Automated Text Mining for Requirements Analysis of Policy Documents

Aaron K. Massey*, Jacob Eisenstein*, Annie I. Antón*, Peter P. Swire†
*School of Interactive Computing, Georgia Institute of Technology, Atlanta, GA, USA
{akmassey, jacobe, aianton}@gatech.edu
†Moritz College of Law, Ohio State University, Columbus, OH, USA
{swire.1}@osu.edu

***Abstract*—Businesses and organizations in jurisdictions around the world are required by law to provide their customers and users with information about their business practices in the form of policy documents. Requirements engineers analyze these documents as sources of requirements, but this analysis is a time-consuming and mostly manual process. Moreover, policy documents contain legalese and present readability challenges to requirements engineers seeking to analyze them. In this paper, we perform a large-scale analysis of 2,061 policy documents, including policy documents from the Google Top 1000 most visited websites and the Fortune 500 companies, for three purposes: (1) to assess the readability of these policy documents for requirements engineers; (2) to determine if automated text mining can indicate whether a policy document contains requirements expressed as either privacy protections or vulnerabilities; and (3) to establish the generalizability of prior work in the identification of privacy protections and vulnerabilities from privacy policies to other policy documents. Our results suggest that this requirements analysis technique, developed on a small set of policy documents in two domains, may generalize to other domains.**

I. Introduction

Information and information systems are increasingly important to modern economies and societies. Marc Andreessen believes that "software is eating the world" because of the importance of information systems to the economy and the transition of traditional businesses to software-based organizations [1]. Andreesson cites leading companies in traditional sectors such as books (Amazon), movies (Netflix, Pixar), music (Apple, Spotify, Pandora), marketing (Google), communications (Skype), and recruiting (LinkedIn) as examples of the new information economy [1]. In 2010, Walmart processed more than 1 million customer transactions every hour [2]. The massive amounts of data generated by information systems led Bruce Schneier to say that "data is the pollution of the information age" because many processes create it and it stays around [3]. Just as regulators have sought to protect the environment from pollution, organizations such as the U.S. Federal Trade Commission (FTC) are now seeking to protect consumers from information pollution.

In the United States, the FTC has statutory powers to ensure that the data created or collected by organizations is managed in ways that protect consumers.[1] Although no blanket federal law in the U.S. requires organizations to post such a notice, some states, including California,[2] passed state laws requiring organizations to post privacy policies, and virtually all organizations operating in the United States, above a minimal size, post notices of their privacy practices. Other commonly posted notices of information practices include Terms of Use, Terms and Conditions, and Terms of Service documents. These policy documents have recently become the subject of requirements engineering analysis [4]–[8].

To protect consumers from unfair or deceptive trade practices, the FTC has the power to hold organizations accountable for their public statements. If an organization promises a practice publicly but does not live up to that promise, then the FTC can impose significant penalties. Google,[3] Facebook,[4] Microsoft,[5] and Twitter[6] have all agreed to consent decrees with the FTC based on violations of their own consumer privacy promises. Google's settlement included a record $22.5 million civil penalty, the largest penalty the FTC has ever levied against a single company, for misrepresenting its tracking policy for users of Apple's Safari web browser.

Organizations seeking to comply with FTC recommendations and California state law must ensure their software systems support the promises made in their policy documents. The FTC needs to be able to easily and quickly assess the requirements and obligations stated in numerous policy documents spanning entire industries. Individual consumers may be interested in comparing the policy documents from a small number of organizations to determine which most closely matches their own preferences. Finally, requirements engineers must ensure that organizations comply with these policies in their software systems, and meet the expectations of both the FTC and individual consumers.

In this paper, we perform a large-scale analysis of 2,061 policy documents. This corpus includes policy documents from two requirements engineering studies [4], [6], the Google Top 1000 most visited websites, and the Fortune 500 companies. We perform this examination for three purposes: (1) to assess

[1]Federal Trade Commission Act of 1914, 15 U.S.C §§ 41-58

[2]California state law is particularly influential because of the large number of high tech and tech-savvy businesses seeking to operate in the state. The specific law passed was the Online Privacy Protection Act of 2003, Cal. Bus. & Prof. Code §§ 22575-22579

[3]http://ftc.gov/opa/2012/08/google.shtm

[4]http://ftc.gov/opa/2012/08/facebook.shtm

[5]http://www.ftc.gov/opa/2002/08/microsoft.shtm

[6]http://www.ftc.gov/opa/2011/03/twitter.shtm

978-1-4673-5765-4/13 © 2013 IEEE

Accepted for publication by IEEE. © 2013 IEEE. Personal use of this material is permitted. Permission from IEEE must be obtained for all other uses, in any current or future media, including reprinting/republishing this material for advertising or promotional purposes, creating new collective works, for resale or redistribution to servers or lists, or reuse of any copyrighted component of this work in other works.

the readability of these policy documents for requirements engineers; (2) to determine if automated text mining can indicate whether a policy document contains requirements expressed as either privacy protections or vulnerabilities; and (3) to establish the generalizability of prior work [5] in the identification of privacy protections and vulnerabilities from privacy policies to other policy documents. Our results suggest that this requirements analysis technique, developed on a small set of policy documents in two domains, may generalize to other domains.

The remainder of this paper is organized as follows: Section II defines the research questions that guide this work. Section III discusses the research background upon which our work is based. We discuss how we collected and analyzed our policy document sets in Section IV. Section V details the methodology we used to address our research questions. Our results are described in Section VI. We discuss threats to the validity of this work in Section VII and provide a summary of this paper in Section VIII.

II. Research Questions

Regulators, consumers, and requirements engineers share an interest in the content of policy documents, but there are few tools to assist in their analysis of these documents. Herein, we present a methodology for analyzing the requirements specified in 2,061 policy documents. In particular, we address three research questions:

- RQ1: How similar, with respect to readability, are policy documents of different types, organizations, and industries?
- RQ2: Can automated text mining help requirements engineers determine whether a policy document contains requirements expressed as either privacy protections and vulnerabilities?
- RQ3: Can topic modeling be used to confirm the generalizability of the Antón-Earp privacy protections and vulnerabilities taxonomy [5]?

The first research question addresses whether policy documents are similarly challenging to read and understand. If so, then requirements engineers, regulators, and consumers need support in understanding these critical pieces of the information economy. The second research question addresses concerns shared by regulators and individuals: do policies contain stated protections and implicit vulnerabilities? Both regulators and concerned individuals need tools to perform this assessment on many policy documents. In addition, requirements engineers need to be able to mitigate vulnerabilities and verify protections. The third research questions addresses the regulatory and consumer desire to broadly and consistently evaluate the goals built into the software systems that manage their personal information. The first step towards engineering systems that respect consumer protection goals is to understand current industry practices and the requirements they entail. This is particularly important to technologists tasked with regulating whole industries [9]. The Antón-Earp taxonomy [5] was developed on privacy policies in two domains, but preliminary results suggest that it may provide requirements guidance more broadly. However, additional work is needed to further verify this.

III. Related Work

Several areas of related work serve as relevant background for the analyses described in this paper. In particular, we discuss the use of policy documents in requirements engineering, large-scale analyses of privacy policies, and the use of natural language processing techniques in requirements engineering.

A. Policy Documents in Requirements Engineering

Requirements compliance with policy documents is a growing field of interest to the research community. Antón and Earp developed an integreated strategy that focuses on the initial specification of security and privacy policies and their operationalization into policy-compliant system requirements [10]. Additionally, they developed techniques for early conflict identification as well as prevention of incongruous behavior, misalignments, and unfulfilled requirements, so that security and privacy are built in rather than added on as an afterthought [11].

Breaux et al. employ an approach called semantic parameterization to derive semantic models from goals mined from privacy policy documents [12]. Allison et al. model privacy policy elements for information systems in Service-Oriented Architectures seeking to comply with the FIPPs [13]. Robinson developed a framework called REQMON that monitors requirements compliance with policy documents at runtime [14]. These researchers focused on the specification and management of requirements from policy documents, but they are not suited to large-scale policy analysis as we attempt in this paper.

Young et al. use a theory of commitments, privileges, and rights to identify software requirements based on the commitments that organizations express in their policy documents (e.g. privacy notices, terms of use, etc.). Their main objective is to ensure the software requirements comply with an organization's commitments and business practices [7], [8]. Similarly, we seek to aid requirements engineers in honoring the commitments expressed in organizational policy documents.

Otto and Antón found both the identification of relevant laws and regulations as well as the difficulty of navigating and searching laws and regulations as important challenges for requirements engineering in a legal domain [15]. Herein, we provide initial research that addresses both concerns for requirements engineers concerned with policy compliance.

B. Large-Scale Analysis of Privacy Documents

The readability of privacy policies has been studied extensively. In 2004, Antón et al. examined 40 financial privacy policies from nine different institutions for clarity, readability, and compliance with relevant laws and regulations [4]. They found the average readability of these policies to be higher

than the average education level of the United States population. In addition, using Goal-Based Requirements Analysis, they found compliance with relevant laws and regulations to be questionable at best. In a similar study, Antón et al. examined 24 policy documents from healthcare institutions both before and after passage of the U.S. Health Insurance Portability and Accountability Act (HIPAA) [6].[7] They found similarly poor readability results. Herein, we include these policy documents in our text mining analysis to determine whether they exhibit themes found in other policy documents.

Several researchers studied methods for improving comprehension of policy documents. Vail et al. compared four techniques for presenting privacy policies and found that users believe natural language policies to be more secure than other representations [16]. However, they also found that users comprehend natural language policies less than other representations when asked to answer questions about the policy document's content [16]. McDonald et al. surveyed over 700 individuals on their understanding of policy documents from six companies in three different formats [17]. With a layered policy, individuals sacrificed accurate understanding of policy content for speed when compared to natural language [17]. More importantly, they found that participants disliked each format similarly [17]. These studies found policy documents to be challenging for average individuals to read and understand, but neither examined more than a couple dozen policies. Also, these studies focused on the end-user experience, whereas our study seeks to aid requirements engineers and regulators.

McDonald and Cranor demonstrated that it would take the average individual around 244 hours per year (roughly six full 40-hour work weeks) to read and understand the privacy policies for every company with which they interacted [18]. Their estimate is based on multiple factors including a study of the privacy policies from 75 of the most popular websites [18]. Clearly, the amount of information is overwhelming, and the cost of obtaining that information is non-trivial. Even if time were not a concern and individuals understood the content of privacy policies, Acquisti and Grossklags believe the amount of information necessary would make completely rational decision making impossible [19]. Regulators, such as the FTC, must also cope with the complexity and immensity of these policy documents. Herein, we propose the use of text mining techniques to dramatically reduce the scope and cost of obtaining actionable information from many policy documents for both requirements engineers and regulators.

C. Text Mining in Requirements Engineering

Topic modeling is a text mining technique that can discover the themes in massive document collections [20], [21]. Topic modeling allows individuals to determine their particular interest first and then analyze a document collection to determine what it may say about that interest [20], [21]. The topics identified represent latent themes; specifically, they are probability distributions of the words in the vocabulary defined by the document collection. The only inputs required are the number of topics and the text of the documents. No manual annotation or analysis is needed to perform topic modeling, but it is not intended to eliminate the need for examination by individuals. This is consistent with Ryan's assertion that natural language processing cannot replace the role of engineers in the specification and validation of requirements [22]. Rather, topic modeling augments and amplifies individual analysis [23]. Topic models have been applied to many domains, including political science [24].

The particular topic modeling approach we apply in this paper is Latent Dirichlet Allocation (LDA) with variational estimation [25]. LDA is a statistical model of a document collection that attempts to reveal a document's underlying structure in the form of the topics it discusses [25]. For example, a policy document may contain descriptions of services available, information storage and access, privacy protections, privacy vulnerabilities, and other themes or topics. LDA formally defines a topic as a distribution of words over a fixed vocabulary, which captures the intuition that some words in a document refer to a particular topic in that document while other words refer to other topics in the same document [25]. The key distinguishing characteristic of LDA is that the model assumes that all documents in the collection are about the same set of topics, but that each document individually may vary in the amount it discusses those topics [25]. For example, all policy documents are about the various policies an organization upholds, but a particular privacy policy may focus on privacy protections to a greater extent than other policy documents, including another privacy policy. Indeed, Antón and Earp previously found this to be true [5].

Herein, we apply topic models to privacy policies for two purposes. First, we seek to determine if topic models can reveal whether a policy document contains software requirements artifacts. Second, we seek to determine if topic models can confirm the validity of the Antón-Earp taxonomy [5] across multiple domains.

IV. Data Sets and Collection

To perform our analysis, we first needed to collect several large sets of policy documents to examine. The vast majority of websites do not follow a standard protocol for disseminating their policy documents. The Platform for Privacy Preferences (P3P)[8] is a standard for machine-accessible privacy notices, but it has not been widely adopted [26], [27]. This lack of consistency makes collection of large sets of policy documents a time-intensive and laborious process.

We selected our data from three basic sources:

1) **Our prior requirements engineering work** in policy document analysis that demonstrates the use of both goal-based requirements analysis and commitment-based requirements analysis for policy documents [4], [5], [7], [8].

[7] Pub.L. 104-191, 110 Stat. 1936

[8] http://www.w3.org/P3P/

2) **The Google Top 1000 websites** are the most-visited sites on the Internet based on Google's estimates of Internet traffic.[9]
3) **The Fortune 500** companies for 2012 are the largest companies in the United States sorted by revenues.[10]

Our first source of policy documents came from our prior analysis on financial privacy policies and on the effects of HIPAA on healthcare privacy notices [5], [6]. We chose these policies because the previous analyses included both an evaluation of readability and a goal-based requirements analysis resulting in the development of a validated taxonomy of privacy goals and vulnerabilities.

For our second source of policy documents, we employed the Google Top 1000 websites according to Google's estimates of total Internet traffic. We employed this set because online policy documents from these sites cover numerous popular destinations on the Internet receiving billions of users per year and directly applying to nearly 80% of North Americans [28]. Of the original 1000 sites, 58 targeted non-English-speaking populations. These policies were not included, but because some sites had more than one policy document, we collected 1,063 total policy documents for the set. All of these policies were effective in early December 2012.

For our third set of policy documents, we selected all policies of the companies listed in the Fortune 500. These companies represent many of the most influential and powerful organizations in the United States, and thus, their stated policies are critically important for American consumers. Of the original 500 companies, all had a homepage targeted towards English-speaking populations, and as before some had more than one policy document linked on their homepage, so we collected a total of 891 policy documents for the set. All of these policies were effective in early January 2013.

For each data set, we visited the homepage for each website and manually collected any policy documents for each organization. These policy documents included Privacy Policies, Privacy Notices, Terms of Use, Terms of Service, Terms and Conditions, and similarly titled documents. We only collected policy documents that were linked directly on the homepage of the organizations. Essentially, we collected any formal statement of policies pertaining to an Internet user's use of their web-based services. We then saved the full HTML served to us and manually excerpted[11] only the relevant policy document as plain text. For tables, we manually copied and pasted the text row by row to create a coherent plain text representation.

[9]More information about the Google Top 1000 can be found here: http://www.google.com/adplanner/static/top1000/

[10]More information about the Fortune 500 can be found here: http://money.cnn.com/magazines/fortune/fortune500/2012/full_list/

[11]Extracting plain text from html may be automated in the future, but current approaches popularized by tools like Instapaper (http://www.instapaper.com) and Readability (http://readability.com) do not handle policy documents reliably.

V. Methodology

Our analysis methodology consists of three steps: (1) readability analysis of policy documents, (2) building and validating a topic model of the policies, and (3) exploring privacy protection goals and vulnerabilities using the topic model.

A. Readability of Policy Documents

Policy documents are notoriously difficult to understand [4], [18]. However, to our knowledge, the largest readability studies of policy documents consist of fewer than 100 policies. Our own prior research was conducted on documents collected in 2003 and again in 2007 [4]–[6]. Since that time, the FTC conducted several major investigations of deceptive practices in policy documents resulting in settlements with Google, Facebook, Twitter, and Microsoft among other companies. Herein, we perform a comprehensive readability analysis of 2,060 policy documents. If these policy documents are all similarly challenging to read, then development of requirements analysis techniques is justified.

We measure readability of policy documents using five metrics: Flesch Reading Ease [29], Flesch Grade Level [29], FOG [30], SMOG [31], and the Automated Readability Index (ARI) [32], [33]. The Flesch Reading Ease (FRE) metric produces a score from 0 to 100, with 0 representing a challenging document to read and 100 representing a document that is easy to read [29]. The Flesch Grade Level (FGL) metric computes the estimated years of education needed to be capable of reading and understanding the meaning of a given document [29]. Both measures have been used in our prior studies [4]–[6]. Three newer metrics: the FOG [30], SMOG [31], and Automated Readability Index (ARI) [32], [33] were all developed in part to address different aspects of readability than the original Flesch metrics. All five metrics account for differences in document length by normalizing based on the number of sentences. These metrics begin to become less accurate for documents containing fewer than 100 words [34], but only one policy document[12] of the 2,060 included in our analysis contained fewer than 100 words. We removed this document from our analysis.

We chose these metrics for three reasons. First, they are commonly used metrics for assessing the readability of privacy policies [4]–[6], [17], [18]. Second, they do not use a language ontology or language model,[13] which matches our approach to topic modeling as described in Section V-B. Third, they are established readability metrics for regulatory scenarios. In particular, the Automated Readability Index was developed for use with technical materials and has been employed by the United States Navy [32]. Section VI-A presents our readability study results.

[12]The Morgan Stanley Privacy Pledge from our first study.

[13]More recent readability measures that use language models include the new Dale-Chall formula [35] and unigram language models [36], [37].

B. A Topic Model for Policy Documents

Probabilistic topic models are designed to uncover the hidden themes in large document collections that would otherwise be impossible to analyze through human annotation [20]. They have been successfully employed in many scenarios, including bioinformatics, political science, and information retrieval [20], [24], [38]. Topic models assume that documents are comprised of some number of "topics" or distributions over words from a fixed vocabulary related to a single theme. For example, the words "healthcare," "hospital," and "medicine" are all related to a similar theme. The goal of topic modeling is to discover these topics and their proportions across the document set.

The most important assumption made by topic models is that word order is unimportant when determining the topics discussed in a document. Consider the words "healthcare," "hospital," and "medicine" mentioned earlier. Even when randomized, they still share the same theme. This assumption is commonly referred to as the "bag of words" assumption. Topic modeling makes the following additional assumptions:

1) Documents are made of topics, and topics are made of words.
2) The topics are identified automatically rather than being manually specified.
3) Topics are shared across documents.

The second assumption is particularly important for the analysis of policy documents. Organizations, not regulators, choose what to include in their policy documents. Once chosen, their statements are binding and enforceable. Thus, it is up to regulators to determine the content of the policy documents to ensure the organization is held accountable for their statements. Topic modeling automates the process of identifying these topics.

The third assumption states that the documents in a collection examined using a topic model all share the same set of topics, but they may have different proportions of those topics. For example, all the documents in the collection may share some proportion of a topic including the terms "shoe," "shirt," and "glove," but an article devoted almost entirely to scarves may not include that topic.

We apply topic modeling to determine the underlying topics in our policy document collection. Our previous work in this area discovered a validated taxonomy of privacy goals and vulnerabilities [5]. This taxonomy is based on goal-based requirements analysis and consists of two broad classifications of policy goals: (1) protection goals that should be operationalized into functional requirements that preserve user privacy and (2) vulnerabilities that should be operationalized as avoidance goals to avoid compromises to user privacy [5]. Associated with these goals are keywords that represent the basic action which would be completed by any functional requirement. For example, consider the goal keywords `COLLECT`, `COMPLY`, `NOTIFY`, `REMOVE`, `MAINTAIN`, `STORE`, and `TRACK` [4]. These keywords may be represented as a topic in a topic model.

The particular topic modeling algorithm we apply herein is paper is the Latent Dirichlet Allocation algorithm developed by Blei et al. [25]. Although an intricate explanation of this algorithm is beyond the scope of this paper, we will present an overview of its operation and the assumptions it makes. LDA is a joint probability distribution over both observed and hidden random variables. The words and documents themselves are the observed variables, whereas the topic structure is a hidden variable. To identify the hidden variables, we compute the conditional probability distribution. Thus, we are able to identify the hidden variables in this model.

More formally, LDA is defined by the formula in Figure 1. Note that $\beta_{1:K}$ represents all of the topics across all of the documents and $W_{d,n}$ represents a single observed word in a particular document. The θ_d represents the topic proportions for document d. Thus, an observed word, $W_{d,n}$ depends on both the set of all topics, $\beta_{1:K}$ and the distribution of topics for a given document $Z_{d,n}$. These dependencies are built into the model.

LDA requires two provided inputs. The first is a set of documents. The second is the number of topics, represented by K, which controls the granularity of topic modeling. A model with too many topics will overfit the training data, and have a poor likelihood on additional held-out data. However, with too few topics, the model will underfit, and perform poorly on both training and test data.

LDA has two key unknown quantities: the document-topic proportions θ, and the topic-word probabilities β. The maximum-likelihood criterion suggests that we select both θ and β so as to maximize the likelihood of the observed data, $P(W|\theta, \beta)$. Unfortunately, the structure of the LDA model makes this maximization impossible [39]. We choose a coordinate-ascent approach known as variational inference [25], which greedily maximizes a lower bound on the data likelihood.[14]

We built our topic model using the R statistical computing environment [42]. There are two major topic modeling packages available for R: the `lda` package and the `topicmodels` package. We chose the `topicmodels` package, which is based on the original LDA implementation by Blei et al. [25]. An outline of our method is as follows:

1) Preprocess the policy documents collected.
2) Select a subset of the data to hold out for validation of the model.
3) Build a series of topic models.
4) Perform a best fit validation on the held out data to determine which model to use for our requirements engineering analysis.
5) Determine the extent to which the model may help a requirements engineering effort.

We now discuss each step, starting with our preprocessing procedures. We removed numbers, punctuation, and extraneous white space from the documents. We also converted all

[14] A popular alternative is a randomized algorithm known as Gibbs sampling [40]; results are comparable [41].

$$P(\beta_{1:K}, \theta_{1:D}, Z_{1:D}, W_{1:D}) = \prod_{k=1}^{K} P(\beta_k) \prod_{d=1}^{D} P(\theta_d)(\prod_{n=1}^{N} P(Z_{d,n}|\theta_d)P(W_{d,n})|\beta_{1:K}, Z_{d,n}) \quad (1)$$

Fig. 1. LDA's Joint Distribution of Hidden and Observed Variables

text to lower case. We also stemmed the words to consolidate inflected word forms to their root form. For example, "collecting," "collected," and "collection" all refer to the same root concept: "collect." We used the Porter stemming algorithm [43], which is available as a part of the `Snowball` package in R. We removed stopwords using the list of common English stopword provided by the `tm` text mining package[15] for R. In addition to dropping extremely common words, we also dropped extremely rare words. We used term-frequency inverse-document-frequency (tf-idf) analysis to do so [44]. Essentially, tf-idf describes the importance of a word to a document while also being tempered by its importance to the rest of the corpus. We removed words that fell below the mean tf-idf value.

After preprocessing, we randomly chose 10% of the corpus as our set of data to hold out of the model building so that we could use it to validate the candidate models. To build a topic model using LDA with variational estimation, the only parameter that must be selected a priori is the number of topics, K. To determine a best value for K built 35 topic models over the common range of values for K. Once these models were built, we calculated the perplexity of each model as applied to our held out data. *Perplexity* is a measure of the predictive likelihood of a model of text, which can be applied to topic models. When used to evaluate language models, perplexity is often evaluated per word. For example, the lowest perplexity published for the Brown Corpus, a large and diverse corpus of English, is around 247 per word [45]. Although there is some subtlety in applying perplexity, lower values of perplexity on the held out data set demonstrate a stronger model performance [46].

VI. Results

We discuss the results of our analysis in the same order as our methodology: (1) readability, (2) building and validating a topic model of the policies, and (3) exploring privacy protection goals and vulnerabilities using the topic model.

A. Readability Results

Our prior work consisted of two studies. The first study of 40 online financial policy documents revealed a median FRE of 35.16 and a standard deviation of 9.33 [4]. The second study of 24 online healthcare policy documents revealed a median FRE of 32.16 and a standard deviation of 12.51 [6]. For this study, we treat the policy documents from the Google Top 1000 websites and the Fortune 500 as separate sets for comparison. The only preprocessing done was to manually extract the plain text so that they could be analyzed using the `koRpus` statistical readability package.[16] The Google Top 1000 had a median FRE of 31.82 with a standard deviation of 13.52 and the Fortune 500 had a median FRE of 27.49 with a standard deviation of 12.43. Recall that lower scores of the FRE indicate a document that is more difficult to read. A score of 65 is considered to be "plain English," and a score of below 40 is considered to be "difficult to read" [29]. None of the document sets could be considered to be "plain English."

We also examined the FGL, FOG, SMOG, and ARI for each of the four sets of policy documents (e.g. our first study, our second study, the Google top 1000, and the Fortune 500). Table I presents a summary of our results. For each metric, we present the fifth percentile, the mean, and the ninety-fifth percentile with the standard deviation in parenthesis next to the mean. All of these metrics are designed to produce a grade level indicator for the education system in the United States. Grade levels 9 through 12 correspond to freshman year to senior year in high school, and grade levels 13 through 16 correspond to freshman year to senior year in college.

These results indicate that policy documents remain extremely difficult to read. Both the Google Top 1000 and the Fortune 500 policy documents are rated more challenging to read than the policy documents in the first two studies. This may be the result of regulatory influence in the five years since our second study was conducted. Regardless of the cause, the implications are clear: requirements engineers need tools and techniques to analyze these documents and ensure that software deployed by organizations lives up to the promises in their policies. Official policy documents should reflect an organizational commitment and serve as a mutually understandable agreement between the organization and the consumer. The challenge of interpreting these policies does not fall on requirements engineers alone. Regulators and customers also need to evaluate and understand these policies. Even if these policies were easily readable and coherent, which is clearly not the case, the sheer number and length of policies would remain an obstacle to overcome [18]. For all of these reasons, we believe the use of text mining techniques, which can improve and augment both requirements engineering analysis and regulatory understandability, are justifiable and worthwhile pursuits.

B. Topic Modeling Results

There are two required inputs to building a topic model: (1) the corpus of documents and (2) the number of topics assumed to comprise the corpus as a whole. To build our topic model, we combined all the policy documents collected into a single corpus of documents. This corpus consisted of

[15] http://cran.r-project.org/web/packages/tm/

[16] http://cran.r-project.org/web/packages/koRpus/index.html

TABLE I
GRADE LEVEL READABILITY OF POLICY DOCUMENTS

Document Set	FGL 5%, Mean (σ), 95%	FOG 5%, Mean (σ), 95%	SMOG 5%, Mean (σ), 95%	ARI 5%, Mean (σ), 95%
First Study [4]	8.7, 13.5 (2.34), 16.4	10.8, 14.9 (2.23), 18.2	12.2, 15.2 (1.72), 17.3	7.6, 13.7 (2.87), 17.0
Second Study [6]	10.7, 13.9 (2.81), 19.1	12.3, 15.5 (2.08), 19.0	13.0, 15.6 (2.10), 19.4	9.7, 13.6 (2.96), 18.6
Google Top 1000 Sites	12.2, 15.4 (3.27), 21.5	12.5, 16.0 (2.90), 21.0	14.0, 16.6 (2.15), 20.6	11.4, 15.3 (4.00), 22.6
Fortune 500	11.2, 14.8 (3.67), 20.1	11.9, 15.7 (3.28), 20.2	13.3, 15.9 (2.09), 19.0	10.5, 14.7 (4.47), 21.1

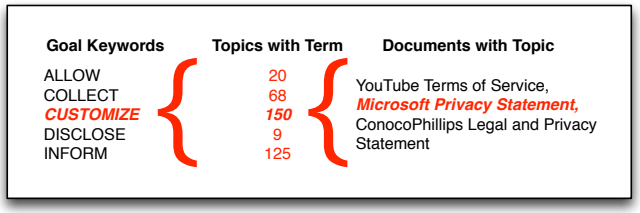

Fig. 2. Application of Topic Model to Requirements Engineering

TABLE II
NUMBER OF POLICY DOCUMENTS (OUT OF 2,061) IDENTIFIED AS POTENTIALLY CONTAINING GOAL STATEMENTS

Key-word	Docu-ments	Key-word	Docu-ments	Key-word	Docu-ments
access	904	apply	331	change	31
collect	202	comply	339	connect	121
display	308	help	61	honor	19
inform	23	limit	52	notify	347
opt-in	32	opt-out	76	post	76
request	31	reserve	51	share	300
specify	38	store	38	use	525

2,061 documents, all of which were preprocessed according to the description in our methodology section. We randomly selected 207 of these to serve as our held out document set for validation of the model, leaving 1,854 documents to serve as the first input used to build our topic model.

A common approach to determining the number of topics needed to build an effective topic model is to build a series of models and determine which produces the best fit on a held out set of data. To build our series of topic models, we first constructed 20 topic models based on a value of k (the number of topics chosen prior to building the model) equally spaced over the range 10 to 160. Once these models were built, we then identified the model that had the lowest perplexity value against the held out data set. We then build another 15 models using that model's value of k as a middle point. Once those models were built, we again selected the lowest perplexity value (129.78) of those models against the held out data set, which gave us a model built with $k = 154$ topics.

Having built the model, we then sought to validate it by using it to identify policy documents likely to contain privacy protections and vulnerabilities when analyzed using goal-based requirements engineering. Figure 2 demonstrates the process for applying our topic model for a requirements engineer seeking to identify documents that may contain privacy protections or vulnerabilities. First, the requirements engineer selects a goal keyword of interest to study. In this example, we have chosen the `CUSTOMIZE` goal keyword. This corresponds to several goals specified in our previous analysis [5], including G_{109}: `CUSTOMIZE` content to specific customer using demographic/profile data. The topic model indicates several topics that contain 'customiz,' which is the stemmed version of this keyword. These are all shown in red on the figure. Then, the requirements engineer selects one or more of these topics to identify documents that are likely to contain this topic. In this example, we have selected only topic number 150 because this is the topic most likely to be associated with the keyword. Note that topics are assigned numbers rather than semantically meaningful names because the topic model does not guarantee that the topic has a semantically meaningful name. The only guarantee provided by the topic model is that these words are significantly associated with one another as a topic in the collection of documents. Finally, the requirements engineer identifies documents in the model likely to contain the topic associated with the selected keyword. In this example, all of the documents listed are associated with topic 150. Microsoft's Privacy Statement[17] contains the following text:

> Microsoft collects and uses your personal information to operate and improve its sites and services. These uses include providing you with more effective customer service; making the sites or services easier to use by eliminating the need for you to repeatedly enter the same information; performing research and analysis aimed at improving our products, services and technologies; and displaying content and advertising that are customized to your interests and preferences.

When these statements are analyzed using the goal-based requirements analysis heuristics developed in our prior work, they yield, among other goals, the original goal (G_{109}).

We conducted searches as described in Figure 2 on the stemmed versions of all 57 of the goal keywords published

[17] http://privacy.microsoft.com/en-us/fullnotice.mspx

in our prior work [4]. To ensure more meaningful results, we limited our search to the 50 most likely terms per topic and the 20 most likely topics per document. Twenty-one of the 57 goal keywords identified a subset of policy documents more likely to contain relevant privacy protection goals or vulnerabilities than the remainder of the document collection. Table II displays the number of documents identified as potentially containing goal statements by each of these 21 goal keywords. There is no guarantee that the documents actually contain goal statements related to the keywords; that analysis must be conducted by a requirements engineer. Topic models are probabilistic, and thus, we only know that these documents have been identified as more likely to contain goal statements, according to the LDA algorithm related, to the keywords than the other documents in the collection. Note that the results described in Table II are significantly different than what a word search would have revealed. When searching for a stemmed goal keyword using this topic model, any resulting documents may contain a topic in which this word plays a significant role. A plain word search would simply return all documents containing the word without an indication of the significance of that word in the document.

To confirm these policy documents contain privacy protection goals or vulnerabilities would have been impractical due to the number of policy documents involved. We were able to randomly check several policy documents, including the Microsoft Privacy Statement discussed earlier, but we still need to examine a reasonably large sample of the documents to verify that they contain goals indicated by the keyword search. Although our preliminary analysis is limited, we believe it demonstrates that this approach is useful in large-scale requirements analysis. Consider that a requirements engineer could limit the scope of a search from 2,061 policy documents to, in several cases, fewer than 100 policy documents that may contain discuss topics related to a particular goal keyword.

VII. Threats to Validity

The evaluation of topic models remains an open research question [23], [47]. Construct validity, which deals with the question of whether the topics identified should truly be thought of as "topics" discussed in the documents, is a particularly important concern. Most researchers validate their topic model by holding out a subset of their corpus and fitting the model generated to that subset, as we have done in this paper. In addition, topic models are often used for exploring, organizing, or summarizing large document collections, which is how we have used them in this work. Although there is no accepted solution to this threat to validity, it is an active area of research that we intend to pursue [23], [47].

In addition to construct validity, we must also consider threats to internal and external validity. Internal validity refers to the validity of causal relationships established. Because we model and describe policies rather than make claims about causality in this study, internal validity is not a concern. External validity refers to the generalizability of the results to other cases. To our knowledge, this is the larest content analysis of privacy policies completed to date. Still, more organizations are not included in the study than are included. We plan to extend this work to cover more policies and organizations in the future.

Another concern is potential loss of contextual information when using a goal keyword as a search term. Other elements of goals, such as objects or actors, may have an important role to play in the discovery of new goals in unrelated policy documents. In addition, the goal heuristics developed along with the Antón-Earp taxonomy call for using keywords that may not appear in the text of the policy document. For example, consider the policy statement:

Our cookies will never be used to track your activity on any third party Web sites or to send spam, ...

Using the goal heuristics [5], this statement was translated as a prevention goal and stated as follows:

G_{53}: PREVENT use of cookies to send spam.

Our approach may miss goals of this nature for two reasons. First, because the heuristics sometimes require the use of keywords that do not appear in the policy text, they cannot be detected by the model. We may be able to mitigate this threat in the future by considering semantically equivalent word forms or using a language ontology as we build our topic model. In this initial investigation, we simply sought to determine whether the standard "bag of words" approach would yield actionable results. Second, limiting searches to the 50 most likely terms per topic and limiting the topics to the 20 most likely per document may result in missed goals. However, these simplifying assumptions allowed us to readily visualize relevant documents. Visualizing topic models is an open research question, and not the subject of this study. However, we intend to investigate this in our future work. For example, we would like to develop a search interface that accounts for the actual likelihoods of terms or of topics rather than using an ordinal list with a cutoff.

We would also like to compare our approach to more straightforward attempts to limit the number of policy documents, such as searching for a single keyword that appears a certain number of times in a document. We believe that topic modeling offers two advantages over such an approach. First, our keyword search expands to topics that contain the term before searching for documents that contain the topic. Thus, this approach may identify documents discussing a topic without using the keyword. Second, topic models do not make the simplifying assumption that individual words represent individual topics, which may be misleading.

In the future, we would like to incorporate the heuristics in our application of the topic model to identifying relevant policy documents. For example, we could use the complete set of goals in the Privacy Goals Management Tool repository [4]. This would have allowed us to search for words like spam, which may reveal the policy document from which G_{53} was identified.

VIII. Summary and Future Work

To our knowledge, our examination of 2,061 policy documents is the largest readability analysis of requirements source documents conducted. Our results demonstrate that policy documents are similarly challenging to read and understand (RQ1). Additional tools and techniques are needed to support the software requirements engineers building systems that must uphold the promises these documents make to end users. The results of our work also indicate that topic models can indicate whether a document contains software requirements expressed as privacy protections or vulnerabilities (RQ2). These requirements have serious implications for requirements engineers or regulators seeking to build or evaluate software systems that must comply with these policies. Clearly, topic models cannot replace requirements engineering analysis conducted by trained individuals. Applying the heuristics [5] needed to extract goals from these documents requires trained engineers. This matches the common understanding that natural language processing techniques are not capable of specifying software requirements [22]. Finally, our results provide preliminary support for the generalizability of the Antón-Earp taxonomy to multiple domains (RQ3). However, further research is needed to confirm these early findings. We plan to identify a significant subset of randomized results from searches conducted on our topic model and perform a complete goal-based requirements analysis on them to determine the precise number of goals these policies contain. In addition, we will compare the results of this analysis across domains.

Another important consideration for future work is whether topic modeling can reveal specific software requirements expressed as privacy protections or vulnerabilities in policy documents. In this paper, we only seek to identify documents for which a goal-based requirements analysis may prove fruitful, but with expanded visualization and search procedures, we may be able to narrow these results to documents likely to contain specific privacy protections or vulnerabilities. Topic modeling is recognized to be valuable in part because of the highly modular design of the algorithm [20]. By relaxing some of the assumptions of topic modeling, we may be able to increase the accuracy and utility of our approach.

Topic modeling enables analysis of policy documents at a scale that would be impossible through individual annotation alone, but these documents must be available in a *machine-readable* format. In the past, efforts such as P3P [48] focused on *machine-accessible* formats for privacy policies that could automatically negotiate privacy preferences with individuals. Consider the robots.txt standard, which calls for a plain text file in a standard location to communicate which links are safe for automated web crawlers. Most websites adopt the robots.txt standard, and it is generally considered successful. A similar standard for plain text versions of policy documents would, if widely adopted, make data collection and processing much less time consuming and potentially more accurate.

Acknowledgments

We would like to thank Allison Massey for her help in the data collection process. This work was partially funded by NSF Grant #121769.

References

[1] M. Andreessen, "Why Software is Eating the World," Wall Street Journal, http://online.wsj.com/article/SB10001424053111903480904576512250915629460.html, Aug. 2011.

[2] K. Cukier, "Data, data everywhere," The Economist, http://www.economist.com/node/15557443, Feb. 2010.

[3] B. Schneier and M. Pasiewicz, "On People, the Death of Privacy, and Data Pollution," EDUCAUSE Review, http://www.educause.edu/ero/article/people-death-privacy-and-data-pollution, March / April 2008.

[4] A. Antón, J. Earp, Q. He, W. Stufflebeam, D. Bolchini, and C. Jensen, "Financial privacy policies and the need for standardization," *Security and Privacy, IEEE*, vol. 2, no. 2, pp. 36–45, Mar - Apr 2004.

[5] A. I. Antón and J. B. Earp, "A requirements taxonomy for reducing web site privacy vulnerabilities," *Requirements Engineering*, vol. 9, no. 3, pp. 169–185, 2004.

[6] A. I. Anton, J. B. Eart, M. W. Vail, N. Jain, C. M. Gheen, and J. M. Frink, "Hipaa's effect on web site privacy policies," *Security and Privacy, IEEE*, vol. 5, no. 1, pp. 45–52, Jan.-Feb. 2007.

[7] J. D. Young and A. I. Antón, "A Method for Identifying Software Requirements Based on Policy Commitments," *18th International IEEE Requirements Engineering Conference*, 2010.

[8] J. D. Young, "Commitment Analysis to Operationalize Software Requirements from Privacy Notices," *Requirements Engineering*, vol. 16, no. 1, pp. 33–46, March 2010.

[9] I. Rubinstein, "Regulating privacy by design," *Berkeley Technology Law Jounral*, vol. 26, p. 1409, 2012.

[10] A. I. Antón and J. B. Earp, "Strategies for developing policies and requirements for secure e-commerce systems," in *Recent Advances in E-Commerce Security and Privacy*, A. Gosh, Ed. Kluwer Academic Publishers, 2001, pp. pp. 29–46.

[11] A. I. Antón, J. B. Earp, and R. A. Carter, "Precluding incongruous behavior by aligning software requirements with security and privacy policies," *Information and Software Technology*, vol. 45, no. 14, pp. 967–977, 2003.

[12] T. Breaux and A. Antón, "Deriving semantic models from privacy policies," *Policies for Distributed Systems and Networks, 2005. Sixth IEEE International Workshop on*, pp. 67–76, June 6-8, 2005 2005.

[13] D. S. Allison, H. F. El Yamany, and M. Capretz, "Metamodel for privacy policies within SOA," in *IWSESS '09: Proceedings of the 2009 ICSE Workshop on Software Engineering for Secure Systems*. Washington, DC, USA: IEEE Computer Society, 2009, pp. 40–46.

[14] W. Robinson, "Implementing rule-based monitors within a framework for continuous requirements monitoring," *System Sciences, 2005. HICSS '05. Proceedings of the 38th Annual Hawaii International Conference on*, pp. 188a–188a, Jan. 2005.

[15] P. N. Otto and A. I. Antón, "Addressing Legal Requirements in Requirements Engineering," *Requirements Engineering Conference, 2007. RE '07. 15th IEEE International*, pp. 5–14, 15-19 Oct. 2007.

[16] M. W. Vail, J. B. Earp, and A. I. Antón, "An empirical study of consumer perceptions and comprehension of web site privacy policies," *Engineering Management, IEEE Transactions on*, vol. 55, no. 3, pp. 442–454, Aug. 2008.

[17] A. McDonald, R. Reeder, P. Kelley, and L. Cranor, "A comparative study of online privacy policies and formats," in *Privacy Enhancing Technologies*, ser. Lecture Notes in Computer Science, I. Goldberg and M. Atallah, Eds. Springer Berlin Heidelberg, 2009, vol. 5672, pp. 37–55.

[18] A. M. McDonald and L. F. Cranor, "The Cost of Reading Privacy Policies," *I/S: A Journal of Law and Policy for the Information Society*, vol. 2008 Privacy Year in Review Issue, 2008.

[19] A. Acquisti and J. Grossklags, "Privacy and rationality in individual decision making," *IEEE Security and Privacy*, vol. 3, no. 1, pp. 26–33, Jan.-Feb. 2005.

[20] D. M. Blei, "Probabilistic topic models," *Commun. ACM*, vol. 55, no. 4, pp. 77–84, Apr. 2012.

[21] M. Steyvers and T. Griffiths, "Probabilistic topic models," in *Latent Semantic Analysis: A Road to Meaning.*, T. Landauer, D. Mcnamara, S. Dennis, and W. Kintsch, Eds. Laurence Erlbaum, 2006.

[22] K. Ryan, "The role of natural language in requirements engineering," in *Requirements Engineering, 1993., Proceedings of IEEE International Symposium on*, Jan 1993, pp. 240–242.

[23] J. Grimmer and B. M. Stewart, "Text as Data : The Promise and Pitfalls of Automatic Content Analysis Methods for Political Texts," *Political Analysis (forthcoming)*, 2013.

[24] J. Grimmer, "A bayesian hierarchical topic model for political texts: Measuring expressed agendas in senate press releases," *Political Analysis*, vol. 18, no. 1, pp. 1–35, 2010.

[25] D. M. Blei, A. Ng, and M. Jordan, "Latent dirichlet allocation," *JMLR*, vol. 3, pp. 993–1022, 2003.

[26] L. F. Cranor, "Necessary But Not Sufficient: Standardized Mechanisms for Privacy Notice and Choice," *Journal of Telecommunications and High Technology Law*, vol. 10, no. 2, 2012.

[27] I. Reay, P. Beatty, S. Dick, and J. Miller, "A survey and analysis of the p3p protocol's agents, adoption, maintenance, and future," *Dependable and Secure Computing, IEEE Transactions on*, vol. 4, no. 2, pp. 151–164, April-June 2007.

[28] The World Bank, "Internet Users (per 100 people)," *The World Bank Group, http://data.worldbank.org/indicator/IT.NET.USER.P2?cid=GPD_44*, 2012.

[29] R. F. Flesch, "A New Readability Yardstick," *Journal of Applied Psychology*, vol. 32, pp. 221–233, 1948.

[30] R. Gunning, *The Technique of Clear Writing*. McGraw-Hill International Book Co, 1952.

[31] G. McLaughlin, "SMOG grading – a new readability," *Journal of Reading*, May 1969.

[32] J. Kincaid, J. R.P. Fishburne, R. Rogers, and B. Chissom, "Derivation of new readability formulas (Automated Readability Index, Fog Count and Flesch Reading Ease Formula) for Navy enlisted personnel," *Research Branch Report 8-75, Naval Technical Training, U. S. Naval Air Station, Memphis, TN.*, 1975.

[33] J. Kincaid and E. Smith, "Derivation and Validation of the Automated Readability Index for use with Technical Materials," *Human Factors*, vol. 12, pp. 457–464, 1970.

[34] E. Fry, "A readability formula for short passages," *Journal of Reading*, vol. 33, no. 8, pp. 594–597, 1990.

[35] J. Chall and E. Dale, *Readability Revisited: The New Dale-Chall Readability Formula*. Brookline Books, Cambridge, MA., 1995.

[36] L. Si and J. P. Callan, "A statistical model for scientific readability," *CIKM 2001*, pp. 574– 576, 2001.

[37] K. Collins-Thompson and J. P. Callan, "A language modeling approach to predicting reading difficulty," *HLT-NAACL*, pp. 193–200, 2004.

[38] C. Templeton, "Topic Modeling in the Humanities: An Overview," *Maryland Institute for Technology in the Humanities, http://mith.umd.edu/topic-modeling-in-the-humanities-an-overview/*, 2011.

[39] D. Sontag and D. M. Roy, "Complexity of Inference in Latent Dirichlet Allocation," *Twenty-Fifth Annual Conference on Neural Information Processing Systems*, 2011.

[40] T. Griffiths and M. Steyvers, "Finding scientific topics," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 101, no. Suppl 1, pp. 5228–5235, 2004.

[41] A. Asuncion, M. Welling, P. Smyth, and Y. W. Teh, "On smoothing and inference for topic models," in *Proceedings of the Twenty-Fifth Conference on Uncertainty in Artificial Intelligence*, ser. UAI '09. Arlington, Virginia, United States: AUAI Press, 2009, pp. 27–34.

[42] R Development Core Team, "R: A Language and Environment for Statistical Computing," *R Foundation for Statistical Computing, http://www.R-project.org*, 2012.

[43] M. Porter, "An Algorithm for Suffix Stripping," *Program*, vol. 14, no. 3, pp. 130–137, 1980.

[44] K.S. Jones, "A statistical interpretation of term specificity and its application in retrieval," *Journal of Documentation*, vol. 28, no. 1, pp. 11–21, 1972.

[45] P. F. Brown, S. A. D. Pietra, V. J. D. Pietra, J. C. Lai, and R. L. Mercer, "An Estimate of an Upper Bound for the Entropy of English," *Computational Linguistics*, vol. 18, no. 1, 1992.

[46] H. Wallach, I. Murray, R. Salakhutdinov, and D. Mimno, "Evaluation methods for topic models," in *Proceedings of the 26th Annual International Conference on Machine Learning*. ACM, 2009, pp. 1105–1112.

[47] D. M. Blei, "Introduction to Probabilistic Topic Models," *Available: http://www.cs.princeton.edu/~blei/topicmodeling.html*, 2012.

[48] L. Cranor, "P3P: making privacy policies more useful," *IEEE Security and Privacy*, vol. 1, no. 6, pp. 50–55, Nov.-Dec. 2003.