and Christopher D Manning. 2014. GloVe: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Oha, Qatar, 1532–1543.

[50] David J Phillips. 2004. Privacy policy and PETs: The influence of policy regimes on the development and social implications of privacy enhancing technologies. *New Media & Society* 6, 6 (2004), 691–706.

[51] Travis Pinnick. 2011. Privacy short notice design. TRUSTe blog.

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

[53] Joel R Reidenberg, Jaspreet Bhatia, Travis Breaux, and Thomas B Norton. 2016. Automated comparisons of ambiguity in privacy policies and the impact of regulation. http://papers.ssrn.com/sol3/papers.cfm.

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

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

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

[57] Kanthashree Mysore Sathyendra, Florian Schaub, Shomir Wilson, and Norman Sadeh. 2016. Automatic extraction of opt-out choices from privacy policies. In *Proceedings of the 2016 Advancement of Artificial Intelligence Fall Symposium Series*. AI Access Foundation, Arlington, United States, 270–275.

[58] Florian Schaub, Rebecca Balebako, Adam L Durity, and Lorrie Faith Cranor. 2015. A design space for effective privacy notices. In *Proceedings of the 11th USENIX Conference on Usable Privacy and Security*. USENIX Association, Ottawa, Canada, 1–17.

[59] Paulo Silva, Carolina Gonçalves, Carolina Godinho, Nuno Antunes, and Marilia Curado. 2020. Using NLP and machine learning to detect data privacy violations. In *Proceedings of the 2020 IEEE Conference on Computer Communications Workshops*. IEEE, Toronto, Canada, 972–977.

[60] Daniel Smullen, Yaxing Yao, and N Sadeh. 2021. Managing intrusive practices in the browser: A user centered perspective. *Proceedings on Privacy Enhancing Technologies (PoPETs)* 2022, 4 (2021), 500–527.

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

[62] V Thong Ta and M Hashem Eiza. 2021. DataProVe: Fully automated conformance verification between data protection policies and system architectures. *Proceedings on Privacy Enhancing Technologies* 2022, 1 (2021), 565–585.

[63] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Proceedings of the 31st Conference on Neural Information Processing System*. Curran Associates, California, United States, 5998–6008.

[64] Paul Voigt and Axel Von dem Bussche. 2017. *The EU general data protection regulation (GDPR): A practical guide*. Springer International Publishing, Springer International Publishing.

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

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

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

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

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

A DATA PRACTICE CATEGORIES

The 12 data practice categories used in the classification have the following generic meaning [66].

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

B ADDITIONAL TABLES

Table B1: Frequency of sentences in each category in the annotated OPP-115 data set.

Category	Frequency
Policy Change	373
Third Party Sharing/Collection	2328
User Choice/Control	1267
Data Retention	107
User Access Edit and Deletion	307
Practice not covered	502
International and Specific Audiences	793
Data Security	589
Privacy contact information	308
Do Not Track	59
Introductory/Generic	1176
First Party Collection/Use	2908

Table B2: XLNet prediction performance on 1,858 sentences from the latest available policies of top-10 Alexa-ranked websites in PPCrawl. Pr: Precision, Re: Recall, F1: F1-score.

Category	Pr	Re	F1
Policy Change	0.95	1.00	0.97
Third-Party Sharing/Collection	0.94	0.92	0.93
User Choice/Control	0.85	0.93	0.89
Data Retention	0.96	0.86	0.91
User Access, Edit, and Deletion	0.94	0.92	0.93
Practice Not Covered	0.92	0.88	0.90
International and Specific Audiences	0.90	0.96	0.93
Data Security	0.84	0.96	0.89
Privacy Contact Information	0.71	0.96	0.82
Do Not Track	0.86	0.86	0.86
Introductory/Generic	0.95	0.85	0.90
First-Party Collection/Use	0.91	0.92	0.92
micro avg	**0.91**	**0.91**	**0.91**
macro avg	**0.89**	**0.92**	**0.90**

Table B3: XLNet prediction performance on 2,000 randomly sampled sentences in PPCrawl. Pr: Precision, Re: Recall, F1: F1-score.

Category	Pr	Re	F1
Policy Change	0.99	0.97	0.98
Third-Party Sharing/Collection	0.96	0.98	0.97
User Choice/Control	0.87	0.98	0.92
Data Retention	0.94	0.89	0.92
User Access, Edit, and Deletion	0.97	0.93	0.95
Practice Not Covered	0.95	0.87	0.91
International and Specific Audiences	0.92	1.00	0.96
Data Security	0.96	0.98	0.97
Privacy Contact Information	0.97	0.95	0.96
Do Not Track	0.71	1.00	0.83
Introductory/Generic	0.98	0.91	0.95
First-Party Collection/Use	0.97	0.98	0.97
micro avg	**0.96**	**0.96**	**0.96**
macro avg	**0.93**	**0.95**	**0.94**

Table B4: Flesch reading ease score interpretation.

Score range	Readability level	Grade
≥90	Very Easy	5th grade
80-89	Easy	6th grade
70-79	Fairly Easy	7th grade
60-69	Standard	8th and 9th grade
50-59	Fairly Difficult	10th to 12th grade
30-49	Difficult	In college
≤29	Very Confusing	College graduate