Analyzing Privacy Policies at Scale: From Crowdsourcing to Automated Annotations

SHOMIR WILSON, FLORIAN SCHAUB, FREDERICK LIU,
KANTHASHREE MYSORE SATHYENDRA, DANIEL SMULLEN, SEBASTIAN ZIMMECK,
ROHAN RAMANATH, PETER STORY, FEI LIU, and NORMAN SADEH,
Carnegie Mellon University
NOAH A. SMITH, University of Washington

Website privacy policies are often long and difficult to understand. While research shows that Internet users care about their privacy, they do not have the time to understand the policies of every website they visit, and most users hardly ever read privacy policies. Some recent efforts have aimed to use a combination of crowdsourcing, machine learning, and natural language processing to interpret privacy policies at scale, thus producing annotations for use in interfaces that inform Internet users of salient policy details. However, little attention has been devoted to studying the accuracy of crowdsourced privacy policy annotations, how crowdworker productivity can be enhanced for such a task, and the levels of granularity that are feasible for automatic analysis of privacy policies. In this article, we present a trajectory of work addressing each of these topics. We include analyses of crowdworker performance, evaluation of a method to make a privacy-policy oriented task easier for crowdworkers, a coarse-grained approach to labeling segments of policy text with descriptive themes, and a fine-grained approach to identifying user choices described in policy text. Together, the results from these efforts show the effectiveness of using automated and semi-automated methods for extracting from privacy policies the data practice details that are salient to Internet users' interests.

CCS Concepts: • **Information systems → Crowdsourcing**; • **Security and privacy → Human and societal aspects of security and privacy**; • **Human-centered computing → Human computer interaction (HCI)**; • **Social and professional topics → Privacy policies**; • **Computing methodologies → Natural language processing**;

Additional Key Words and Phrases: Privacy, privacy policies, crowdsourcing, machine learning, natural language processing, human computer interaction (HCI)

This research has been partially funded by the National Science Foundation under grant agreement CNS-1330596.
Authors' addresses: S. Wilson, College of Information Sciences and Technology, Westgate Building, Pennsylvania State University, University Park, PA 16802 USA; email: shomir@psu.edu; F. Schaub, University of Michigan, School of Information, 105 S State St, Ann Arbor, MI 48109, USA; email: fschaub@umich.edu; K. M. Sathyendra, D. Smullen, R. Ramanath, P. Story, and N. Sadeh, School of Computer Science, Carnegie Mellon University, 5000 Forbes Ave, Pittsburgh, PA 15213 USA; emails: kanthashree.ms@gmail.com, dsmullen@cs.cmu.edu, ronramanath@gmail.com, pstory@andrew.cmu.edu, sadeh@cs.cmu.edu; S. Zimmeck, Department of Mathematics and Computer Science, Wesleyan University, Science Tower 655, 265 Church St, Middletown, CT 06459-0128 USA; email: szimmeck@wesleyan.edu; F. Liu, Computer Science Department, University of Central Florida, 4328 Scorpius St, Orlando, FL 32816-2362 USA; email: feiliu@cs.ucf.edu; N. A. Smith, Paul G. Allen School of Computer Science & Engineering, University of Washington, Box 352350, 185 E Stevens Way NE, Seattle, WA 98195 USA; emails: nasmith@gmail.com.

ACM 1559-1131/2018/12-ART1 $15.00
https://doi.org/10.1145/3230665

ACM Reference format:
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 Trans. Web* 13, 1, Article 1 (December 2018), 29 pages.
https://doi.org/10.1145/3230665

1 INTRODUCTION

Privacy policies are verbose, complex legal documents that provide notices about the data practices of websites and online service providers. McDonald and Cranor (2008) showed that if users were to read the privacy policies of every website they access, they would spend an unreasonable fraction of their time doing so; additionally, they found that study participants were largely unable to answer basic questions about what these privacy policies say. Unsurprisingly, many people do not read website privacy policies (Federal Trade Commission 2012), which are often drafted to ensure legal and regulatory compliance rather than to effectively inform users (Schaub et al. 2017; Reidenberg et al. 2015a). Despite these limitations, website privacy policies remain Internet users' primary sources of information on how companies collect, use, and share their data.

Efforts to codify privacy policies, such as the development of the Platform for Privacy Preferences (P3P) standard (Cranor et al. 2006) or more recent initiatives like Do Not Track (DNT) (Doty et al. 2016), have been met with resistance from website operators (McDonald 2013; Cranor 2012). While the vast majority of prominent websites have natural language privacy policies (some required by legal regulation, such as the California Online Privacy Protection Act (Official California Legislative Information 2003)), many service providers are reluctant to adopt machine-implementable solutions that would force them to further clarify their privacy practices or to commit to more stringent practices.

In response to this issue, recent efforts have focused on the development of approaches that rely on crowdsourcing and natural language processing to annotate important elements of privacy policies. This includes PrivacyChoice (acquired by AVG), Terms of Service; Didn't Read (ToS;DR) (Tos;DR 2012), work by Zimmeck and Bellovin (2014), and the Usable Privacy Policy Project (Sadeh et al. 2013), by far the most comprehensive effort of its kind to date.[1] Crowdsourcing is typically applied to tasks that are still difficult for computers to solve, but can be easily solved by humans (Quinn and Bederson 2011). Crowdsourcing the extraction of data practices from privacy policies faces a particular challenge: it has been shown that the length and complexity of privacy policies makes them difficult to understand and interpret by most Internet users (Jensen and Potts 2004; McDonald and Cranor 2008). Even experts and trained analysts may disagree on their interpretation (Reidenberg et al. 2015a).

In this article, we describe our efforts toward automating the analysis of privacy policies at a large scale. We begin with an annotation procedure based entirely on crowdworker efforts and then describe subsequently developed procedures that require diminishing amounts of human labor. Four main contributions represent this trajectory:

(1) **A demonstration of the feasibility of crowdsourcing to answer questions about privacy policies** (Section 3): We demonstrate that, by requiring a high level of agreement within a group of crowdworkers (≥80%), their answers to questions about privacy policies can be aggregated to produce results with high accuracy (>96%).

[1]This article describes research conducted within the Usable Privacy Policy Project (www.usableprivacy.org). It includes and expands upon research that previously appeared in conference proceedings (Wilson et al. 2016c; Sathyendra et al. 2017b; Wilson et al. 2016a).

(2) **A technique to highlight a small number of paragraphs in a given privacy policy to help crowdworkers answer questions** (Section 4): We use relevance models for each question to identify the paragraphs in a privacy policy that are most applicable for answering the question. Through a second crowdworker study, we show that highlighting those paragraphs results in a slight reduction in mean time for task completion and no significant loss of accuracy.
(3) **A method for automatically assigning categories of data practices, as defined by legal experts, to text segments in privacy policies** (Section 5): We use the OPP-115 Corpus of privacy policies (Wilson et al. 2016b) to train and test classifiers to label policy text with common themes (i.e., the *categories* in the OPP-115 annotation scheme). Paragraph labeling and sentence labeling are both possible, though performance on the former is slightly greater.
(4) **A method for automatically finding user choices in privacy policy text** (Section 6): Privacy policies often describe how a website or app user can make choices about how their personal information is collected or used. We show results from a multistage system that identifies when a sentence contains a choice and then categorizes it based on common attributes.

Finally, in Section 7, we describe some challenges for future research, motivated by both the practical value of communicating privacy information and the potential for related basic research in natural language processing.

2 RELATED WORK

Notice and choice are core principles of information privacy protection under the Fair Information Practice Principles (Federal Trade Commission 2000). However, privacy policies are often long and complex, and they make use of technical jargon not readily understandable by average Internet users (Cranor 2012; Cate 2010; Reidenberg et al. 2015a; Schaub et al. 2015, 2017). Readability can be used as an approximation for comprehensibility, and the readability of privacy policies has been studied extensively (Jensen and Potts 2004). Privacy policies have been evaluated with readability metrics in different domains, such as energy companies' terms and conditions (Luger et al. 2013), online social networks (Meiselwitz 2013), or health care notices (Ermakova et al. 2015). Findings suggest that understanding privacy policies requires reading skills and patience that exceed those of the average Internet user. Some structured privacy policies, such as financial institutions' annual privacy notices, lend themselves to automated analysis (Cranor et al. 2016). However, most privacy policies are unstructured and contain substantial vagueness and ambiguity (Bhatia et al. 2016b; Reidenberg et al. 2016).

Various efforts to extract data practices from privacy policies have used crowdsourcing techniques (Schaub et al. 2016). For instance, ToS;DR (Tos;DR 2012) is a community-driven effort to analyze websites' privacy policies and grade their respect for users' privacy. However, ToS;DR's open-ended approach to annotations is difficult to automate or analyze at scale. The manual analysis of privacy policies has been recognized as a serious bottleneck to modeling their contents, and some prior efforts have aimed to increasingly automate the annotation process (Reidenberg et al. 2015a; Wilson et al. 2016c).

A common approach to crowdsourcing is to split a complex task into smaller subtasks that are easier to solve (Chilton et al. 2013; Kittur et al. 2011; Negri et al. 2011). This approach works well for labeling tasks, such as tagging or categorizing images, but privacy policies are substantially more complex. Descriptions of a particular data practice may be distributed throughout a policy. For example, in one section, a policy may claim that data is not shared with third parties, and later it may list exceptional third parties that receive data. This complexity makes it difficult to

correctly interpret a policy's meaning without reading it in its entirety. Thus, a policy's text cannot be trivially partitioned into smaller reading tasks for crowdworkers to annotate in parallel, since integrating contradictory annotations becomes a difficult problem.

Few efforts have been made to crowdsource tasks as complex as annotating privacy policies. André et al. (2014) investigate crowdsourcing of information extraction from complex, high-dimensional, and ill-structured data. However, their focus is on classification via clustering, rather than on human interpretation to answer questions. Breaux and Schaub (2014) take a bottom-up approach to annotating privacy policies by asking crowdworkers to highlight specific action words and information types in a privacy policy. However, this creates the challenge of reconciling results from multiple questions and segments of policy text into a coherent representation of a website's data practices (Bhatia et al. 2016a). A way forward could be the automatic assignment of category labels to policy segments (Wilson et al. 2016b).

Similarly, few efforts have measured the accuracy of policy annotations obtained from crowdworkers. Reidenberg et al. (2015a) studied how experts, trained analysts, and crowdworkers disagree when interpreting privacy policies. They conducted a qualitative analysis based on six privacy policies and found that even experts are subject to notable disagreements. Moreover, data practices that involve sharing with third parties appeared to be a particular source of disagreement among the annotation groups. On the other hand, Breaux and Schaub (2014) found that crowdworkers working in parallel identified more keywords than expert annotators. Both studies were based on a small number of privacy policies (six and five, respectively). In contrast, we assess crowdworkers' accuracy and agreement with trained analysts using a larger set of privacy policies.

The potential for the application of natural language processing (NLP) and information retrieval techniques to legal documents has been recognized by the legal community (Mahler 2015). In this regard, the analysis of privacy policies can benefit from the advances in applying NLP techniques to other types of legal documents. Notably, Bach et al. (2013) use multi-layer sequence learning models and integer linear programming to learn logical structures of paragraphs in legal articles. Galgani et al. (2012) present a hybrid approach to summarization of legal documents based on creating rules to combine different types of statistical information about text. Montemagni et al. (2010) investigate the peculiarities of the language in legal text with respect to other types of text by applying shallow parsing.

Prior computational work on privacy policy text has used information extraction techniques to identify information types collected by websites (Costante et al. 2013) or to answer categorical questions about privacy (Ammar et al. 2012). Similarly, Zimmeck and Bellovin (2014) complement ToS;DR data with supervised policy analysis. Their analysis addresses six binary questions (e.g., whether a policy provides for limited retention or allows ad tracking), reaching F_1 scores between 0.6 and 1. Other approaches have applied topic modeling to privacy policies (Chundi and Subramaniam 2014; Stamey and Rossi 2009) and have automatically grouped related sections and paragraphs of privacy policies (Liu et al. 2014). Ramanath et al. (2014) introduce an unsupervised model for the automatic alignment of privacy policies and show that hidden Markov models are more effective than clustering and topic models. Liu et al. (2016a) modeled the language of vagueness in privacy policies using deep neural networks. Some research has focused on identifying semantic relationships among concepts and information types expressed in privacy policies (Hosseini et al. 2016; Evans et al. 2017). The Usable Privacy Policy Project (Sadeh et al. 2013) has created a corpus of 115 privacy policies annotated by law students (Wilson et al. 2016b; Oltramari et al. 2017) to facilitate the development of automated privacy policy analysis approaches. Other work has combined information extracted from privacy policies with code analysis techniques to assess whether mobile apps adhere to their own privacy policies (Zimmeck et al. 2017; Slavin et al. 2016). However, since the complexity and vagueness of privacy policy

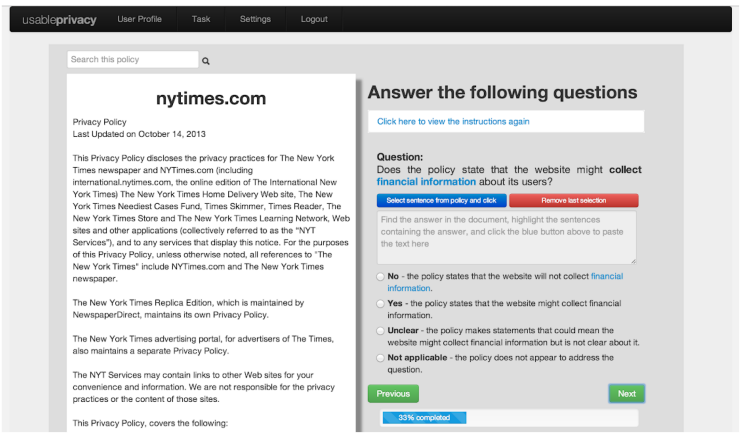

Fig. 1. The privacy policy annotation tool. It displays a privacy policy (*left*) and one of the annotation questions (*right*). Annotators select phrases from the policy text to support their responses and can jump between questions before submitting the completed annotation task.

language makes it difficult to automatically extract complex data practices from privacy policies, we propose to use relevance models to select paragraphs that pertain to a specific data practice and to highlight those paragraphs for annotators.

3 ANALYZING PRIVACY POLICIES WITH CROWDSOURCING

We developed an annotation tool to enable crowdworkers and skilled annotators to annotate privacy policies online (Reidenberg et al. 2015a; Wilson et al. 2016c). In this section, we first describe the online annotation tool and our privacy policy annotation scheme, followed by a study to assess the reliability of crowdsourced privacy policy annotations in comparison to skilled annotators (i.e., law and public policy graduate students) (Wilson et al. 2016c). In particular, we studied whether combining annotations from multiple crowdworkers can approximate expert interpretations, and what levels of agreement should be set to achieve reliable outcomes.

3.1 Privacy Policy Annotation Tool

We developed our online annotation tool for privacy policies in order to provide annotators with an effective interface and workflow to read a privacy policy and answer questions about specific data practices described in the privacy policy. The annotation tool was developed in an iterative user-centered design process that included multiple pilot studies and interview sessions.

The annotation tool, shown in Figure 1, displays a scrollable privacy policy on the left and one annotation question with multiple response options on the right. When selecting an answer, an annotator also selects one or more passages in the policy text that informed their answer before proceeding to the next question, except when selecting "not applicable." These phrase selections serve as supporting evidence for provided annotations. Multiple text segments can be added to (and removed from) the selection field. The selection field is intentionally located between question and response options to integrate it into the annotator's workflow. Additionally, the annotation tool features a search box above the policy, which enables annotators to search for key terms or phrases within the privacy policy before selecting an answer. While annotators must answer all questions before they can complete a policy annotation task, they can jump between questions, answer them in any order, and edit their responses until they submit the task. This flexibility allows users to

account for changes in their interpretation of policy text as they read and understand the privacy policy to answer successive questions.

The policy annotation tool provides users with detailed instructions before they start the task. Users are asked to answer the annotation questions for the main website and to ignore statements about mobile applications or other websites. As part of our study, users were also instructed to ignore statements applying to a limited audience (e.g., Californians, Europeans, or children) in order to obtain answers that consistently reflect an interpretation of the privacy policy for the same jurisdiction (the United States in our case). As part of the annotation interface, we provide definitions for privacy-specific terms used in the questions and the response options (e.g., third parties, explicit consent, core service). Those clarifications are provided as pop-ups when the user hovers over a term highlighted in blue (see Figure 1).

The online annotation tool, the instructions, and the wording of the questions and the response options were refined over multiple iterations. We conducted pilot testing with students and crowdworkers. We also conducted pilot annotations and semi-structured interviews with five skilled annotators to gather feedback, assess the tool's usability, and allow the skilled annotators to familiarize themselves with the tool. Because the skilled annotators provided the gold standard data in our main study, exposing them to the annotation interface at this stage did not affect the results. More specifically, we were interested in eliciting their most accurate interpretations of policies rather than evaluating their interaction with the annotation tool. Pilot tests were conducted with a set of privacy policies different from those used in the actual study. The iterative design resulted in substantial usability improvements. For instance, although we started with a much simpler set of instructions, user tests revealed the need for additional instructions to support the users' interpretation process by reducing ambiguity.

3.2 Privacy Policy Annotation Scheme

We based our annotation scheme on a literature analysis. We identified a small number of data practices and information types that prior work determined to be primary concerns for users. We focused on data practices most frequently mentioned in federal privacy litigation and FTC enforcement actions (Reidenberg et al. 2015b), namely collection of personal information, sharing of personal information with third parties, and whether websites allow users to delete data collected about them. In addition, we were interested in how clearly these practices were described with respect to particularly sensitive information types (Ackerman et al. 1999; Joinson et al. 2010; Leon et al. 2013): contact information, financial information, current location, and health information.

Based on relevant data practices, we devised a set of nine annotation questions: four questions about *data collection* (Q1–Q4, one for each information type above), four questions about *sharing collected information with third parties* (Q5–Q8), and one question about *deletion of user information* (Q9). For collection and sharing, the provided response options allowed users to select whether a given policy explicitly stated that the website engaged in that practice ("Yes"), explicitly stated that it did not engage in that practice ("No"), whether it was "Unclear" if the website engaged in the practice, or if the data practice was "Not applicable" for the given policy. The sharing questions further distinguished sharing for the sole purpose of fulfilling a core service (e.g., payment processing or delivery), for purposes other than core services, or for purposes other than core services but only with explicit consent. The response options for the deletion question were "no removal," "full removal" (no data is retained), "partial removal" (some data may be retained), "unclear," and "not applicable." Users were instructed to ignore statements concerning retention for legal purposes, as our interest was in annotating retention practices that were questionably motivated but not legally obliged. For all nine questions, each response option was accompanied by an explanation to support its understanding. Throughout the questions, the "unclear" option allowed users to

Table 1. Privacy Policies from 26 Shopping and News Websites Were Annotated by Crowdworkers and Skilled Annotators to Assess Crowdworkers' Annotation Accuracy

sfgate.com	costco.com	accuweather.com
money.cnn.com	drudgereport.com	chron.com
bloomberg.com	tigerdirect.com	jcpenney.com
examiner.com	*hm.com*	*washingtonpost.com*
nike.com	ticketmaster.com	*wunderground.com*
abcnews.go.com	bodybuilding.com	*overstock.com*
time.com	*lowes.com*	*barnesandnoble.com*
zappos.com	*shutterfly.com*	latimes.com
bhphotovideo.com	*staples.com*	

The twelve policies in italics were used in the second experiment to evaluate the effectiveness of highlighting relevant paragraphs.

indicate when a policy was silent, ambiguous, or self-contradictory with regard to a specific data practice. See Appendix A for the full text of the annotation questions and their response options.

3.3 Analyzing Annotation Quality

We conducted a user study with the objective of determining to what extent it is possible to reliably crowdsource meaningful privacy policy annotations, specifically for the annotation scheme introduced in the previous section. To this end, we compared the annotations of our crowdworkers with those produced by our skilled annotators on a dataset of 26 privacy policies. Carnegie Mellon University's institutional review board approved our study.

3.3.1 Study Design. For our study, we selected the privacy policies of 26 news and shopping websites, listed in Table 1. They were selected based on traffic rankings from Alexa.com to provide a cross-section of frequently visited websites. All policies were collected in December 2013 and January 2014.

We recruited two participant groups for our study: *skilled annotators*, to obtain gold standard interpretations of privacy policies, and *crowdworkers* to evaluate the accuracy and utility of crowdsourcing privacy policy annotations. Both groups used the same online annotation tool.

The skilled annotators were five graduate students with a background in law and public policy, who concentrated on privacy research and were experienced in reading and interpreting privacy policies. They were recruited from Fordham University, Carnegie Mellon University, and the University of Pittsburgh. They were hired as research assistants for the duration of the annotation study. Three of them were female and two were male. They were 23 to 35 years old (median age: 24). Each of the five skilled annotators annotated all 26 policies by answering the nine questions, resulting in 1,170 question responses in total.

Crowdworkers were recruited on Amazon Mechanical Turk. Participants were required to be U.S. residents and to have at least a 95% approval rating for 500 completed tasks. Crowdworkers provided demographics information in an exit survey. Of the crowdworkers, 50.2% were male and 49.4% female (100); one crowdworker did not provide their gender.[2] They were 18 to 82 years old (median age: 32.5). The crowdworkers were somewhat less educated than the skilled annotators, all of whom had at least a college degree (bachelor's or higher). Of all the crowdworkers, 51.3% had

[2]The survey instrumentation only permitted two gender options and an option not to disclose gender, and we acknowledge that this set of answers was overconstrained.

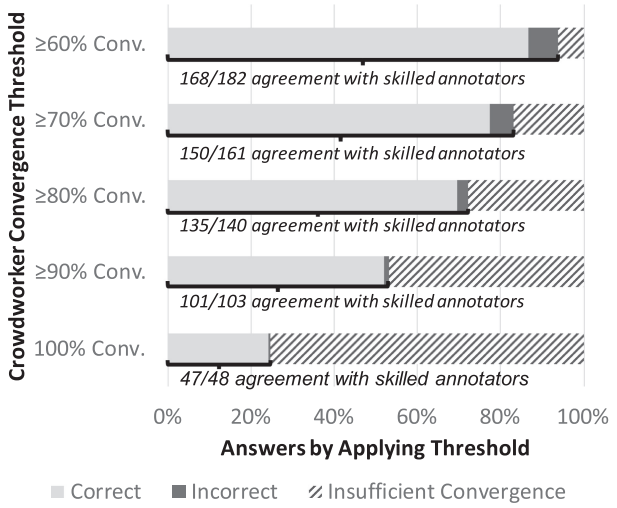

Fig. 2. Accuracy of annotations produced by 10 crowdworkers, as measured against skilled annotators, on a set of 194 policy-question pairs. Skilled annotators' answers were held to an 80% agreement threshold (i.e., at least four of five skilled annotators must agree on the same answer to each policy-question pair to merit its inclusion in the comparison). From top to bottom, the bars show crowdworkers' answers when held to a series of progressively higher agreement thresholds.

at least a college diploma, 47.2% had only a high school diploma, and 1.5% did not complete high school. Primary occupations of the crowdworkers were diverse. The most frequently named occupations were administrative support (12.7%); business, management, or financial (12.4%); computer engineering or information technology (10.6%); service industry (10.1%); student (8.7%); and unemployed (7.8%). The vast majority of crowdworker participants had no legal training (76.6%). Some (11.5%) indicated that their background provided them with some legal experience. 8.3% indicated they were knowledgeable in legal matters but had no formal legal training. Only 2.3% (5) studied law and 1.4% (3) received other legal or paralegal training. Crowdworkers with legal training were not excluded from participation because our goal was to assess how accurately crowdworkers as a population annotate privacy policies. This population happens to include a small percentage of legally trained persons.

Crowdworkers were paid USD$6 per annotated privacy policy, and each policy was annotated by 10 crowdworkers. The average time for task completion was 31 minutes for the skilled annotators[3] and 24 minutes for the crowdworkers. A total of 218 crowdworkers participated in our study and the vast majority (88.5%) annotated only one policy. We screened task submissions and checked whether question responses were accompanied by meaningful text selections. The rate of bogus answers was extremely low, perhaps due to the approval rating requirements and the relatively high pay.

3.3.2 Overall Accuracy Results. In Figure 2, we provide a high-level summary of the accuracy of crowdworker annotations as measured on the 26 privacy policies. In our analysis, we grouped "unclear" and "not addressed in the policy" annotations, since crowdworkers struggled to differentiate between these two options. To consolidate the five skilled annotators' responses, we held them to an 80% *agreement threshold*: for each policy-question pair, if at least four of the five skilled

[3]This average excludes six assignments with outlier durations greater than 10 hours, where we assume that the skilled annotators stepped away from the task for an extended period of time.

annotators agreed on an answer we considered it to be sufficiently confident for the evaluation standard. Otherwise, it was excluded from the comparison. We show results from consolidating crowdworkers' answers using agreement thresholds ranging from 60% to 100% at 10% intervals. Unsurprisingly, higher agreement thresholds yield progressively fewer answers. Crowdworkers' consolidated answers are deemed *correct* if they match the skilled annotators' consolidated answers and *incorrect* otherwise. All crowdworker agreement thresholds demonstrate strong accuracy when evaluated against skilled annotators' answers, with accuracies ranging from 92% (i.e., 168/182 at the 60% crowdworker agreement threshold) up to 98% (47/48 at the 100% crowdworker agreement threshold).

The 80% crowdworker agreement threshold (with 96% accuracy) seems to provide a reasonable balance between accuracy and coverage over the annotations available for analysis. We reached similar conclusions about the skilled annotator agreement threshold, and for the results in the remainder of this article, both agreement thresholds are set at 80%. This suggests that crowdsourcing produces meaningful privacy policy annotations, as they match the skilled annotators' interpretations with high accuracy if sufficiently high agreement thresholds are set. Most notably, given a sufficiently high agreement threshold ($\geq$80%), crowdworkers produce a very low number of false positives when they meet the threshold. This means that when crowdworker responses meet the agreement threshold, the response is, with high likelihood, consistent with an interpretation of the privacy policy by skilled annotators. If crowdworkers do not meet the agreement threshold, this also provides useful information about the privacy policy. It suggests that the privacy policy is sufficiently ambiguous to hinder consistent interpretation by an untrained population.

However, the fact that crowdworkers reach that agreement threshold and match the skilled annotators' interpretation for a large fraction of policy-question pairs should not be seen as an indication that privacy policies are clear. Instead, this reflects the fact that annotators were offered answer options that included "unclear" and "not addressed in the policy." For a number of policy-question pairs, skilled annotators and crowdworkers simply agreed with a high level of confidence that the policy was indeed unclear or that an issue was simply not addressed in the policy. Next, we take a detailed look at the results for each of the nine questions.

3.3.3 Question-Specific Results. Table 2 and Figure 3 provide detailed comparisons of answers from our skilled annotators and our crowdworkers, with both groups held to 80% agreement thresholds. Some questions appear to be substantially easier to answer than others; for example, our skilled annotators and the crowdworkers found it easy to answer questions about the collection of contact information. However, answering questions about the sharing of financial information or location information seems to be particularly difficult for crowdworkers who fail to meet the agreement threshold on 20 out of the 26 policies for each of those questions. It is worth noting that some questions seem to be challenging for skilled annotators as well. In particular, skilled annotators fail to converge on 15 of the 26 policy-question pairs dealing with the sharing of financial information. Overall, we observe that crowdworkers are able to converge on annotations in a majority of cases.

4 SUPPORTING CROWDWORKERS WITH RELEVANCE MODELING

4.1 Highlighting Paragraphs

Our results show that crowdworkers can provide highly accurate privacy policy annotations for most questions but that they struggle to converge on answers for questions pertaining to sharing practices, which are typically more spread out in the policy. An exacerbating factor is the length of privacy policies. Policies in our dataset contained 40.8 paragraphs on average, with a standard deviation of 15.8 paragraphs. To fully capture all aspects relating to an annotation question,

Table 2. Distributions of Skilled Annotations and Crowdsourced Annotations Collected for all Nine Questions Across all 26 Policies, Calculated with an 80% Agreement Threshold for Both Groups of Annotators

Question	Skilled Annotators			Crowdworkers		
	Yes	Unclear or N/A	No Conv.	Yes	Unclear or N/A	No Conv.
Collect Contact Info.	26			25		1
Collect Financial Info.	21	4	1	13	4	9
Collect Location Info.	10	12	4	14		12
Collect Health Info.	1	25		1	25	
Share Contact Info.	17		9	22		4
Share Financial Info.	7	4	15		6	20
Share Location Info.	1	19	6	2	4	20
Share Health Info.		25	1		24	2
Deletion of Info.	9	13	4	7	8	11
Total	92	102	40	84	71	79

"No conv." indicates a lack of sufficient agreement among the skilled annotators or crowdworkers. "Yes" indicates that the policy does allow the practice. "Unclear" indicates a "policy is unclear" annotation. Neither the skilled annotators nor the crowdworkers converged on a "no" answer (i.e., indicating that the policy does not allow the practice) for any of the policies.

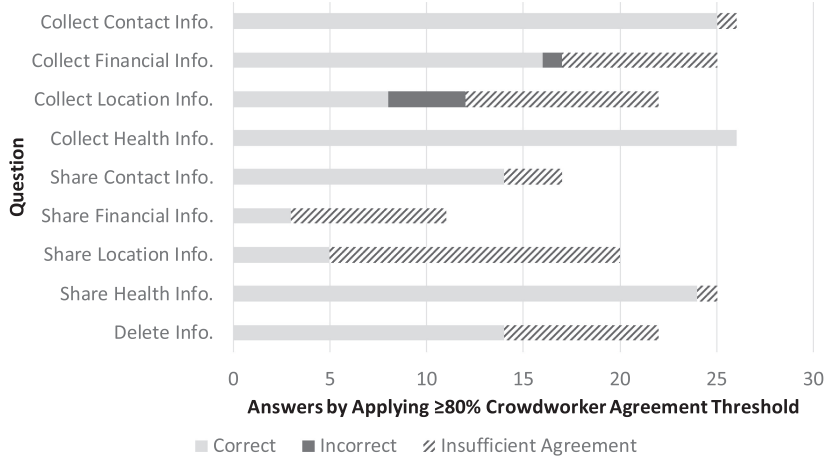

Fig. 3. Crowdworkers' annotation accuracy broken down by question. For the sake of this comparison, crowdworkers' answers and skilled annotators' answers were held to 80% agreement thresholds within their cohorts.

crowdworkers must read or at least skim the entire policy. This is both time-consuming and sub-optimally efficient, since they must read or skim many paragraphs multiple times as they answer multiple questions. Due to the length of policies, navigating them can be unwieldy, potentially causing a reader to miss relevant passages.

As noted before, splitting a policy into smaller parts could reduce reading time, but it bears the risk of losing context and the necessary holistic view on data practices. Instead, we propose a technique to identify and highlight paragraphs in the policy that are likely to be relevant to the given annotation question in order to support annotators in answering the respective question (Wilson

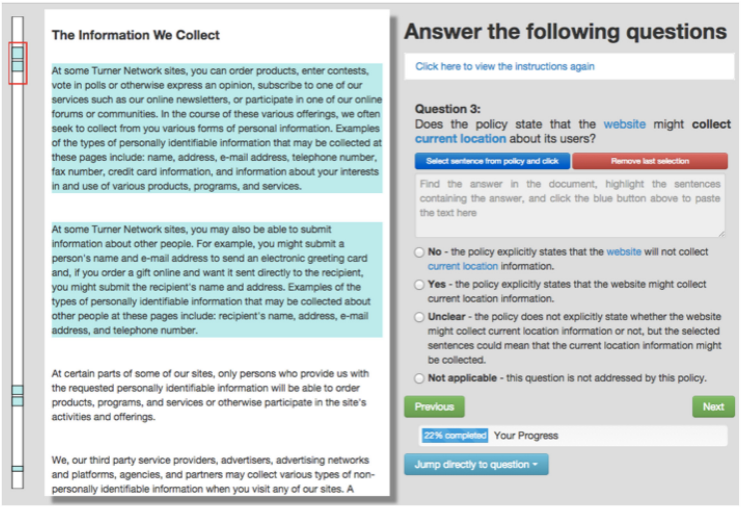

Fig. 4. Privacy policy annotation tool with paragraph highlighting. The paragraphs most relevant to the shown question are highlighted, and an overview bar (*left*) supports navigation between them. Rather than highlighting only the matched key phrases in the policy, we highlight entire paragraphs to reduce visual clutter and to encourage crowdworkers to read relevant context and thus gain a better understanding of the respective data practice.

et al. 2016c). A study evaluating the effects of highlighting paragraphs on annotation accuracy follows in Section 4.2.

4.1.1 Identifying Relevant Paragraphs. Our method predicts the top k paragraphs in a policy relevant to answering a given question. These relevant paragraphs are highlighted in the annotation tool, as shown in Figure 4, to provide annotators with cues about which parts of the policy they should focus on.

We created a separate classifier for each question and applied it to predict each paragraph's relevance to the question. Our approach involves developing regular expressions for a given data practice, which are then applied to a policy's paragraphs. The text selections provided by the skilled annotators were analyzed by a group of five law and privacy graduate students, who picked out phrases (4–10 words) that captured the essence of the response to a specific data practice question. For example, one phrase they chose was "*we obtain ... information we need to*" (the ellipsis being a placeholder for one or more words). These phrases were first normalized (for stemming and capitalization) and then converted into a list of 110 regular expressions, such as:

```
(place|view|use)(.*?)(tool to collect)(\w+){,3}(inform)
```

In this example, a word with the normalized form *place*, *view*, or *use* must occur in the same sentence as *tool to collect*, and a word with normalized form *inform* (e.g., *information*) must occur within three words of *collect*.

If a regular expression matched one or more paragraphs, those paragraphs were extracted for further feature engineering. After removing stopwords and stemming the selected paragraphs, we used normalized tf-idf values of lower order n-grams as features. Thus, for a paragraph, our feature set was comprised of two types of features: (1) *regex features*, i.e., a binary feature for every regular expression in the above constructed list; and (2) *n-gram features*, i.e., tf-idf values for uni-, bi-, and trigrams from the extracted paragraphs.

Based on the sentences selected by skilled annotators, we used the respective paragraphs as labels in supervised training. We trained nine classifiers—one for each question—using logistic regression. These classifiers predicted the probability that a given paragraph was relevant to the question for which it is trained. Logistic regression is a standard approach for combining a set of features that might correlate with each other to predict categorical variables. Additionally, it performs well with a low number of dimensions and when the predictors do not suffice to give more than a probabilistic estimate of the response.

Since we were working with a relatively small dataset, we used L_1 regularization to prevent the model from overfitting the data. We used five-fold cross-validation to select the regularization constant. If there are N paragraphs in the corpus, for each of the nine questions, we represent the i^{th} paragraph in the corpus as a feature vector (x_i). Depending on whether it was selected by the skilled annotator or not, we set the label (y_i) as 1 or 0, respectively. The parameters (θ) are learned by maximizing the regularized log likelihood:

$$l(\theta) = \sum_{i=1}^{N} y_i \log h_\theta(x_i) + (1 - y_i) \log(1 - h_\theta(x_i)) - \lambda \|\theta\|_1.$$

We then pick the top 5 or top 10 paragraphs ordered by probability to constitute the `TOP05` and `TOP10` relevant paragraph sets for a given policy-question pair.

4.1.2 Model Validation. To ensure that our model was indeed selecting relevant paragraphs, we calculated the recall of the `TOP05` and `TOP10` models against the paragraphs selected by the skilled annotators. Across all questions, the average recall rate was 0.91 with a standard deviation of 0.70 for `TOP05`, and it increased to 0.94 (standard deviation .07) for `TOP10`. We chose recall as an internal evaluation metric because our goal was to ensure that most of the relevant paragraphs for a question-policy pair were included in the highlights. Highlighting too few paragraphs may decrease annotation quality, as crowdworkers may ignore key text outside of the highlights. Thus, we preferred to potentially highlight some non-relevant paragraphs rather than omitting relevant ones.

4.2 Study: Effects of Highlighting

We integrated the relevance model into our privacy policy annotation tool by color-highlighting the top k-relevant paragraphs in each policy, as shown in Figure 4. We also added an overview bar to indicate which parts of the policy were highlighted. Annotators could click on the bar to directly jump to highlighted paragraphs or use buttons above the policy to navigate between highlighted paragraphs. We then conducted a between-subjects study on Mechanical Turk to evaluate the effects of highlighting on annotation accuracy as productivity (Wilson et al. 2016c). We found that highlighting relevant paragraphs can reduce task completion time without impacting annotation accuracy. Below, we describe our study design and results in detail.

4.2.1 Study Design. Our between-subjects study consisted of a control condition and two treatment conditions that highlighted different numbers of paragraphs (5 and 10), in order to investigate the effects of the number of highlights on annotation accuracy and productivity. We named these conditions as follows:

`NOHIGH`. This control condition consisted of the crowdworkers' responses for the 12 selected policies in the original privacy policy annotation task (cf. Figure 1). Crowdworkers were shown a privacy policy and asked to complete the nine annotation questions. No parts of the policy were highlighted.

Table 3. Demographics of Participants in the Highlighting Study

	Gender			Age		Education	
	Male	Female	Undisclosed	Range	Median	No College Degree	College Degree
NOHIGH	50.2%	49.4%	0.4%	18–82	32.5	48.7%	51.3%
TOP10	58.0%	42.0%	0%	19–68	30.9	42.9%	57.1%
TOP05	58.3%	41.0%	0.8%	20–65	31.4	45.8%	54.2%

TOP05. This condition was identical to NOHIGH, except that for each annotation question, the five most relevant paragraphs were highlighted (cf. Figure 4), based on our relevance model.

TOP10. This condition was identical to TOP05, except that the 10 paragraphs most relevant to the shown question were highlighted.

Crowdworkers were recruited on Mechanical Turk and randomly assigned to one of the treatments. If they had participated in the control, they were excluded from further participation, and we ensured that crowdworkers could not participate in more than one condition. In each condition, participants completed the privacy policy annotation task and a short exit survey that gathered user experience feedback and demographic information. We further asked participants to complete a Cloze test—an English proficiency test in which they had to fill in missing words in a short passage (University of Cambridge 2013, p. 14). Each participant annotated only one privacy policy, and we required 10 crowdworkers to annotate a given privacy policy. Participants were compensated with $6USD. They were required to be US residents with at least a 95% approval rating on 500 HITs. This study received IRB approval.

In order to balance overall annotation costs and annotation scale, we ran the study for a subset of 12 privacy policies randomly selected in equal parts from news and shopping websites. The 12 policies used in the highlighting study are marked in *italics* in Table 1. In total, we obtained annotations from 360 participants.

4.2.2 Results. We first discuss participant demographics followed by an analysis of the conditions' effect on productivity, accuracy, and usability.

Table 3 summarizes basic demographics for the three participant groups. The three groups exhibited similar characteristics in terms of gender, age, and education level.

Participants reported diverse occupations across all groups. Only 3.6% (NOHIGH), 1.6% (TOP10), and 5% (TOP05) of the crowdworkers reported to work in a position that required legal expertise. As Figure 5 shows, there was also little difference in terms of self-reported legal training between groups. Additionally, the fraction of correct answers in the English proficiency test were 0.55 (NOHIGH, SD=.23), 0.56 (TOP10, SD=.24), and 0.55 (TOP05, SD=.23), suggesting that English proficiency was comparable across groups.

A major concern with drawing annotators' attention to a subset of highlighted paragraphs is that it may negatively impact annotation accuracy, as annotators may miss relevant details in other parts of the policy due to over-reliance on the provided highlights. We evaluated the annotation accuracy of crowdworkers (≥80% agreement threshold) against the data previously collected from skilled annotators, focusing on those policy-question pairs from the 12 policies for which at least four of five skilled annotators agreed on the same interpretation (≥80% threshold). This was the case for 90 policy-question pairs.

Figure 6 shows the annotation accuracy for each condition. The annotation accuracy is similar across conditions: 98.4% for NOHIGH, 97.0% for TOP10, and 96.8% for TOP05. This suggests that highlighting relevant paragraphs does not affect annotation accuracy, especially not negatively. In the

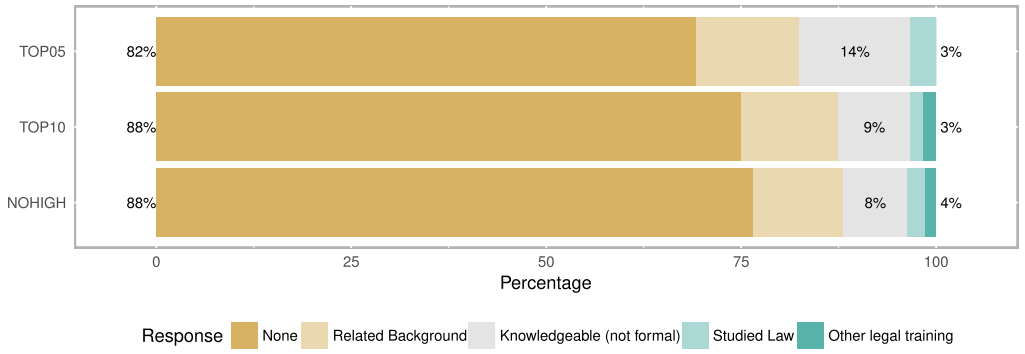

Fig. 5. Self-reported level of legal training.

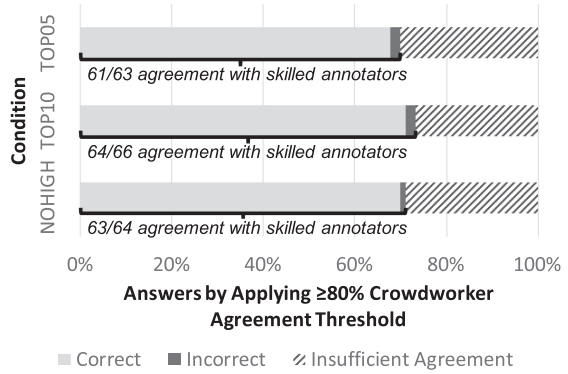

Fig. 6. Annotation accuracy by crowdworkers in the highlighting study, as measured against skilled annotators' answers (which were held to a ≥80% agreement threshold).

TOP10 condition, crowdworkers further reached 80% agreement for slightly more policy-question pairs. However, this effect is too small to be directly attributed to the highlighted paragraphs.

We further investigated if highlighting paragraphs affected the crowdworkers' text selections. The goal was to determine whether participants focused solely on the highlighted regions of text, ignoring the rest of the policy, or if they also considered potentially relevant information in non-highlighted parts of the policy. Almost all participants in the treatment conditions self-reported that they either "always read some of the non-highlighted text in addition to the highlighted sections before answering the question" (46.7% TOP05, 46.7% TOP10) or that they "read the non-highlighted text only when [they] did not find the answer within the highlighted text" (53.3% TOP05, 51.7% TOP10). Only 1.6% of participants in the TOP10 group and none in TOP05 reported that they "never read the non-highlighted text." Additionally, Figure 7 shows the percentage of selections from non-highlighted paragraphs in the policy for each of the nine annotation questions. For a substantial portion of questions, participants selected text from non-highlighted parts of the policy, which confirms that they did not solely focus on the highlights but also considered other policy parts when answering a question. The question-specific variations in Figure 7 suggest that some questions may benefit from the use of different machine learning methods, but highlighting relevant paragraphs does not seem to bias annotators to ignore non-highlighted parts of the policy.

However, while both groups selected text from non-highlighted parts for all questions of the policy, TOP05 participants tended to select more information from non-highlighted parts. This

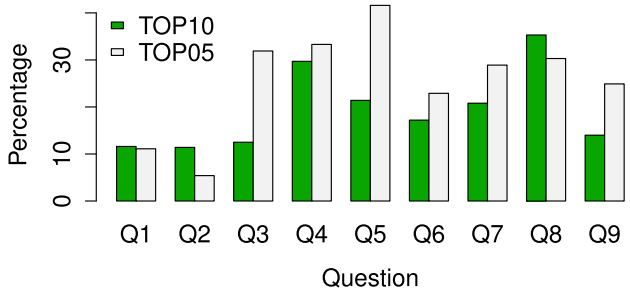

Fig. 7. Text selections from non-highlighted parts of a policy for each of the nine questions. Participants still consider other parts of the policy and do not only focus on highlighted paragraphs.

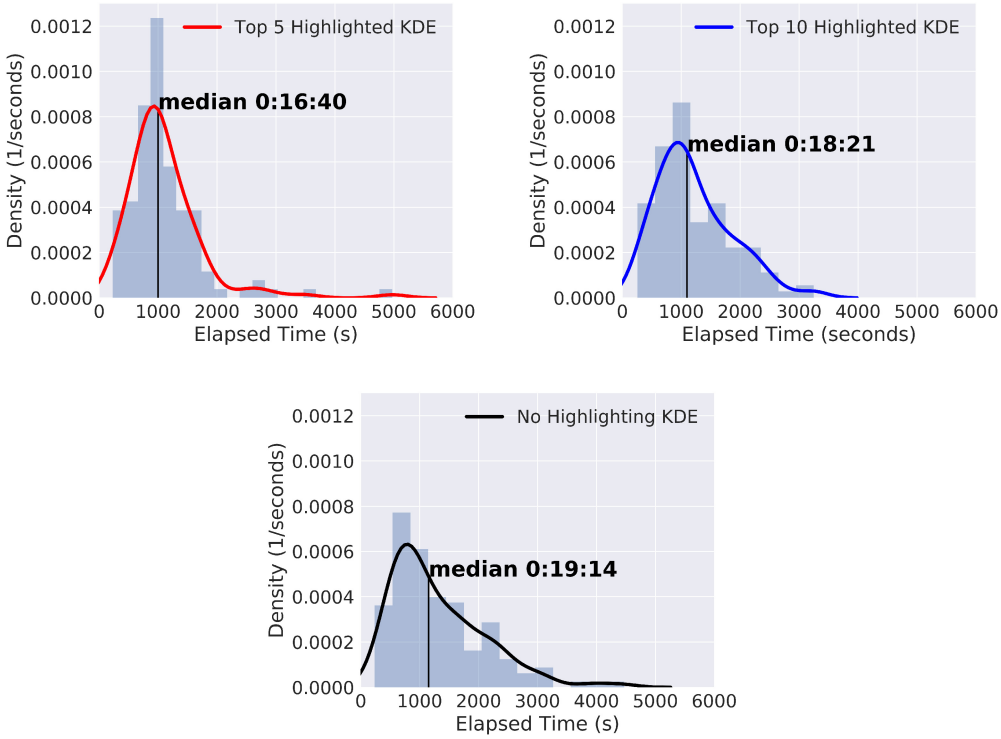

Fig. 8. Task completion time in the highlighting study. Highlighting the five most relevant paragraphs substantially reduces median task completion time.

suggests that, for some questions, more than five paragraphs need to be considered to fully grasp a data practice. We also observe differences for certain annotation questions and data practices. For instance, collection of financial (Q3) and health information (Q4) practices are often not as explicitly and concisely addressed as collection of contact (Q1) or location (Q2) information.

We further analyzed how highlighting paragraphs affected the crowdworkers' productivity in terms of task completion time, as shown in Figure 8. The median task completion times for the three conditions were 19 min 14 sec (`NOHIGH`), 18 min 21 sec (`TOP10`), and 16 min 40 sec (`TOP05`). Although these differences were not statistically significant (Kruskal-Wallis test), we observe that highlighting five paragraphs appeared to substantially reduce median task completion time by

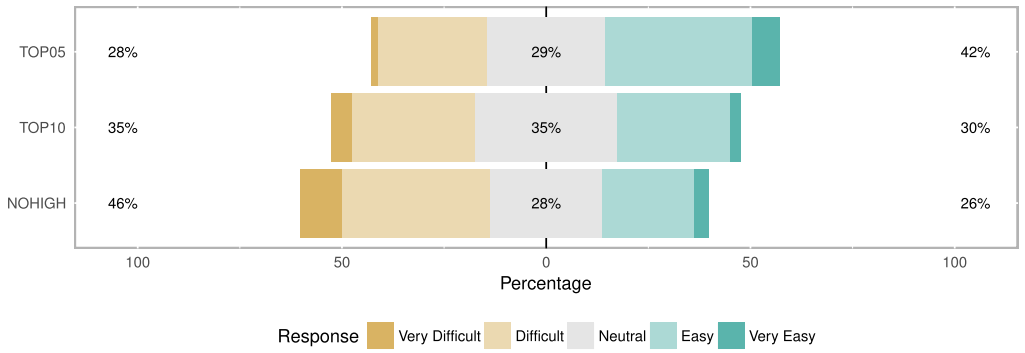

Fig. 9. Participants' responses to the question, *"How easy or difficult is it for you to understand legal texts?"*

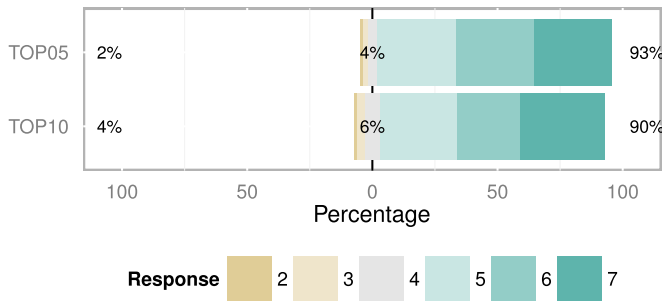

Fig. 10. Perceived usefulness of highlighting paragraphs in the treatment conditions, on a scale from (1) not at all helpful (not shown in the figure due to a lack of participants choosing this answer) to (7) very helpful.

more than 2 minutes without impacting annotation accuracy. Highlighting 10 paragraphs had a lesser effect on task completion time, suggesting that crowdworkers in the control condition may have read or skimmed a similar number of paragraphs.

When asked after the annotation task "*How easy or difficult is it for you to understand legal texts?*," responses for all three groups were normally distributed and centered on neutral, as shown in Figure 9. However, the treatment conditions rated their ability to understand legal text slightly higher compared to participants in the control group (`NOHIGH`). We also asked participants in the `TOP05` and `TOP10` groups to rate the perceived usefulness of paragraph highlighting on a seven-point scale ranging from *"Not at all helpful"* (1) to *"Very Helpful"* (7). Distribution of answer choices are shown in Figure 10. The median answer choice was Helpful (6) for both groups, signifying that the highlighted paragraphs were seen as useful cues and likely supported the annotators in determining the answer for a given data practice question.

Thus, we infer that paragraph highlighting in the annotation tool improved annotation productivity and user experience, which is an important factor for worker retention and cultivating a crowd of experienced annotators. Simultaneously, paragraph highlighting did not negatively impact annotation accuracy.

4.3 Discussion

The results presented in our crowdsourcing studies show promise for using crowdworkers to answer questions about privacy policies. It appears that data practices can be reliably extracted from privacy policies through crowdsourcing, and it is thus a viable mechanism to provide the data re-

quired for privacy policy analysis. This analysis can support new types of browser plug-ins and other user interfaces (e.g., personalized privacy interfaces (Das et al. 2018), interfaces emphasizing unexpected practices (Rao et al. 2016)) aimed at more effectively informing Internet users, who have generally given up on trying to read privacy policies. Furthermore, crowdsourcing could aid analysis to ease the work of regulators, who currently rely on manual inspection of privacy policies by experts in policy sweeps.

Our results further show that crowdsourcing privacy policy annotations is not trivial. We went through multiple iterations to refine our task design, as well as the annotation questions and response options. Given the vagueness of privacy policies, it was essential to provide crowdworkers with annotation options that indicate that a policy is unclear or does not address a given issue. Considering that even privacy experts may not always agree on the interpretation of policies (Reidenberg et al. 2015a), we cannot expect crowdworkers to perform better. From a public policy standpoint, these annotation options could also help identify egregious levels of ambiguity in privacy policies, either in response to particular types of questions or at the level of entire sectors. Finally, policy-question pairs where crowdworkers cannot converge could also be the basis for processes that engage website operators to clarify their data practices.

Although the 80% crowd agreement threshold appears promising, additional experiments with a larger number of policies will need to be conducted to further validate our results. An opportunity also exists for a user study to understand how to meet users' needs more precisely. Additional opportunities for refining this line of inquiry include allowing crowdworkers to rate the difficulty of answering a specific annotation question for a given policy. These ratings could then be considered in the aggregation of results. Such ratings, as well as the performance of individual crowdworkers, could also be used to develop more versatile crowdsourcing frameworks, where crowdworkers are directed to different annotation tasks based on their prior performance and the number of crowdworkers is dynamically adjusted. The longitudinal performance of crowdworkers could be monitored in order to place more weight on high-performing workers. These and similar approaches (Quoc Viet Hung et al. 2013) could be used to dynamically determine and allocate the number of annotations required for a question-policy pair. Additionally, the use of skilled workers on freelancing platforms such as Upwork may reduce the amount of redundancy necessary to reach answers with confidence.

Our research also shows that techniques that highlight paragraphs relevant to specific annotation questions can help increase productivity and may improve the user experience, as workers are provided with cues about which paragraphs they should focus on. This is important given the length of privacy policies and how some data practices are distributed in policy text. The number of highlighted paragraphs plays an essential role. In our study, highlighting the five most relevant paragraphs decreased task completion time, but also resulted in more text being selected from non-highlighted areas compared to highlighting 10 paragraphs. Ideally, we would want to highlight just enough for the annotator to clearly identify the answer. Thus, we are investigating approaches to dynamically adapt the number of highlights to question-specific parameters. For instance, some data practices such as collection of contact information are plainly stated in one part of the policy, while others require annotators to pay attention to multiple policy parts, such as third-party sharing practices. Our relevance models could be fine-tuned further and our approach could be extended to additional data practices to enable a progressive, larger-scale analysis.

5 AUTOMATED SEGMENT AND SENTENCE CLASSIFICATION

In the previous section, we showed how using relevance models to highlight paragraphs improves a crowdsourcing task to answer questions about privacy policies. This motivates us to fully automate the procedure of labeling privacy policy segments with information pertinent to their legal

contents. In this section, we present advances in identifying policy text segments and individual sentences that correspond to expert-identified categories of policy contents.

5.1 The OPP-115 Corpus

For a source of labeled data, we use the Usable Privacy Policy Project's OPP-115 Corpus (Wilson et al. 2016b), which contains detailed annotations for the data practices described in a set of 115 website privacy policies. Viewed at a coarse-grained level, annotations fall into one of ten data practice *categories*, which were developed by a team of privacy and legal experts:

(1) *First-Party Collection/Use*: How and why a service provider collects user information.
(2) *Third-Party Sharing/Collection*: How user information may be shared with or collected by third parties.
(3) *User Choice/Control*: Choices and control options available to users.
(4) *User Access, Edit, and Deletion*: If and how users can access, edit, or delete their information.
(5) *Data Retention*: How long user information is stored.
(6) *Data Security*: How user information is protected.
(7) *Policy Change*: If and how users will be informed about changes to the privacy policy.
(8) *Do Not Track*: If and how Do Not Track signals for online tracking and advertising (see Doty et al. [2016]) are honored.
(9) *International & Specific Audiences*: Practices that pertain only to a specific group of users (e.g., children, residents of the European Union, or Californians).
(10) *Other*: Additional privacy-related information not covered by the above categories.[4]

Privacy policies were divided into *segments*, which were units of text roughly equivalent to paragraphs. Segment boundaries were determined by combining an automated procedure that used features from the text (e.g., punctuation and sentence boundaries) with manual error-checking, which ensured that segment boundaries did not bisect sentences and discouraged the creation of extremely long or short segments. Annotators identified spans of text associated with data practices inside of each segment. Each privacy policy was annotated by three law students, who required a mean time of 72 minutes per document. In aggregate, they produced a total of 23,194 data practice annotations.

We proceed with the observation that the text associated with each category has a distinct vocabulary, even though many of the categories represent closely related themes. Preliminarily, we used weights from logistic regression to identify particularly relevant words for the categories. Table 4 shows the results. The top six words or collocations for each category illustrate the category's topical focus.

5.2 Privacy Policy Text Classification

Below, we describe our procedure for labeling privacy policy text at the sentence and segment levels. Different levels of granularity produce different results on the number of tokens annotated, which would affect reading time if the classification results were used in downstream tasks such as highlighting paragraphs to help human annotators.

First, we explain how we transform the annotations into labels for segments and sentences. Annotations for data practices inside a segment can be effectively "elevated" to cover the entire segment, i.e., a segment receives a binary label for the presence or absence of each data practice category. Wilson et al. (2016b) calculated inter-annotator agreement (Cohen's κ) for segment-level

[4]Because of its indistinct nature, we omit this category from further discussion.

Table 4. Vocabulary for Each Category from Logistic Regression. Words and Collocations Are Sorted in Descending Order from Left to Right According to Their Weights

Category	Vocabularies
First-Party Collection/Use	use, collect, demographic, address, survey, service
Third-Party Sharing/Collection	party, share, sell, disclose, company, advertiser
User Choice/Control	opt, unsubscribe, disable, choose, choice, consent
User Access, Edit, and Deletion	delete, profile, correct, account, change, update
Data Retention	retain, store, delete, deletion, database, participate
Data Security	secure, security, seal, safeguard, protect, ensure
Policy Change	change, change privacy, policy time, current, policy agreement
Do Not Track	signal, track, track request, respond, browser, advertising for
International & Specific Audiences	child, California, resident, European, age, parent

labels to be 0.76 for the first two categories listed above, which comprised 61% of all data practices in the corpus, and found a variety of lower and higher κ-values for the remaining categories. For our present work, we use segment-level labels produced by a simple majority vote: if two annotators agree that a segment contains at least one data practice in a given category, then we apply that category to the segment as a label. We use a similar method to produce sentence-level labels: if at least two annotators label any part of a sentence with a given category, we label the sentence with that category. Note that the labels are not mutually exclusive, and a segment or sentence may be labeled with zero categories or any combination of them.

Second, we explain the methods we used for classifying privacy policy text on the sentence or segment level. For our experiment, we split the 115 policies of the OPP-115 corpus into 80% training and 20% testing sets. Since each segment or sentence can contain information for multiple categories, we built binary classifiers for each category with three models, respectively, logistic regression, support vector machines, and convolutional neural networks (CNNs) (Kim 2014). We used a bigram term frequency—-inverse document frequency (tf–idf) for logistic regression and support vector machines. The parameters for each model are tuned with five-fold cross validation. The parameters for the CNN follow Kim's (2014) CNN-non-static model, which uses pre-trained word vectors. We used 20% of the training set as a held-out development set to refine these models.

5.3 Results and Discussion

The results of segment- and sentence-level classification are shown in Table 5. At the segment level, we observe a mean (across all categories) precision of 0.80, recall of 0.77, and micro-F_1 of 0.78 with SVMs using tf-idf features, which outperforms previous results using word-embeddings as features (Wilson et al. 2016b). The sentence level results are around 0.1–0.12 below the segment level results. One potential explanation for this is that although annotators have access to the context that surrounds a sentence (e.g., prior and subsequent sentences), our models do not. We also observe that the CNN model favors precision while the other two models favor recall. This difference can be taken into consideration for downstream tasks with different objectives (e.g., governmental regulators might be interested in manually verifying results; hence, not missing instances is more important than the false positive rate).

All three models show similar performances after careful parameter tuning, which motivates us to look at the data in more detail to find reasons for errors. For example, the corpus does not contain many privacy policies of health care providers. One provider's policy is quoted in Figure 11 showing health-specific language, more of which would encourage improved performance. During our evaluation, we recognized that our classifiers' performances are also impacted by the context

Table 5. Classification Results (Precision/Recall/F_1 Score) for Sentences and Segments Using Logistic Regression (LR), Support Vector Machines (SVM), and Convolutional Neural Networks (CNN)

Category	Sentence			Segment		
	LR	SVM	CNN	LR	SVM	CNN
First-Party Collection/Use	.62 /.76/.69	.64/.71/.67	.78/ .58/.66	.83/.76/.79	.84/.77/.81	.87/.70/.78
Third-Party Sharing/Collection	.57/.73/.64	.61/.72 /.66	.86/.40/.55	.71/.85/.77	.74/.81/.78	.79/.80/.79
User Choice/Control	.45/.72/.55	.42/.71 /.53	.57/.33/.42	.75/.62/.68	.70/.69/.70	.78/.56/.65
User Access, Edit, & Deletion	.57/.66/.61	.65/.52/.58	.93/.22/.36	.83/.78/.81	.77/.89/.82	.93/.68/.78
Data Retention	.68/.40/.51	.70/.31/.43	.75/.23/.35	.59/.33/.43	.80/.27/.40	0.0/0.0/0.0
Data Security	.62/.74/.67	.60 /.71/.65	.67/.71/.69	.67/.79/.73	.70/.85/.77	.77/.85/.80
Policy Change	.66/.80/.72	.75/.78/.77	.86/.65/.74	.95/.74/.83	.95/.67/.78	1.0/.74/.85
Do Not Track	.71/.77/.74	.69/.69 /.69	.83/.38/.53	1.0/1.0/1.0	1.0/1.0/1.0	1.0/1.0/1.0
International & Specific Audiences	.75/.74/.74	.75/.75/.75	.77/.69/.73	.72/.86/.79	.88/.82/.85	.79/.84/.81
Micro–Average	.61/.73/.66	.63/.70/.66	.78/.51/.60	.77/.76/.76	.80/.77/.78	.80/.71/.75

> *As Kaleida Health is a teaching facility, we may disclose your health information for training and educational purposes to faculty physicians, residents and medical, dental, nursing, pharmacy or other students in health-related professions from local colleges or universities affiliated with Kaleida Health.*

Fig. 11. Example of a classification error: our models failed to detect the Third-Party Sharing/Collection category for this text fragment.

or lack thereof during the production of the annotations. For example, section headings were only shown to annotators for the segments that immediately followed them, but segments were presented for annotation in order. Features that encode context around each segment or sentence should be investigated to avoid this problem.

Overall, these results indicate the strength of these methods toward enabling downstream tasks, such as filtering for more detailed data practices, extracting salient details to present to users (e.g., (Das et al. 2018; Rao et al. 2016)), or summarization of privacy practices.

6 AUTOMATICALLY EXTRACTING PRIVACY CHOICES

6.1 Choice Instances

Although Internet users are concerned about their privacy and would like to be informed about the privacy controls they can exercise, they are not willing or able to find these choices in policy text (Reidenberg et al. 2015a). Choices for privacy controls, which are the most actionable pieces of information in these documents, are frequently "hidden in plain sight" among other information. However, the nature of the text and the vocabulary used to present choices provide us with an opportunity to automatically identify choices.

We define a *choice instance* as a statement in a privacy policy that indicates that the user has discretion over aspects of their privacy. An example (which notably features a hyperlink) is the following:

> If you would like more information on how to opt out of information collection practices by many third parties, visit the Digital Advertising Alliance's website at www.aboutads.info.[5]

[5]http://www.nurse.com/privacy/. Retrieved March 12, 2018.

Some examples of choices offered to users include opt-outs or controls for the sharing of personal information with third parties, receiving targeted ads, or receiving promotional emails. Analyzing these choice instances in aggregate will help to understand how notice and choice is implemented in practice, which is of interest to legal scholars, policy makers and regulators. Furthermore, extracted choice options can be presented to users in more concise and usable notice formats (Schaub et al. 2015), such as a browser plug-in or a privacy-based question answering system.

We also used the OPP-115 Corpus (Wilson et al. 2016b) to train and evaluate our models for identifying opt-out choices. In the OPP-115 Corpus, attributes representing choice instances are present in multiple categories of data practices, namely "First-Party Collection/Use," "Third-Party Sharing/Collection," "User Access, Edit, and Deletion," "Policy Change," and "User Choice/Control." The dataset contains annotations for different types of user choice instances, namely "opt-in," "opt-out," "opt-out link," "opt-out via contacting company," "deactivate account," "delete account (full)," and "delete account (partial)."

We treat the identification of choice instances as a binary sentence classification problem, in which we label each sentence in the privacy policy text as containing a choice instance or not, based on the presence of text spans highlighted by the annotators. We focus on extracting hyperlinks indicating opt-out choices (coarse-grained classification) and further devise methods to classify these hyperlinks based on the type of opt-out (fine-grained classification). Using the coarse- and fine-grained classification models, we develop a composite two-tier classification model to identify opt-out choices along with their types (Sathyendra et al. 2017b).

6.2 Coarse-Grained Classification

We divided the dataset into training and testing sets of 85 and 30 privacy policies, respectively. We experimented with a variety of features for *coarse-grained classification*, to separate choice instances from negative instances:

- **Stemmed Unigrams and Bigrams.** We removed most stop words from the feature set, although some were retained for the modal verb and opt out features (described below). Bigrams are important to capture pertinent phrases such as *opt out*.
- **Relative Location in the Document.** This was a ratio between the number of sentences appearing before the sentence instance and the total number of sentences in the privacy policy.
- **Topic Model Features.** We represented the OPP-115 segment (roughly, a paragraph) containing the sentence instance as a topic distribution vector using latent Dirichlet allocation (Blei et al. 2003) and non-negative matrix factorization (Xu et al. 2003) with 8 and 10 topics, respectively. Previous work on vocabulary intersections of expert annotations and topic models for data practices in privacy policies (Liu et al. 2016b) inspired us to take this approach.
- **Modal Verbs and Opt-Out Specific Phrases.** We observed vocabulary cues in positive instances that suggested a domain-independent "vocabulary of choice." Many positive instances were imperative sentences and contained modal words such as *may*, *might*, or *can*. We also identified key phrases in the training set such as *unsubscribe* and *opt-out* that were indicative of opt-out choices.
- **Syntactic Parse Tree Features.** We obtained constituency parse trees for sentences using the Stanford Parser (Manning et al. 2014) and extracted production rules and non-terminals as features. We also included the maximum depth and average depth of the parse tree as features, as these are indications of specificity.

Table 6. Distribution of Different Annotation Types

Annotation		
Party offering choice	**Purpose**	**# Instances**
Third Party (TH)	Advertising (AD)	52
First Party (FI)	Communications (CM)	19
First Party (FI)	Advertising (AD)	15
First Party (FI)	Data Sharing (DS)	6
Third Party (TH)	Analytics (AN)	4
Browser (BR)	Cookies (CK)	2
Third Party (TH)	Data Sharing (DS)	2
First Party (FI)	Cookies (CK)	1
Third Party (TH)	Cookies (CK)	1

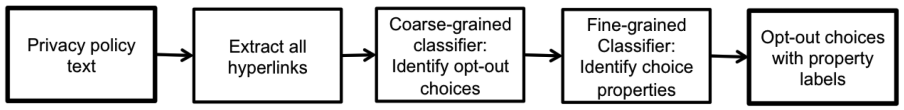

Fig. 12. Two-tier classification model for opt-out choice extraction.

We used logistic regression classification for the coarse-grained classification stage. Model hyper-parameters were tuned based on five-fold cross validation on the training set. The final parameters for the best performing model had the inverse L_2 regularization constant set at $C = 1.3$ and class-weights of 1.5 and 1 for positive and negative class, respectively.

6.3 Fine-Grained Classification

We also developed a *fine-grained* model to differentiate between varieties of opt-out instances. For training data, we annotated a set of 125 positive instances to assign two additional labels to each of them; these were *Party Offering Choice* and *Purpose*. Party Offering Choice could be one of *First Party* (FI), *Third Party*, (TH), or *Browser* (BR). Purpose could be one of *Advertising* (AD), *Data Sharing* (DS), *Communications* (CM), *Analytics* (AN), or *Cookies* (CK). Table 6 shows the distribution of these annotations. To predict these labels, we trained eight binary logistic regression classifiers, one for each of the preceding values. If multiple classifiers in a label set returned positive, we selected the prediction with the higher log likelihood. The features we used for these classifiers were:

- **Stemmed Unigrams and Bigrams.** We collected bags of n-grams from the sentence under consideration and its containing segment.
- **Anchor Text.** The anchor text of the hyperlink in the sentence.
- **Hyperlink URL Tokens.** We split the URL by punctuation (such as "/" and ".") and extracted tokens.
- **Privacy Policy URL Tokens.** We also extracted tokens from the policy URL as features.
- **URL Similarity Measure.** We calculated the Jaccard index between the vocabulary of the policy URL and the hyperlink URL. This feature is used to identify whether the hyperlink was to a first-party page or a third-party page.

Figure 12 illustrates the overall architecture of our system. We first use the coarse-grained step to identify the presence of an opt-out instance, and then use the fine-grained step to ascertain key properties of an opt-out choice if one is present.

Table 7. Results of Ablation Tests for the Coarse-grained Classifier

Features/Models	Precision	Recall	F_1
All	**0.862**	**0.641**	**0.735**
All Minus Unigrams	0.731	0.487	0.585
All Minus Bigrams	0.885	0.590	0.708
All Minus Rel. Location	0.889	0.615	0.727
All Minus Topic Models	0.852	0.590	0.697
All Minus Productions	0.957	0.564	0.710
All Minus Nonterminals	0.913	0.538	0.677
All Minus Max. Depth	0.857	0.615	0.716
All Minus Avg. Depth	0.857	0.615	0.716
Phrase Inclusion—Baseline	0.425	0.797	0.554
Paragraph Vec.—50 Dimensions	0.667	0.211	0.320
Paragraph Vec.—100 Dimensions	0.667	0.158	0.255

6.4 Results and Discussion

For the coarse-grained task, we consider a simple baseline that labels sentences as positive if they contain one or more opt-out specific words, which come from a vocabulary set that we identified by examining positive instances in the training set. The F_1 of the baseline was 0.554.

We performed ablation tests excluding one feature at a time from the coarse-grained classifier. The results of these tests are presented in Table 7 as precision, recall, and F_1 scores for the positive class, i.e., the opt-out class. Using the F_1 scores as the primary evaluation metric, it appears that all features help in classification. The unigram, topic distribution, nonterminal, and modal verb and opt-out phrase features contribute the most to performance, including all the features results in an F_1 score of 0.735. Ablation test without unigram features resulted in the lowest F_1 score of 0.585, and by analyzing features with higher logistic regression weights, we found n-grams such as *unsubscribe* to have intuitively high weights. We also found the syntactic parse tree feature *S→SBAR VP* to have a high weight, indicating that the presence of subordinate clauses (SBARs) helps in classification.

For an additional practical evaluation, we created a second dataset of sentences from the privacy policies of the 180 most popular websites (as determined by Alexa rankings[6]). We selected only those sentences that contained hyperlinks, since they are associated with particularly actionable choices in privacy policy text. We used our model (as trained on the OPP-115 Corpus) to label the 3,842 sentences in this set, and then manually verified the 124 positive predictions, observing perfect precision. Although we were unable to measure recall using this method, the high precision suggests the robustness of the model and the practical applicability of this approach to tools for Internet users.

The results for the opt-out type classification are shown in Table 8. Because of data sparsity, we show performance figures for only the top two most frequent label combinations. These results also demonstrate a practical level of performance for Internet user-oriented tools.

7 FUTURE RESEARCH CHALLENGES

Our work demonstrates the feasibility of automated and semi-automated analysis of privacy policies, but more work remains to fully bridge the gap between these documents and what Internet

[6] http://www.alexa.com/topsites. Retrieved December 2013.

Table 8. Fine-grained Classifier Results

	Precision	Recall	F_1
First-Party Communications (FI, CM)	0.947	0.947	0.947
Third-Party Advertising (TH, AD)	0.905	0.977	0.940

users understand about them (Wilson et al. 2016a). To this end, we challenge the research community to investigate a family of problems related to the analysis of privacy policies. These problems are well-motivated by established topics in natural language processing as well as the difficulties of the "notice and choice" model of online privacy in its current form. Solving them will constitute progress toward helping Internet users understand how their personal information is used and what choices they can make about that usage. Additionally, policy regulators and creators will have tools to help monitor compliance with laws and detect trends that require action.

A central challenge of this research direction is the need to annotate privacy policies in a scalable, cost-efficient manner. We have already observed how machine learning can be used to guide human annotators' efforts; for example, the automatically-generated paragraph highlights made the crowdsourcing task easier for workers. We have also demonstrated how policy segments can be classified into categories and how user choices can be identified. These are steps toward a goal of limiting the need for human annotators to small, self-contained tasks that are optimal for crowdsourcing while natural language processing and machine learning take care of the bulk of the analysis. An ambitious (but not completely unreasonable) goal will be to eliminate the need for human annotators altogether. By producing well-calibrated confidence ratings alongside data practice predictions, an automated system could account for its shortcomings by stating which predictions are very likely to be correct and deferring to crowdworkers for predictions that lack firmness.

Finally, related problems for consideration include:

— *Consolidation of annotations from multiple workers*: Under what criteria do a pair of non-identical data practices produced by two annotators refer to the same underlying data practice in the text? Criteria may be observable (i.e., present in the practices' attributes or text spans) or latent (depending on factors such as policy ambiguity or vagueness, which may cause two annotations of a data practice to be divergent without either being in error).
— *Recombination of data practices into a cohesive body of knowledge about a privacy policy*: How do data practices for a privacy policy relate to each other? The answer to this is not contained chiefly in the annotations. For example, two data practices may appear to contradict each other even though they do not, because the reconciliation cannot be represented by the annotation scheme, and thus it is absent from the annotations. Inconsistencies, generalizations, and implications are other examples of potential relationships between data practices. Adding expressiveness to an annotation scheme comes with the tradeoff of greater complexity.
— *Summarization and simplification*: Can the text of a privacy policy be shortened or reworded so that the average Internet user can understand it? A simple test for content equivalence is whether an annotation procedure (by humans or automated methods) produces the same set of data practices for the simplified text and the original text. In practice, Internet users have already demonstrated limited patience with text-based privacy policies, but this problem is nevertheless motivated by the broader goal of making complex texts easier to understand.
— *Optimizing the balance between human and automated methods for privacy policy annotation*: Human annotators and automated annotation both have strengths and weaknesses.

The ideal combination in an annotation system will depend on the necessary level of confidence in the annotations and the availability of resources. These resources include human annotators, computational power, and training data to create computational models.

— *Identifying sectoral norms and outliers*: Within a sector (e.g., websites for financial services or news), how can we identify typical and atypical practices? A bank website that collects users' health information, for example, would be atypical. When an atypical practice is found, when should it be a cause for concern (or commendation)? Can we recommend websites in a given sector based on an Internet user's expressed privacy preferences?

— *Identifying trends in privacy practices*: The activities that Internet users perform online continue to evolve, and with that evolution, the mechanisms for collecting, using, and sharing their data are subject to change. The Internet of Things (IoT) provides a potent example, as sensors collect and share progressively larger amounts of sensitive data. Finding trends in privacy practices will help guide policy regulators to focus their attention on emerging issues.

8 CONCLUSIONS

We have demonstrated results for a set of tasks that automate the analysis of privacy policies, to assist human readers and to extract pertinent details for them. Our results show that, collectively, crowdworkers can understand privacy policies sufficiently well to answer questions about their contents, and that crowdworkers can also be helped using relevance models that highlight text likely to contain the answer to each question. Moving away from human effort and toward more detailed annotations, we have also shown how privacy policy text can be categorized on a paragraph or sentence basis, and choices embedded in the text can be automatically identified. We forsee this trajectory of automation continuing in future efforts, which will support the development of user-centric tools for understanding and more effectively communicating websites' and apps' privacy practices and choices. Part of our ongoing work in this area includes the development of question answering functionality intended to answer users' privacy queries (Sathyendra et al. 2017a). It also includes work on personalized privacy assistants capable of personalizing privacy notices based on models of what their users care to be notified about (Das et al. 2018).

APPENDIX

A ANNOTATION QUESTIONS

Questions Q1 through Q4 address the collection of contact information, financial information, current location information, and health information, respectively. Their wording is largely identical, and for brevity, only Q1 and its answers are shown below.

Q1. Does the policy state that the website might **collect contact information** about its users?

— **No**; the policy explicitly states that the website will not collect contact information.

— **Yes**; the policy explicitly states that the website might collect contact information.

— **Unclear**; the policy does not explicitly state whether the website might collect contact information or not, but the selected sentences could mean that contact information might be collected.

— **Not applicable**; this question is not addressed by this policy.

Questions Q5 through Q8 address the sharing of contact information, financial information, current location information, and health information, respectively. Their wording is largely identical, and for brevity, only Q5 and its answers are shown below.

Q5. Does the policy state that the website might **share contact information** with **third parties**? Please select the option that best describes how contact information is shared with third parties. Please ignore any sharing required by law (e.g., with law enforcement agencies).
- —**No sharing**; the policy explicitly states that the website will not share contact information with third parties.
- —**Sharing for core service only**; the policy explicitly states that the website might share contact information with third parties, but only for the purpose of providing a core service, either with explicit or implied consent/permission from the user.
- —**Sharing for other purpose**; the policy explicitly states that the website might share contact information with third parties for other purposes. The policy makes no statement about the user's consent/permission or user consent is implied.
- —**Sharing for other purpose (explicit consent)**; the policy explicitly states that the website might share contact information with third parties for a purpose that is not a core service, but only if the user provided explicit permission/consent to do so.
- —**Unclear**
- —**Not applicable**

Finally, Q9 addresses deletion of personal data.

Q9. What is the website's policy about letting its users **delete their personal data**? Please ignore any statements concerning retention for legal purposes.
- —**No removal**; the policy explicitly states that the user will not be allowed to delete their personal data.
- —**Full removal**; the policy explicitly states that users may delete their personal data and that no data will be retained for any purpose, whether the data was provided directly by the user, generated by the user's activities on the website, or acquired from third parties.
- —**Partial removal**; the policy explicitly states that users may delete their personal data, but some/all of the data might be retained for other purposes, whether the data was provided directly by the user, generated by the user's activities on the website or acquired from third-parties.
- —**Unclear**
- —**Not applicable**

REFERENCES

Mark S. Ackerman, Lorrie Faith Cranor, and Joseph Reagle. 1999. Privacy in e-commerce: Examining user scenarios and privacy preferences. In *Proceedings of the 1st ACM Conference on Electronic Commerce (EC'99)*. ACM, New York, NY, 1–8. DOI: https://doi.org/10.1145/336992.336995 00456.

Waleed Ammar, Shomir Wilson, Norman Sadeh, and Noah A. Smith. 2012. Automatic categorization of privacy policies: A pilot study. Technical Report. Carnegie Mellon University.

Paul André, Aniket Kittur, and Steven P. Dow. 2014. Crowd synthesis: Extracting categories and clusters from complex data. In *Proc. CSCW'14*. ACM, 989–998.

Ngo Xuan Bach, Nguyen Le Minh, Tran Thi Oanh, and Akira Shimazu. 2013. A two-phase framework for learning logical structures of paragraphs in legal articles. *ACM Transactions on Asian Language Information Processing (TALIP)* 12, 1 (2013), 3.

Jaspreet Bhatia, Travis D. Breaux, Joel R. Reidenberg, and Thomas B. Norton. 2016b. A theory of vagueness and privacy risk perception. In *Proceedings of the 2016 IEEE 24th International Requirements Engineering Conference (RE)*. 26–35. DOI: https://doi.org/10.1109/RE.2016.20

Jaspreet Bhatia, Travis D. Breaux, and Florian Schaub. 2016a. Mining privacy goals from privacy policies using hybridized task recomposition. *ACM Transactions on Software Engineering and Methodology* 25, 3, Article 22 (May 2016), 24 pages. DOI: https://doi.org/10.1145/2907942

David M. Blei, Andrew Y. Ng, and Michael I. Jordan. 2003. Latent dirichlet allocation. *Journal of Machine Learning Research* 3, Jan (2003), 993–1022.

Travis D. Breaux and Florian Schaub. 2014. Scaling requirements extraction to the crowd. In *Proceedings of the 22nd IEEE International Requirements Engineering Conference (RE'14)*. IEEE Society Press, Washington, D.C.

Fred H. Cate. 2010. The limits of notice and choice. *IEEE Security & Privacy* 8, 2 (2010), 59–62.

Lydia B. Chilton, Greg Little, Darren Edge, Daniel S. Weld, and James A. Landay. 2013. Cascade: Crowdsourcing taxonomy creation. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM, 1999–2008.

Parvathi Chundi and Pranav M. Subramaniam. 2014. An approach to analyze web privacy policy documents. In *KDD Workshop on Data Mining for Social Good*.

Elisa Costante, Jerry den Hartog, and Milan Petković. 2013. What websites know about you: Privacy policy analysis using information extraction. In *Data Privacy Management and Autonomous Spontaneous Security (Lecture Notes in Computer Science)*, Roberto Di Pietro, Javier Herranz, Ernesto Damiani, and Radu State (Eds.), Vol. 7731. Springer, 146–159.

Lorrie Cranor, B. Dobbs, S. Egelman, G. Hogben, J. Humphrey, M. Langheinrich, M. Marchiori, M. Presler-Marshall, J. Reagle, D. A. Stampley, Matthias Schunter, and Rigo Wenning. 2006. *The Platform for Privacy Preferences 1.1 (P3P1.1) Specification*. Working Group Note. W3C. Retrieved March 12, 2018 from http://www.w3.org/TR/P3P11/.

Lorrie Faith Cranor. 2012. Necessary but not sufficient: Standardized mechanisms for privacy notice and choice. *J. on Telecomm. & High Tech. L.* 10 (2012), 273.

Lorrie Faith Cranor, Pedro Giovanni Leon, and Blase Ur. 2016. A large-scale evaluation of U.S. financial institutions' standardized privacy notices. *ACM Transactions on the Web* 10, 3, Article 17 (Aug. 2016), 33 pages. DOI: https://doi.org/10.1145/2911988

Anupam Das, Martin Degeling, Daniel Smullen, and Norman Sadeh. 2018. Personalized privacy assistants for the Internet of Things. *IEEE Pervasive Computing—Special Issue on Securing the IoT*. 17, 3 (2018), 35–46. DOI: http://dx.doi.org/10.1109/MPRV.2018.03367733

Nick Doty, Heather West, Justin Brookman, Sean Harvey, and Erica Newland. 2016. Tracking compliance and scope. *Candidate Recommendation. W3C*.

Tatiana Ermakova, Benjamin Fabian, and Eleonora Babina. 2015. Readability of privacy policies of healthcare websites. In *12. Internationale Tagung Wirtschaftsinformatik (Wirtschaftsinformatik 2015)*.

Morgan C. Evans, Jaspreet Bhatia, Sudarshan Wadkar, and Travis D. Breaux. 2017. An evaluation of constituency-based hyponymy extraction from privacy policies. In *Proceedings of the 2017 IEEE 25th International Requirements Engineering Conference (RE)*. 312–321. DOI: https://doi.org/10.1109/RE.2017.87

Federal Trade Commission. 2000. *Privacy Online: A Report to Congress*. Technical Report. Federal Trade Commission.

Federal Trade Commission. 2012. Protecting Consumer Privacy in an Era of Rapid Change: Recommendations For Businesses and Policymakers. Retrieved March 12, 2018 from https://www.ftc.gov/reports/protecting-consumer-privacy-era-rapid-change-recommendations-businesses-policymakers.

Filippo Galgani, Paul Compton, and Achim Hoffmann. 2012. Combining different summarization techniques for legal text. In *Proceedings of the Workshop on Innovative Hybrid Approaches to the Processing of Textual Data*. Association for Computational Linguistics, 115–123.

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*.

Carlos Jensen and Colin Potts. 2004. Privacy policies as decision-making tools: An evaluation of online privacy notices. In *Proc. CHI'04*. ACM.

Adam N. Joinson, Ulf-Dietrich Reips, Tom Buchanan, and Carina B. Paine Schofield. 2010. Privacy, trust, and self-disclosure online. *Human-Computer Interaction* 25, 1 (Feb. 2010), 1–24. DOI: https://doi.org/10.1080/07370020903586662

Yoon Kim. 2014. Convolutional neural networks for sentence classification. *arXiv Preprint arXiv:1408.5882* (2014).

Aniket Kittur, Boris Smus, Susheel Khamkar, and Robert E. Kraut. 2011. Crowdforge: Crowdsourcing complex work. In *Proc. UIST'11*. ACM, 43–52.

Pedro Giovanni Leon, Blase Ur, Yang Wang, Manya Sleeper, Rebecca Balebako, Richard Shay, Lujo Bauer, Mihai Christodorescu, and Lorrie Faith Cranor. 2013. What matters to users?: Factors that affect users' willingness to share information with online advertisers. In *Proc. SOUPS'13*. ACM. DOI: https://doi.org/10.1145/2501604.2501611

Fei Liu, Nicole Lee Fella, and Kexin Liao. 2016a. Modeling language vagueness in privacy policies using deep neural networks. In *Proceedings of the 2016 AAAI Fall Symposium Series*.

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 (COLING)*.

Frederick Liu, Shomir Wilson, Florian Schaub, and Norman Sadeh. 2016b. Analyzing vocabulary intersections of expert annotations and topic models for data practices in privacy policies. In *Proceedings of the 2016 AAAI Fall Symposium Series*.

Ewa Luger, Stuart Moran, and Tom Rodden. 2013. Consent for all: Revealing the hidden complexity of terms and conditions. In *Proc. CHI'13*. ACM.

Lars Mahler. 2015. What Is NLP and Why Should Lawyers Care? Retrieved March 12, 2018 from http://www.lawpracticetoday.org/article/nlp-lawyers/.

Christopher D. Manning, Mihai Surdeanu, John Bauer, Jenny Finkel, Steven J. Bethard, and David McClosky. 2014. The Stanford CoreNLP natural language processing toolkit. In *Association for Computational Linguistics (ACL) System Demonstrations*. 55–60.

Aleecia M. McDonald. 2013. *Browser Wars: A New Sequel?* The Technology of Privacy. Silicon Flatirons Center, University of Colorado. Presented Jan. 11, 2013.

Aleecia M. McDonald and Lorrie Faith Cranor. 2008. The cost of reading privacy policies. *I/S: Journal of Law and Policy for the Information Society* 4, 3 (2008), 540–561.

Gabriele Meiselwitz. 2013. Readability assessment of policies and procedures of social networking sites. In *Proceedings of OCSC'13*.

Simonetta Montemagni, Wim Peters, and Daniela Tiscornia. 2010. *Semantic Processing of Legal Texts*. Springer.

Matteo Negri, Luisa Bentivogli, Yashar Mehdad, Danilo Giampiccolo, and Alessandro Marchetti. 2011. Divide and conquer: Crowdsourcing the creation of cross-lingual textual entailment corpora. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 670–679.

Official California Legislative Information. 2003. Online Privacy Protection Act of 2003. Retrieved March 12, 2018 from http://leginfo.legislature.ca.gov/faces/billTextClient.xhtml?bill_id=200320040AB68.

Alessandro Oltramari, Dhivya Piraviperumal, Florian Schaub, Shomir Wilson, Sushain Cherivirala, Thomas B. Norton, N. Cameron Russell, Peter Story, Joel Reidenberg, and Norman Sadeh. 2017. PrivOnto: A semantic framework for the analysis of privacy policies. *Semantic Web Journal* Preprint (2017), 1–19.

Alexander J. Quinn and Benjamin B. Bederson. 2011. Human computation: A survey and taxonomy of a growing field. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI'11)*. ACM, New York, NY, 1403–1412. DOI: https://doi.org/10.1145/1978942.1979148 00257.

Nguyen Quoc Viet Hung, Nguyen Thanh Tam, LamNgoc Tran, and Karl Aberer. 2013. An evaluation of aggregation techniques in crowdsourcing. In *Proceedings of WISE'13*. Springer, 1–15. DOI: https://doi.org/10.1007/978-3-642-41154-0_1

Rohan Ramanath, Fei Liu, Norman Sadeh, and Noah A. Smith. 2014. Unsupervised alignment of privacy policies using hidden Markov models. In *Proceedings of the Annual Meeting of the Association of Computational Linguistics (ACL'14)*. ACL, 605–610.

A. Rao, F. Schaub, N. Sadeh, A. Acquisti, and R. Kang. 2016. Expecting the unexpected: Understanding mismatched privacy expectations online. In *Proceedings of the 12th Symposium on Usable Privacy and Security (SOUPS'16)*. USENIX Association, 77–96. DOI: https://www.usenix.org/system/files/conference/soups2016/soups2016-paper-rao.pdf.

Joel R. Reidenberg, Jaspreet Bhatia, Travis Breaux, and Thomas B. Norton. 2016. Automated comparisons of ambiguity in privacy policies and the impact of regulation. *Journal of Legal Studies* 45, 2 (15 Mar 2016), S163–S190.

Joel R. Reidenberg, Travis Breaux, Lorrie Faith Cranor, Brian French, Amanda Grannis, James T. Graves, Fei Liu, Aleecia McDonald, Thomas B. Norton, Rohan Ramanath, N. Cameron Russell, Norman Sadeh, and Florian Schaub. 2015a. Disagreeable privacy policies: Mismatches between meaning and users' understanding. *Berkeley Tech. LJ* 30 (2015), 39.

Joel R. Reidenberg, N. Cameron Russell, Alexander J. Callen, Sophia Qasir, and Thomas B. Norton. 2015b. Privacy harms and the effectiveness of the notice and choice framework. *I/S: Journal of Law & Policy for the Information Society* 11 (2015).

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, and Shomir Wilson. 2013. *The Usable Privacy Policy Project: Combining Crowdsourcing, Machine Learning and Natural Language Processing to Semi-Automatically Answer Those Privacy Questions Users Care About*. Tech. report CMU-ISR-13-119. Carnegie Mellon University.

K.M. Sathyendra, A. Ravichander, P. Story, A.W. Black, and N. Sadeh. 2017a. *Helping Users Understand Privacy Notices with Automated Question Answering Functionality: An Exploratory Study*. Tech. Report CMU-LTI-17-005. Carnegie Mellon University.

Kanthashree Sathyendra, Shomir Wilson, Florian Schaub, Sebastian Zimmeck, and Norman Sadeh. 2017b. Identifying the provision of choices in privacy policy text. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 2774–2779.

Florian Schaub, Rebecca Balebako, and Lorrie Faith Cranor. 2017. Designing effective privacy notices and controls. *IEEE Internet Computing* 21, 3 (May 2017), 70–77. DOI: https://doi.org/10.1109/MIC.2017.75

Florian Schaub, Rebecca Balebako, Adam L. Durity, and Lorrie Faith Cranor. 2015. A design space for effective privacy notices. In *Proceedings of the 11th Symposium On Usable Privacy and Security (SOUPS 2015)*. USENIX Association, Ottawa, 1–17.

Florian Schaub, Travis D. Breaux, and Norman Sadeh. 2016. Crowdsourcing privacy policy analysis: Potential, challenges and best practices. *it–Information Technology* 58, 5 (2016), 229–236.

Rocky Slavin, Xiaoyin Wang, Mitra Bokaei Hosseini, James Hester, Ram Krishnan, Jaspreet Bhatia, Travis D. Breaux, and Jianwei Niu. 2016. Toward a framework for detecting privacy policy violations in android application code. In *Proceedings of the 38th International Conference on Software Engineering (ICSE'16)*. ACM, New York, NY, 25–36. DOI: https://doi.org/10.1145/2884781.2884855

John W. Stamey and Ryan A. Rossi. 2009. Automatically identifying relations in privacy policies. In *Proc. SIGDOC'09*. ACM.

Tos;DR. 2012. Terms of Service Didn't Read. http://tosdr.org/. Retrieved March 12, 2018.

University of Cambridge. 2013. Certificate of Proficiency in English (CPE), CEFR Level C2): Handbook for Teachers.

Shomir Wilson, Florian Schaub, Aswarth Dara, Sushain K. Cherivirala, Sebastian Zimmeck, Mads Schaarup Andersen, Pedro Giovanni Leon, Eduard Hovy, and Norman Sadeh. 2016a. Demystifying privacy policies with language technologies: Progress and challenges. In *Proceedings of LREC 1st Workshop on Text Analytics for Cybersecurity and Online Safety (TA-COS'16)*. ELRA, Portorož, Slovenia.

Shomir Wilson, Florian Schaub, Aswarth Dara, Frederick Liu, Sushain Cherivirala, Pedro Giovanni Leon, Mads Schaarup Andersen, Sebastian Zimmeck, Kanthashree Sathyendra, N. Cameron Russell, Thomas B. Norton, Eduard Hovy, Joel R. Reidenberg, and Norman Sadeh. 2016b. The creation and analysis of a website privacy policy corpus. In *Annual Meeting of the Association for Computational Linguistics, Aug 2016*. ACL.

Shomir Wilson, Florian Schaub, Rohan Ramanath, Norman Sadeh, Fei Liu, Noah A. Smith, and Frederick Liu. 2016c. Crowdsourcing annotations for websites' privacy policies: Can it really work? In *Proceedings of the 25th International World Wide Web Conference*. International World Wide Web Conferences Steering Committee, 133–143.

Wei Xu, Xin Liu, and Yihong Gong. 2003. Document clustering based on non-negative matrix factorization. In *Proceedings of the 26th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. ACM, 267–273.

Sebastian Zimmeck and Steven M. Bellovin. 2014. Privee: An architecture for automatically analyzing web privacy policies. In *Proceedings of the USENIX Security Symposium*.

Sebastian Zimmeck, Ziqi Wang, Lieyong Zou, Roger Iyengar, Bin Liu, Florian Schaub, Shormir Wilson, Norman Sadeh, Steven M. Bellovin, and Joel Reidenberg. 2017. Automated analysis of privacy requirements for mobile apps. In *Proceedings of the 24th Network & Distributed System Security Symposium (NDSS'17)*. Internet Society, San Diego, CA.

Received June 2017; revised June 2018; accepted August 2018