Privacy Policy Analysis with Sentence Classification

Andrick Adhikari
Computer Science
University of Denver
Denver, CO, USA
andrick.adhikari@du.edu

Sanchari Das
Computer Science
University of Denver
Denver, CO, USA
sanchari.das@du.edu

Rinku Dewri
Computer Science
University of Denver
Denver, CO, USA
rinku.dewri@du.edu

***Abstract*—Privacy policies inform users of the data practices and access protocols employed by organizations and their digital counterparts. Research has shown that users often feel that these privacy policies are lengthy and complex to read and comprehend. However, it is critical for people to be aware of the data access practices employed by the organizations. Hence, much research has focused on automatically extracting privacy-specific artifacts from the policies, predominantly by using natural language classification tools. However, these classification tools are designed primarily for the classification of paragraphs or segments of the policies. In this paper, we report on our research where we identify the gap in classifying policies at a segment level, and provide an alternate definition of segment classification using sentence classification. To this aid, we train and evaluate sentence classifiers for privacy policies using BERT and XLNet. Our approach demonstrates improvements in prediction quality of existing models and hence, surpasses the current baselines for classification models, without requiring additional parameter and model tuning. Using our sentence classifiers, we also study topical structures in Alexa top 5000 website policies, in order to identify and quantify the diffusion of information pertaining to privacy-specific topics in a policy.**

***Index Terms*—Automated Classification, Privacy Policy, Usable Privacy and Security, NLP.**

I. Introduction

Privacy policies are natural language documents that ideally describe all the applicable data practices that an organization follows. The primary purpose of these documents is to notify and educate users about the data practices associated with the use of services offered by these organizations, while complying with data processing standards and regulations [1]. However, the unstructured, incomprehensible, and complicated characteristics of the language usually adopted for these documents pose a significant obstacle for users [2]. Nonetheless, various legal regimes around the world require service providers to notify users about their privacy practices [3]. This leads to many privacy policies that most users are unlikely to read. It is estimated that it would cost users at least 181 hours per year if they decide to read all the policies that apply to them [4]. Researchers have thus adopted natural language processing (NLP), machine learning (ML), and deep learning to develop tools to extract critical, actionable information from privacy policies [5]–[7].

In applying NLP and ML in privacy policy research, automated classification has received recent attention. Automated classification of privacy policies usually involves assigning each segment (paragraph) a category label [8]–[10]. Such categorization provides a high-level overview of the segment and enables users with an option for selective reading of the document, increasing comprehension and readability. Subsequent extraction of information specific to a segment's category can also be attempted, allowing a further breakdown of the information contained in a policy.

Automated classification applications in the domain have delivered several research projects that experiment with different learning models. Most of the previous research focuses on segment-level classification or uses sentence classification only to extract specific types of information, e.g. opt in/out choices. However, a casual look at online privacy policy documents immediately reveals that policy writers do not restrict themselves to assumed notions of categories within a paragraph. For example, a segment discussing first-party data collection practices can include information about third-party sharing, user opt in/out choices, or data security mechanisms, all in the same paragraph. Consequently, machine learning models trained to identify a segment's purpose can miss relevant information, especially when the information is treated as noise by a model. Privacy policy analysis methods also follow a hierarchical structure in extracting information, i.e. once a segment's category is identified, a follow up analysis can be done to extract privacy-specific artifacts that are relevant in the context of the identified category [8]. Clearly, if a prediction misses on an appropriate segment-level category assignment, finer information pertaining to that category is also missed.

To address this research gap, the goal of this study is to shift the focus from segment classification to sentence classification. We redefine segment classification as an aggregation of sentence classification, and conduct our experiments with Polisis [8], BERT [10], and XLNet [9], three state of the art models in the privacy policy classification domain. We observe better or consistent results in predicting the type of information contained in a segment when utilizing sentence-level classification, irrespective of the model. We also conduct a structural analysis of privacy policies at segment and sentence levels using the OPP-115 corpus [7] and Alexa top 5000 website policies. Our analysis with sentence classification exemplifies and brings to light several flaws in the topic organization of privacy policies that are overlooked at paragraph or segment-level analysis. The objective behind our work is to motivate sentence-level as the ideal granularity for future privacy policy automation tools and analysis.

The remainder of the paper is organized as follows. Sections II and III present related work and background on privacy policy classification respectively. Section IV demonstrates the inadequacy of current segment classifiers in capturing the information contained in segments, versus one that derives segment labels from sentence labels. Section V discusses the performance improvements that can be obtained by training a classifier to label sentences, instead of segments. Section VI presents a structural analysis of privacy policies using our sentence classifier. Finally, we conclude in Section VII.

II. Related Work

Privacy policies' readability and usability have been a concern for a few years now. Ammar et al. first explored text categorization as a means to extract information from a privacy policy in a pilot study [6]. Their study focused on determining the feasibility of text classification in the privacy policy domain by checking the presence or absence of predetermined concepts in the entire policy document. Zimmeck et al. also performed document-level classification to determine the presence or absence of practices associated with data collection, encryption, retention, Ad-tracking, profiling, and Ad-disclosure [11].

More models came into the research domain as research moved into the automation of information extraction from the policies. Along those lines, Constante et al. tested several learning algorithms (k-NN, SVM, LSVM, and decision tree) to map paragraphs of a policy to labels derived from EU directives, OECD (Organization for Economic Cooperation and Development) guidelines, and standard policy practices [12].

Wilson et al. presented the OPP-115 corpus containing segment-level annotations and fine-grained annotations of 115 website policies [7]. The OPP-115 corpus does not contain sentence-level category labels; the finer annotations it has are attribute labels on specific phrases/words of a sentence. Most of the work post-OPP-115 corpus creation has utilized these segment-level annotations and hence, implemented segment-level classifiers. For example, Harkous et al. presented a hierarchy of high level and finer-grained classifiers for segments of policies [8]. The classifiers were combined into a framework named 'Polisis,' which users can also use to visualize the predicted categories in an interactive interface. As improvements to the previous models, Mousavi et al. and Mustapha et al. both presented segment-level classifiers utilizing BERT and XLNet models respectively, in pursuit of establishing stronger baselines for privacy policy segment classification [9], [10]. However, it is also pointed out that the segmentation method employed for privacy policies is not standardized, thus making results less reproducible.

As an alternative, sentence classification for privacy policies can be explored, which is rarely done previously. Liu et al. experimented with SVM, LR, and CNN models for both segment and sentence classification [13]. Due to the absence of sentence-level category labels, the sentence classifier in their work was created using attribute level annotations of OPP-115. It is worth noting that domain-specific word embedding data for privacy policies have become available since then, which can significantly enhance the performance of classification models. Other researchers have leveraged sentence classification for tasks such as classifying whether a sentence describes a user choice instance [14]. Sathyendra et al. extend choice classification by employing active learning to clean the OPP-115 data set for opt-out choices and upgrading to a two-classifier architecture [15]. Similarly, Kumar et al. present the development of a tool that automatically extracts and classifies opt-out choices found, and propose techniques to automatically identify opt-out choices for the design, development, and evaluation of a browser extension [16]. Zimmeck et al. focused on categorizing sentences of policies in an Android ecosystem, using classification and static code analysis of 1,035,853 applications [17], [18].

As we see from this prior work, analysis of privacy policies has relied heavily on segment-level classification. Sentence-level classification has been used when the goal is to extract specific types of information from a policy. This work can be considered as a first step towards demonstrating the inadequacy of using segment-level classification in privacy policy analysis, and subsequently the insights that can be gained by adopting sentence classification as the primary step instead.

III. Methods and Data

A. Data Used

1) OPP-115 corpus: Wilson et al. [7] presented OPP-115, a corpus of 115 website policies annotated with 12 high-level data practice categories. The policies are also annotated with 22 distinct attributes specific to each category. The annotation scheme and tools were developed through careful analysis of methods for labeling policy segments suitable for crowd workers to generate detailed policy annotations [19]. The high-level categories are 'First Party Collection/Use', 'Third-Party Sharing/Collection', 'User Choice/Control', 'User Access, Edit, & Deletion', 'Data Retention', 'Data Security', 'Policy Change', 'Do Not Track', 'International & Specific Audiences', and 'Other'. The 'Other' category encompasses three categories – 'Introductory/Generic,' 'Practice not covered,' and 'Privacy contact information.' We selected the OPP-115 corpus as the primary data set for this study as it is the most widely used data set in the privacy policy classification research.

2) Princeton Privacy Crawl: Amos et al. developed a crawler that discovers, downloads, and extracts archived privacy policies from the Internet Archive's Wayback Machine [20]. Using the crawler, a repository of 1,071,488 English language privacy policies was created, curating policies of 130,000 different websites organized by policy date and the website's Alexa ranking. This corpus addresses the issue that prior research has been limited to analyzing privacy policies from a single point in time or short periods, by providing access to a large-scale, longitudinal, curated data set. We use

Princeton's privacy crawl repository[1] to download the latest available policies of Alexa top 5000 websites.

B. Learning Models

Polisis is one of many projects using CNN-based classifiers to improve privacy policy interpretability [8]. Recent studies with BERT and XLNet based models have also shown outstanding performance in the domain [9], [10]. We have thus selected Polisis, BERT, and XLNet classifiers for our study, and this section lays down the background for Polisis, BERT, and XLNet models.

1) Polisis: Polisis uses Convolutional Neural Network (CNN) for all the classifiers, and integrates a policy specific embedding trained using Fasttext on 130,000 privacy policies crawled from the Google Play store [21]. Since the custom word embedding uses texts from privacy policies, Polisis trained models can better capture the language information specific to the privacy policy domain. For example, words like 'email,' 'location,' 'name,' etc. that usually represent 'collected information,' will not have vectors that represent this privacy policy domain-specific information if pre-trained generic word embeddings such as Word2vec [22] or GloVe [23] are used. Custom domain-specific word embedding thus leads to better text classification.

Fasttext learns representations for character n-grams (of size 3 to 6) and represent words as the sum of the n-gram vectors, using an extension of the continuous skip-gram model [22]. Fasttext can efficiently handle rare or unseen words as well by combining vectors of constituent subwords, which Word2vec and GloVe cannot due to their inability to handle unknown or out-of-vocabulary words. Fasttext uses a neural network for computing word embeddings, and require substantial computational resources and time. Fasttext captures semantic relationships better than traditional embedding methods, but real-time computation is not an option. Quality is also dependent on the corpora used for generating the embeddings as they are static and do not alter once learned.

CNN is used in the Polisis classifier, as CNN models are good at recognizing tokens that are good indicators for a class. The architecture adopted by Polisis is described in detail in [8]. The first layer is an embedding layer that outputs the vectors of each word in the input segment. This layer is frozen to prevent any alteration of the custom embedding generated using Fasttext. The convolutional layer then applies a Rectified Linear Unit (ReLU) function over a window of k words. Next, vectors from all the windows are combined with a max-pooling layer that passes through a fully connected layer (dense) with a ReLU activation function. After passing through a second fully connected layer, a sigmoid operation at the final layer generates the class probabilities. Polisis established a good baseline for our research which we implemented for our analysis (discussed later), but for training our sentence classifier we chose to explore newer models like BERT and XLNet, which outperforms CNN models in text classification.

2) BERT: Devlin et al. introduced BERT (Bidirectional Encoder Representations from Transformers). This model can be fine-tuned with just one additional output layer to create state-of-the-art models for a wide range of tasks, such as question/answering and language inference, without substantial task-specific architecture modifications [24]. There are two steps in building the model: pre-training and fine-tuning. The model is trained on unlabeled data over different pre-training tasks during pre-training. For fine-tuning, the BERT model is first initialized with the pre-trained parameters. Then, the parameters are fine-tuned using labeled data from the downstream tasks, such as multi-label classification in our case. BERT is trained to minimize a combined Masking Language Modelling (MLM) and Next Sentence Prediction (NSP) loss function.

3) XLNet: XLNet is a BERT-like model which uses Transformer-XL [25], except XLNet is an autoregressive (AR) language model, whereas BERT is categorized as an autoencoder (AE) language model. AE language models consider the context in both the forward and backward directions, and aims to reconstruct the original data given a corrupted input. BERT achieves that by replacing tokens with "mask" tokens during pre-training. A disadvantage with such an approach is that mask tokens are not present during fine-tuned training, leading to discrepancies. Masking also assumes that tokens are independent of each other, which is not always the case. For example, in the sentence *"We store personal information as long as needed."*, tokens 'personal' and 'information' are masked by BERT even though the masked tokens have an implicit relation. However, the AR model will aim to predict 'information' given unmasked tokens and predict 'personal' given unmasked tokens separately.

IV. Segment Classification With Sentence Labels

Privacy policy segments can have one or more categories, and categories are not mutually exclusive. Consider this segment from Amazon's privacy policy: *"Information You Give Us: We receive and store any information you enter on our Web site or give us in any other way. You can choose not to provide certain information, but then you might not be able to take advantage of many of our features. We use the information that you provide for such purposes as responding to your requests, customizing future shopping for you, improving our stores, and communicating with you."*, which classifies as 'First Party Collection/Use', 'Data Retention' and 'User Choice/Control' categories according to OPP-115 annotations. Automated classification of privacy policy segments is therefore a multi-label classification problem.

Polisis, like most multi-label classifiers utilize a threshold of 0.5, on probability. During prediction, for each category, Polisis outputs an independent probability indicating whether the category is an appropriate label for the input text. All categories that have a probability greater than the threshold are assigned as predictions to the input text. If no category has a probability greater than the threshold, no label is assigned

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

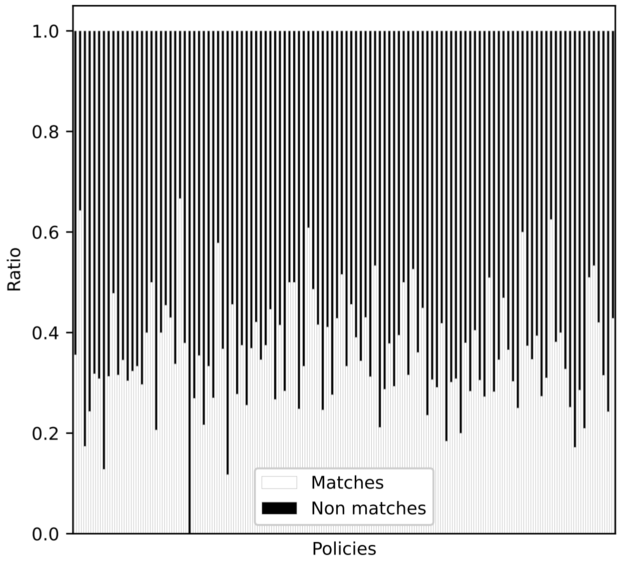

Fig. 1. Ratio of sentences (relative to total number of sentences) where sentence category matches or does not match a high probability category assigned to the corresponding segment.

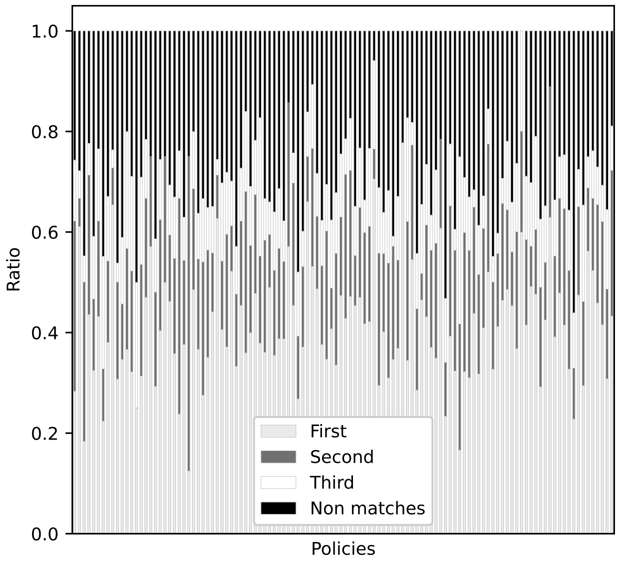

Fig. 2. Distribution of sentences where the category matches one of the top three predictions for the corresponding segment. Segment predictions are assumed to be done without the restriction of probability thresholds.

to the text. Please note that, unlike in single-label classification, probabilities in multi-label classification are computed independently for each category.

To help comprehend the loss of information when utilizing a segment-level classifier instead of a sentence-level classifier, we used Polisis to classify individual sentences from 115 policies of the OPP-115 corpus. Since the Polisis classifier was trained at a segment level, to have high confidence on our evaluation, we eliminated examples where the sentence category assignment probability was less than 0.9. We predicted categories for the segments in the same policies and evaluated what fraction of sentences in a policy have a predicted category that is included in the categories assigned to the corresponding segment where the sentence is present. The vertical axis in figure 1 shows the ratio of sentences matching some assigned segment category (with respect to the total number of sentences in a policy), and the ratio of sentences not matching any of their assigned segment categories. The horizontal axis represents each policy from OPP-115 corpus.

It is evident from the plot that more than 50% of sentences are not represented in the assigned segment category. In other words, the classifier potentially misses a relatively large portion of information at the segment level. When a segment-level classification loses most of the sentence-level categorical information, it also leads to degradation in classification performance at segment level. Policy writers often do not group same category sentences in a segment. Consider the aforementioned segment example from Amazon's privacy policy composed of sentences from different categories. Even when tokens or words exist that are strong indicators for a particular class (like 'choose' that strongly suggest the 'User Choice/Control' practice), when grouped together with indicator tokens of other classes, the effect gets dampened. Thus, the segment classifier could not assign high probabilities to any of the classes and hence, no category was assigned.

If we ignore the threshold used during prediction, and consider the three classes with highest probabilities instead, we observe a better representation of sentence categories at segment level. Figure 2 represents this phenomenon, where the y-axis represents the ratio of sentences in a policy (represented by x-axis) matching either of the top-three segment classes (marked as 'First', 'Second' and 'Third' in the plot). The plot also represents the ratio of sentence categories not matching segment categories, marked as 'Non matches'.

Comparing the plots in figure 1 and 2, the first observation is the significant reduction in unrepresented sentence categories. The portion of non-matches has decreased significantly, implying that segment classification can capture sentence categories but with low probability. The portion of 'First' in figure 2 has high resemblance with 'Matches' in figure 1, implying that classification with threshold only captures the first highest probability category for most segments. Removing the threshold clearly demonstrates that even the second and third highest probability categories of segments hold sentence categorical information. Fraction of sentences matching the second highest segment category is almost equal to first highest segment category matches. Even the third segment category matches a significant portion of the category of sentences. **Derived segment classification.** To generate further empirical evidence that segment classificiation is more accurate when it is derived from sentence labels, we took Polisis's trained high level classifier as it is, without any modifications, and used it to classify all the segments from the OPP-115 corpus. As an alternative method for obtaining predictions of the segments,

TABLE I
POLISIS: PERFORMANCE METRICS OF DIRECT VERSUS DERIVED SEGMENT CLASSIFICATION

Category	**Direct**			**Derived**		
	F-1	**Prec.**	**Rec.**	**F-1**	**Prec.**	**Rec.**
First Party Collection/Use	0.31	0.89	0.18	0.53	0.78	0.40
Third Party Sharing/Collection	0.54	0.73	0.43	0.56	0.73	0.46
User Access, Edit and Deletion	0.05	0.86	0.03	0.06	0.64	0.03
Data Retention	0.01	1.00	0.01	0.00	0.00	0.00
Data Security	0.26	0.77	0.16	0.32	0.67	0.21
International and Specific Audiences	0.69	0.91	0.55	0.71	0.88	0.60
Do Not Track	0.17	1.00	0.09	0.06	1.00	0.03
Policy Change	0.34	0.85	0.21	0.49	0.86	0.35
User Choice/Control	0.23	0.77	0.14	0.24	0.60	0.15
Introductory/Generic	0.36	0.62	0.25	0.41	0.37	0.45
Practice not covered	0.14	0.45	0.08	0.17	0.57	0.10
Privacy contact information	0.38	0.54	0.29	0.33	0.48	0.25
micro avg	0.36	0.73	0.24	0.43	0.62	0.33
macro avg	0.29	0.78	0.20	0.32	0.63	0.25

named as derived segment classification here, we first tokenize the segments into individual sentences, and then classified the sentences with the Polisis classifier. We then derived a multi-label prediction for each segment, comprising of the labels assigned to one or more sentences in a segment. Both set of segment predictions were then compared against the ground truth annotations in OPP-115.

For the traditional segment classification, we use a threshold of 0.5 during prediction (similar to as in [8]), and for our derived segment classification method, we use a threshold of 0.9. Precision and recall computations are also mutually exclusive for each category. For example, if a segment is annotated as 'A' and 'C', and the prediction is 'A' and 'D', false negative increments for class 'C' and hence, the recall for 'C' ($\frac{TP}{TP+FN}$) decreases. In case of class 'D', false positive is incremented and the precision ($\frac{TP}{TP+FP}$) decreases. Class 'A' is unaffected as true positive is incremented independently.

Table I shows the comparison between direct segment classification and segment classification derived from sentence classes. We observe an increase in recall for most categories for derived segment classification except for 'Data Retention', 'Do Not Track' and 'Privacy contact information'. Increase in recall supports the idea that segment-level classification misses predictions that a sentence-level classifier would not miss. The purpose of training a machine learning model is to learn differentiating parameters between categories. Training with segments containing a mix of different categorical sentences does not help the training purpose. Consider this segment: *"We will share personal information with companies, organizations or individuals outside of Google when we have your consent to do so. We require opt-in consent for the sharing of any sensitive personal information."*, annotated as both 'Third Party Sharing/Collection' and 'User Choice/Control'. Using such a segment for training, where the sentence *"We require opt-in consent for the sharing of any sensitive personal information."* is (mis)representing 'Third Party Sharing/Collection', introduces a flaw during training itself, possibly attributing to low recall in direct segment classification.

Precision either increased or remained same for 'Third Party Sharing/Collection', 'Policy Change', and 'Practice not covered' categories. Increase in false positive instances reduced precision in derived segment classification for other categories. Since Polisis was not trained for sentence-level classification, it can be an attributing factor in lower precision when using it at a sentence level. For example, if a 'User Choice/Control' sentence is grouped together with a different category sentence in a segment, as seen in the above segment example, then the language features characterizing 'User Choice/Control' could get mistaken as characteristics of other categories. Note that the threshold we used for sentence classification is quite high (0.9). Even then, instances of false classification is significant, which further shows the extent to which grouping different categories has affected the trained classifier.

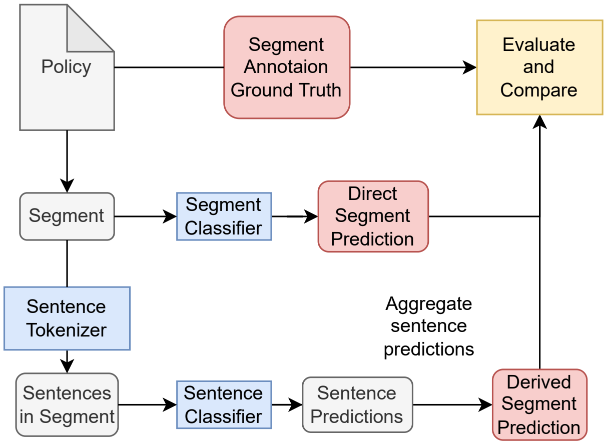

Fig. 3. Schematic of sentence and segment-level classifier evaluation.

The derived segment classification approach had higher recall and higher F-1 score with a slight decrease in precision. The higher F-1 score is a strong indication that segment classification by aggregating sentence categories improves segment classification, despite using a common classifier for both sentences and segments. The next step was thus to implement a proper sentence-level classifier and use that directly for deriving segment categories.

V. SENTENCE CLASSIFICATION

We next assess if a classifier trained specifically for sentence-level classification can improve the performance of derived segment classification. However, obtaining data for training sentence-level classifiers has been a challenge. OPP-115 corpus has segment annotations and attribute annotations for partial sentences in a segment. The partial sentences in attribute annotations are extracted segment fragments and not an ideal representation of a policy sentence. As a way forward, we instead used Polisis and created a baseline data set with sentence-level category labels for the policies in OPP-115. To do so, we first ran each sentence in a segment through the Polisis classifier, and obtained the corresponding probabilities for each possible category. Next, for a specific sentence, we discard any category that does not appear in the ground truth OPP-115 annotations for the segment in which the sentence

TABLE II
BERT: PERFORMANCE METRICS OF DIRECT VERSUS DERIVED SEGMENT CLASSIFICATION

Category	**Direct**			**Derived**		
	F-1	**Prec.**	**Rec.**	**F-1**	**Prec.**	**Rec.**
First Party Collection/Use	0.87	0.82	0.92	0.63	0.95	0.47
Third Party Sharing/Collection	0.79	0.89	0.72	0.72	0.86	0.62
User Access, Edit and Deletion	0.61	0.82	0.48	0.23	1.00	0.13
Data Retention	0.07	0.75	0.04	0.53	0.91	0.38
Data Security	0.74	0.88	0.64	0.70	0.81	0.61
International and Specific Audiences	0.87	0.94	0.81	0.80	0.93	0.69
Do Not Track	0.87	0.93	0.81	0.84	1.00	0.72
Policy Change	0.72	0.91	0.60	0.62	0.76	0.53
User Choice/Control	0.69	0.74	0.65	0.68	0.74	0.63
Introductory/Generic	0.62	0.52	0.77	0.76	0.77	0.75
Practice not covered	0.40	0.75	0.27	0.80	0.72	0.89
Privacy contact information	0.71	0.86	0.61	0.85	0.74	0.99
micro avg	0.73	0.77	0.69	0.71	0.81	0.62
macro avg	0.66	0.82	0.61	0.68	0.85	0.62

TABLE III
XLNET: PERFORMANCE METRICS OF DIRECT VERSUS DERIVED SEGMENT CLASSIFICATION

Category	**Direct**			**Derived**		
	F-1	**Prec.**	**Rec.**	**F-1**	**Prec.**	**Rec.**
First Party Collection/Use	0.87	0.95	0.79	0.89	0.82	0.96
Third Party Sharing/Collection	0.88	0.95	0.81	0.87	0.79	0.96
User Access, Edit and Deletion	0.71	0.86	0.61	0.80	0.88	0.74
Data Retention	0.67	0.93	0.52	0.67	0.74	0.62
Data Security	0.73	0.96	0.58	0.84	0.78	0.91
International and Specific Audiences	0.91	0.95	0.88	0.91	0.88	0.93
Do Not Track	0.98	1.00	0.97	0.90	1.00	0.81
Policy Change	0.73	0.87	0.63	0.82	0.81	0.82
User Choice/Control	0.70	0.91	0.57	0.79	0.73	0.86
Introductory/Generic	0.67	0.93	0.53	0.79	0.81	0.77
Practice not covered	0.37	0.81	0.24	0.67	0.70	0.65
Privacy contact information	0.77	0.95	0.64	0.82	0.84	0.81
micro avg	0.77	0.93	0.65	0.83	0.80	0.86
macro avg	0.75	0.92	0.65	0.81	0.82	0.82

appears. In the remaining categories, the one with the highest prediction probability is assigned as the sentence's category. We assigned a single category to a sentence as a sentence should ideally communicate a single data practice. There are $11,033$ sentences in the OPP-115 corpus, manual annotation of which requires contribution from domain experts, and is out of scope in this paper.

BERT and XLNet multi-label classifiers have recently shown improved performance in segment-level classification of privacy policies [9], [10]. With that in mind, we chose both BERT and XLNet to build our sentence-level classifier. We then conducted a similar evaluation as we did for Polisis in section IV. We create two versions of the two chosen models, where the first version (segment classifier) is trained using segment-level annotations from OPP-115, and the second version (sentence classifier) is trained using sentence-level annotations from the baseline data set. The first version can directly predict a segment's labels, while the second version is used within our derived segment classification approach (derive a segment's prediction from predicted sentence labels). Using the annotations on segments from the OPP-115 corpus as ground truth, we then compare the prediction performance metrics of the two classifier versions. The evaluation for both models is illustrated in figure 3. We use 80% of the data sets for training, 10% for validation, and the remaining 10% for testing.

A. BERT Classifier

The BERT classifier is built on a pre-trained model having a 12-layer, 768-hidden layers and 12-heads architecture, with a total of 110M parameters[2]. The segment classifier and sentence classifier are both trained with the same parameters. For the segment classifier, we use a sigmoid function to compute probabilities as it is a case of multi-label classification. For the sentence classifier, we use a softmax function as we want to assign a single category label to a sentence. After both versions of the model are trained, we compare them using performance metrics shown in table II. To be consistent with our earlier evaluation using Polisis, we have used a threshold of 0.5 for segment-level classification, and 0.9 when using the sentence classifier in derived segment classification.

BERT direct and derived versions performed better than both versions of Polisis. The major improvement that BERT has is in recall, where Polisis was lacking. The direct segment classification with BERT lowers precision for 'First Party Collection/Use', 'User Access, Edit and Deletion', 'Do Not Track' and 'User Choice/Control' in comparison to Polisis direct segment classification. Decrease in precision for these categories is relatively small when compared with the overall precision gain and a significant increase in recall. The overall micro average precision increases from 73% to 78% for BERT and micro average recall increases from 24% to 69%.

For BERT, we observe an increase in precision while using derived segment classification. Micro average precision observed a 4% increase, while macro average precision increased by 3%. Recall increased by 1% for macro average, but decreased by 7% for micro average. As the classes are not balanced, a micro average decrease correctly represents the potential loss in recall while using the sentence classifier. Classes like 'Data Retention' and 'Practice Not covered', which has significantly less examples and most often occur with other category sentences, saw an increase in both precision and recall. This is another example of how low frequency categories get affected when trained at a segment level, due to the presence of other categories in the segment. But treating such categories individually helped in their correct classification.

Even though sentence-level classification increased performance for some of the classes (as indicated by increase in macro average F-1 score), yet some of the classes suffered as well, which is indicated by a decrease in the micro average F-1 score. The decline in recall is the attributing factor. We have kept a high threshold for prediction probability (0.9) to be consistent with the Polisis evaluation, which has removed many of the correctly predicted tags. Decreasing this threshold can increase recall at the expense of precision; if the threshold is dropped to 0.8, then precision drops by around 4% and recall increases by 3%.

[2] https://tfhub.dev/google/bert_uncased_L12_H768_A12/1

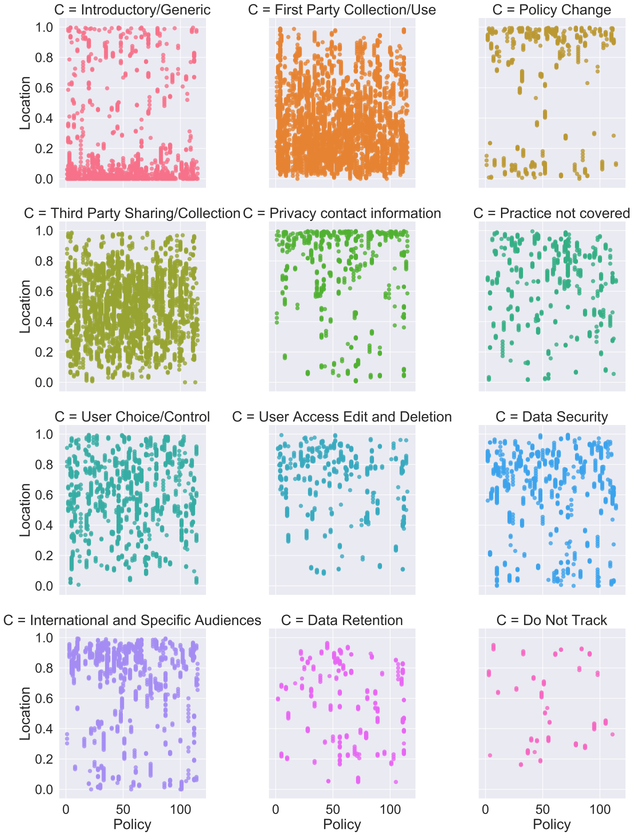

Fig. 4. Location of *segment* categories across 115 policies in the OPP-115 corpus. Location zero is first segment; location one is last segment.

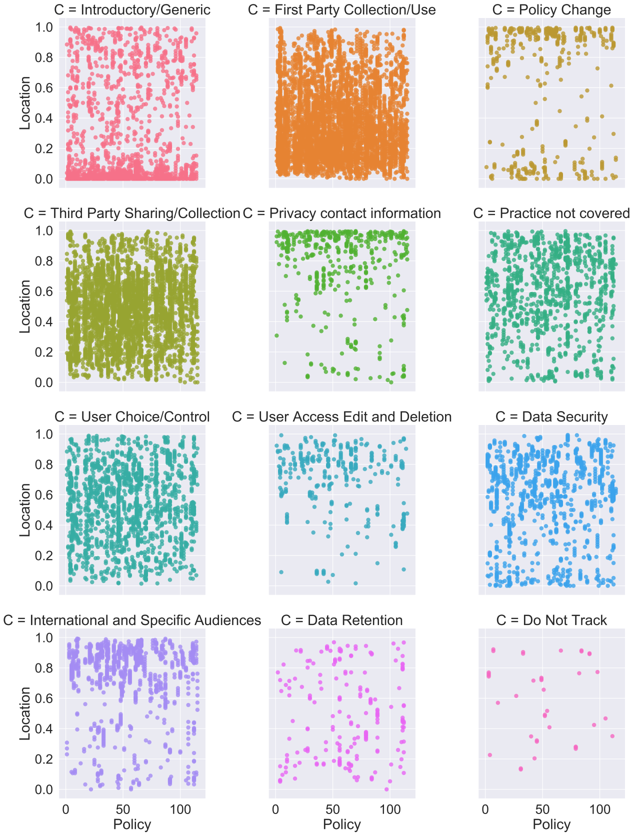

Fig. 5. Location of *sentence* categories across 115 policies in the OPP-115 corpus. Location zero is first sentence; location one is last sentence.

B. XLNet Classifier

The XLNet model is adopted from the work done by Mustapha et al. [9], using the code provided in their GitHub repository[3]. The model is trained for 5 epochs with batch sizes of 8 for training and 16 for validation, where the learning rate is set to 0.001. Similar to our BERT implementation, two versions of classifiers were trained, one for segment and another for sentence level. The data sets used for training and thresholds used are also same.

Table III shows the result from XLNeT models where the direct column presents the performance of the segment-level classifier and the derived column presents performance from segment classification using the sentence-level classifier. Both versions of the XLNet model outperformed all the previous models in this study, confirming the baseline established in [9]. The F-1 score increased for almost all the categories except 'Third Party Sharing/Collection' (1% decline) and 'Do Not Track' (8% decline). But overall, the average F-1 score increased by 6%. Increase in F-1 score is attributed to very significant increase in recall (21% micro and 17% macro). The increase in recall is a strong indicator of how much information can get lost while analyzing policies at a segment level with automated tools.

Precision decreased for the sentence-level classifier, indicating an increase in false classification. The most probable cause for decrease can be attributed to not using standardized sentence-level annotations. For training data, Polisis's classifier was used to assign sentence categories and used as ground truth. Since Polisis's classifier faces misrepresentation of categorical characteristics (due to segment-level training), the created sentence-level training data can also inherit the flaws. Another problem could be the annotation in OPP-115 itself, where the segment-level annotations might have missed the categories of some sentences.

VI. Policy Categorical Structure

The primary objective of implementing a sentence classifier is to capture categorical information otherwise missed by a segment classifier. Since the XLNet classifiers for both segment and sentence classification had the best performance, we used them to delve deeper into our analysis of privacy policy categorical structure analysis. Figures 4 and 5 show the distribution of predicted segment classes and sentence classes across the 115 policies in OPP-115 using the XLNet segment and sentence classifiers respectively. The vertical axis labeled as 'Location' represents the position where a category occurs in the document. The position of a category occurrence, given by the segment or sentence number, is normalized relative to

[3] https://github.com/euranova/privacy-policy-classification-xlnet

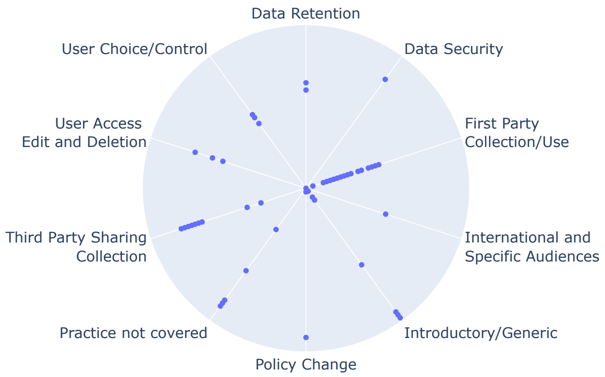

Fig. 6. Category wise *segment* location in a 2015 Google policy. The center is the beginning of the document.

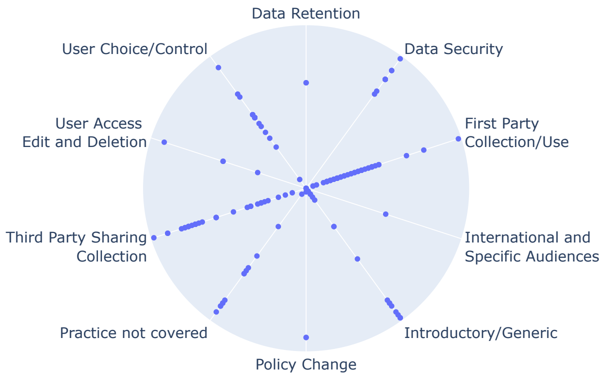

Fig. 7. Category wise *sentence* location in a 2015 Google policy. The center is the beginning of the document.

the total number of segments or sentences respectively for a given policy. A standard scale of 0 to 1 can thus represent the location of the category across all the policies, despite them having variable lengths. The horizontal axis labeled as 'Policy' represents each of the 115 policies of the corpus (identified with numbers from 1 to 115). The organization name is removed to reduce clutter. For easier visualization, the plot for each category is separated. To exemplify, the points vertical to point 50 on the horizontal 'Policy' axis in the 'First Party Collection/Use' plot in figure 4 represents the distribution of the said category at a segment level across the 50th policy.

A. Organization of Information

Looking at both the sentence-level (figure 5) and segment-level (figure 4) category distribution, we notice that there is a lack of organization of categories in a policy. If we merge all the categorical segment plots, we observe a significant amount of overlap between all the categories. There are overlapping data practice concepts across a document, without segregation and structure, which can contribute to low interpretability. Reading a document while frequently switching between categorical contexts does not help readers grasp the complete information.

B. Categorical Characteristics: Segment vs Sentence

A noteworthy observation in the same class plots of figures 4 and 5 is the increase in several plotted points at sentence-level classification. This increase provides evidence of the superiority of sentence classifiers in retrieving most categorical information. When each category is examined individually, we learn more about crucial prevailing privacy design practices, which is critical for tackling usability issues associated with the current state of privacy policies. We summarize below few of our observations using both segment-level classification and sentence-level classification, and compare the two kinds of classification.

1) Introductory/Generic, Policy Change, and Privacy contact information: Segments and sentences describing 'Policy Change' and 'Privacy contact information' practices prominently occur towards the end of most policies. Position-wise, both categories occur near each other, suggesting an implicit relationship between the two classes. Policy change information may also be found in the beginning of a document in few cases, interspersed with introductory or generic text.

2) First Party Collection/Use and Third-Party Sharing/Collection: 'First Party Collection/Use' and 'Third Party Sharing/Collection' form the backbone of a privacy policy. Segment-level classification shows that 'First-party collection/Sharing' occupies the beginning to middle portion and 'Third Party Sharing/Collection' mostly the middle of a policy.

The sentence-level classification reveals that the two categories span almost the entirety of a policy. This implies that for a user to be educated on how their data is being collected, shared, and used, they must read the entire policy. An example of Google's 2015 privacy policy which is classified at the segment and sentence levels, and represented as polar plots (figure 6 and 7 respectively), further elaborates the aforementioned issue. A dot is one sentence or segment in a document. The segment categories for Google's privacy policy show a note of structure where 'First Party Collection/Use' occupies the beginning of the policy, followed by 'Third Party Sharing/Collection' (figure 6). The sentence-level classification (figure 7) demonstrates that 'Third Party Collection/Sharing' is spread throughout the policy. Regulations such as GDPR requires complete communication of sharing practices to users. Even when organizations are describing these practices, they are often dispersed throughout a document and mixed with other data practice information. As seen in the plot in figure 7, some of the 'Third Party Sharing/Collection' sentences occur alongside 'User Access, Edit and Deletion' and 'User Control/Choice' sentences. Mixing of these different categorical sentences in a segment reduces segment classification performance evidenced by the low recall for 'User Access, Edit and

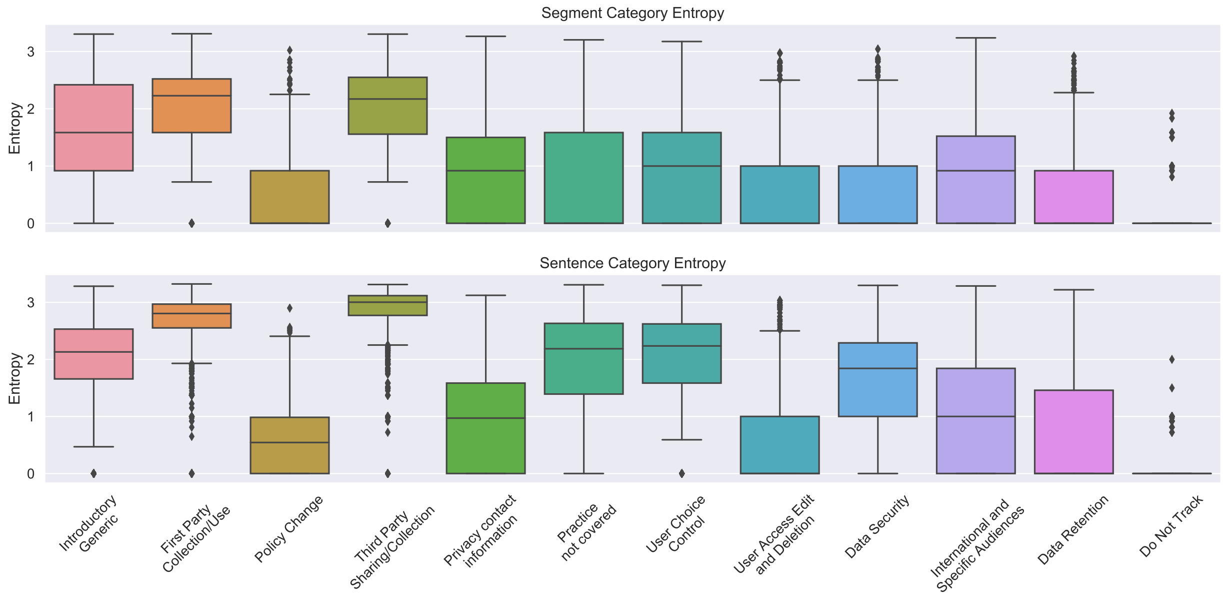

Fig. 8. Box plot of segment and sentence category entropy across privacy policies from Alexa top 5000 websites.

Deletion' and 'User Control/Choice' in table II (direct).

3) User Choice/Control: The 'User Choice/Control' practices empower users with options to control their data collection and use. It is one of the most relevant pieces of information for users. The fact that we see a drastic increase in 'User Control/Choice' statements from segment classification to sentence classification also raises questions about how these policies are authored. With or without intent, 'User Choice/Control' statements are hidden among other categorical statements, making it hard for users to know about their options over how their data is being handled. Segment classification with XLNet had a recall of 0.57 for the category, whereas sentence classification achieves a recall of 0.86. This shows that at segment level, most of this information gets lost due to noise added by other categories. Opt in/out choice identification research have thus employed sentence-level categorization to overcome the noise [14], [26]. 'User Access, Edit and Deletion' also shows a high co-occurrence with 'Privacy contact information' during sentence classification. This may suggest that access, edit and deletion options given to users often require contacting the service providers.

Overall, all categories saw a boost in recall when classified at the sentence level. Positioning of most categories is not well defined leading to lower readability. Sentence-level classification also reveals implicit relationship between categories that is otherwise missed by segment classification.

C. Information Distribution Entropy

The positional distribution plots of categories across policies in the OPP-115 corpus demonstrated uncertainty, with little or no structure in organization. For a larger analysis, we classify the policies from the Alexa top 5000 websites. We consider each policy document to be composed of 10 regions made up of one or more contiguous segments. We selected 10 as the number of regions since domain experts have defined 10 high level categories for policies; alignment researchers have also considered 10 regions in policies while clustering segments [27].

To quantify the extent of positional uncertainty of a particular data practice, we utilize Shannon's entropy. If P_i^c is the probability that a category C can be found in region i of a policy, then we calculate the entropy for category C as $\sum_{i=1}^{i=10} -P_i^c log(P_i^c)$. The entropy values for each category are presented as box plots in figure 8, observed at both the segment and sentence levels.

Segment-level entropy values may seem to indicate that at least 50% of the policies restrict certain categories such as 'Policy Change', 'User Access, Edit and Deletion', 'Data Security' and 'Data Retention' to one region. Sentence-level entropy values indicate higher uncertainty in positioning across most categories.

Policy topics such as 'Policy Change' and 'User Access, Edit, and Deletion' are observed to be the most organized in a policy. In about 25% of the policies, this information is commonly located within one region. The lower inter-quartile range implies that, for majority of policies, the two categories have relatively high alignment. There are observable outliers that exhibit uncertainty as well; in other words, not every policy show the desirable organization.

From the entropy values and positional distribution of 'First Party Collection/Use' and 'Third Party Sharing/Collection', it can be reasoned that all the other categorical statements are implicitly dependent on and cannot be complete without 'First Party Collection/Use' or 'Third Party Sharing/Collection' statements. If this is indeed the case, then logically segregating these categories from other categories may not be possible. Hence, a revision of the categorical definition is due that incorporates the observed dependency in a hierarchical fashion. Segregation between ‘First Party Collection/Use’ and ‘Third Party Sharing/Collection’ still remains an important requirement. Outliers with lower entropy do exists in both categories, suggesting that some policy writers have attempted to better present information in these two categories.

The primary concern for the ‘Do Not Track’ category is not organization, but instead lack of presence. Most websites do not honor ‘Do Not Track’ and ignore this signal.

VII. Conclusion and Future Work

In this paper, we report on the sentence-level classification of privacy policies. We initiated our work with the premise that segment-level study of privacy policies misses on crucial privacy relevant information. We evaluated Polisis, BERT, and XLNet based segment and sentence classifiers with OPP-115, showing that derived segment classification, i.e. inferring a segment label from sentence labels, achieves a finer granularity and outperforms direct segment classification. Using sentence classification to define segment categories, we observed a 6% improvement in F-1 score compared to traditional segment trained classifiers using the same model. Structural analysis of privacy policies with sentence classification also revealed vital usability issues, which are not observable with segment-level analysis. A corpus with sentence-level annotation can help develop better methods to retrieve information at a finer granularity, since usability and organization issues are most visibly evident at the sentence level. In the future, we aim to focus on tackling these issues, towards the design of better sentence-level classifiers that can identify usability issues and generate effective guidance for policy writers.

References

[1] L. F. Cranor, “Necessary but not sufficient: Standardized mechanisms for privacy notice and choice,” *Journal on Telecommunications and High Technology Law*, vol. 10, p. 273, 2012.

[2] F. Schaub, R. Balebako, A. L. Durity, and L. F. Cranor, “A design space for effective privacy notices,” in *Symposium On Usable Privacy and Security*, 2015, pp. 1–17.

[3] S. Zimmeck, “The information privacy law of web applications and cloud computing,” *Santa Clara Computer & High Technology Law Journal*, vol. 29, p. 451, 2012.

[4] A. M. McDonald and L. F. Cranor, “The cost of reading privacy policies,” *Journal of Law and Policy for the Information Society*, vol. 4, p. 543, 2008.

[5] N. Sadeh, A. Acquisti, T. D. Breaux, L. F. Cranor, A. M. McDonald, J. R. Reidenberg, N. A. Smith, F. Liu, N. C. Russell, F. Schaub, and S. Wilson, “The usable privacy policy project,” in *Technical report, CMU-ISR-13-119*. Carnegie Mellon University, 2013.

[6] W. Ammar, S. Wilson, N. Sadeh, and N. A. Smith, “Automatic categorization of privacy policies: A pilot study,” in *Technical Report, CMU-LTI-12-019*. Carnegie Mellon University, 2012.

[7] S. Wilson, F. Schaub, A. A. Dara, F. Liu, S. Cherivirala, P. G. Leon, M. S. Andersen, S. Zimmeck, K. M. Sathyendra, N. C. Russell, T. B. Norton, E. Hovy, J. Reidenberg, and N. Sadeh, “The creation and analysis of a website privacy policy corpus,” in *Annual Meeting of the Association for Computational Linguistics*, 2016, pp. 1330–1340.

[8] H. Harkous, K. Fawaz, R. Lebret, F. Schaub, K. G. Shin, and K. Aberer, “Polisis: Automated analysis and presentation of privacy policies using deep learning,” in *27th USENIX Security Symposium*, 2018, pp. 531–548.

[9] M. Mustapha, K. Krasnashchok, A. Al Bassit, and S. Skhiri, “Privacy policy classification with xlnet (short paper),” in *Data Privacy Management, Cryptocurrencies and Blockchain Technology*. Springer, 2020, pp. 250–257.

[10] N. Mousavi Nejad, P. Jabat, R. Nedelchev, S. Scerri, and D. Graux, “Establishing a strong baseline for privacy policy classification,” in *International Federation for Information Processing: International Conference on Information Systems Security and Privacy Protection*. Springer, 2020, pp. 370–383.

[11] S. Zimmeck and S. M. Bellovin, “Privee: An architecture for automatically analyzing web privacy policies,” in *23rd USENIX Security Symposium*, 2014, pp. 1–16.

[12] E. Costante, Y. Sun, M. Petković, and J. Den Hartog, “A machine learning solution to assess privacy policy completeness: (short paper),” in *Proceedings of the 2012 Association for Computing Machinery Workshop on Privacy in the Electronic Society*, 2012, pp. 91–96.

[13] F. Liu, S. Wilson, P. Story, S. Zimmeck, and N. Sadeh, “Towards automatic classification of privacy policy text,” in *Techical Report CMU-ISR-17-118R and CMULTI-17-010*. Carnegie Mellon University, 2018.

[14] K. M. Sathyendra, F. Schaub, S. Wilson, and N. Sadeh, “Automatic extraction of opt-out choices from privacy policies,” in *2016 Association for the Advancement of Artificial Intelligence Fall Symposium Series*, 2016.

[15] K. M. Sathyendra, S. Wilson, F. Schaub, S. Zimmeck, and N. Sadeh, “Identifying the provision of choices in privacy policy text,” in *Conference on Empirical Methods in Natural Language Processing*, 2017, pp. 2774–2779.

[16] V. Bannihatti Kumar, R. Iyengar, N. Nisal, Y. Feng, H. Habib, P. Story, S. Cherivirala, M. Hagan, L. Cranor, S. Wilson *et al.*, “Finding a choice in a haystack: Automatic extraction of opt-out statements from privacy policy text,” in *Proceedings of The Web Conference 2020*, 2020, pp. 1943–1954.

[17] S. Zimmeck, P. Story, D. Smullen, A. Ravichander, Z. Wang, J. R. Reidenberg, N. C. Russell, and N. Sadeh, “Maps: Scaling privacy compliance analysis to a million apps,” *Proceedings of Privacy Enhancing Technology*, vol. 2019, p. 66, 2019.

[18] P. Story, S. Zimmeck, A. Ravichander, D. Smullen, Z. Wang, J. Reidenberg, N. C. Russell, and N. Sadeh, “Natural language processing for mobile app privacy compliance,” in *Association for the Advancement of Artificial Intelligence: Spring Symposium Series*, 2019.

[19] S. Wilson, F. Schaub, F. Liu, K. M. Sathyendra, D. Smullen, S. Zimmeck, R. Ramanath, P. Story, F. Liu, N. Sadeh, and N. A. Smith, “Analyzing privacy policies at scale: From crowdsourcing to automated annotations,” *ACM Transactions on the Web*, vol. 13, no. 1, pp. 1–29, 2018.

[20] R. Amos, G. Acar, E. Lucherini, M. Kshirsagar, A. Narayanan, and J. Mayer, “Privacy policies over time: Curation and analysis of a million-document dataset,” in *Proceedings of the Web Conference*, 2021, pp. 2165–2176.

[21] P. Bojanowski, E. Grave, A. Joulin, and T. Mikolov, “Enriching word vectors with subword information,” *Transactions of the Association for Computational Linguistics*, vol. 5, pp. 135–146, 2017.

[22] T. Mikolov, K. Chen, G. Corrado, and J. Dean, “Efficient estimation of word representations in vector space,” *arXiv preprint arXiv:1301.3781*, 2013.

[23] J. Pennington, R. Socher, and C. D. Manning, “Glove: Global vectors for word representation,” in *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing*, 2014, pp. 1532–1543.

[24] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, “Bert: Pre-training of deep bidirectional transformers for language understanding,” *arXiv preprint arXiv:1810.04805*, 2018.

[25] Z. Dai, Z. Yang, Y. Yang, J. Carbonell, Q. V. Le, and R. Salakhutdinov, “Transformer-xl: Attentive language models beyond a fixed-length context,” *arXiv preprint arXiv:1901.02860*, 2019.

[26] K. M. Sathyendra, A. Ravichander, P. G. Story, A. W. Black, and N. Sadeh, “Helping users understand privacy notices with automated query answering functionality: An exploratory study,” in *Technical report, CMU-ISR-17-114R*. Carnegie Mellon University, 2017.

[27] R. Ramanath, F. Liu, N. Sadeh, and N. A. Smith, “Unsupervised alignment of privacy policies using hidden markov models,” in *Annual Meeting of the Association for Computational Linguistics*, 2014, pp. 605–610.