Privacy Policy Annotation for Semi-automated Analysis: A Cost-Effective Approach

Dhiren A. Audich, Rozita Dara, Blair Nonnecke

Privacy Policy Annotation for Semi-Automated Analysis: A Cost-Effective Approach

Dhiren A. Audich, Rozita Dara, and Blair Nonnecke

University of Guelph,
50 Stone Rd., Guelph,
ON, N1G 2W1, Canada
{daudich, drozita, nonnecke}@uoguelph.ca

Abstract. Privacy policies go largely unread as they are not standardized, often written in jargon, and frequently long. Several attempts have been made to simplify and improve readability with varying degrees of success. This paper looks at keyword extraction, comparing human extraction to natural language algorithms as a first step in building a taxonomy for creating an ontology (a key tool in improving access and usability of privacy policies).
In this paper, we present two alternatives to using costly domain experts are used to perform keyword extraction: trained participants (non-domain experts) read and extracted keywords from online privacy policies; and second, supervised and unsupervised learning algorithms extracted keywords. Results show that supervised learning algorithm outperform unsupervised learning algorithms over a large corpus of 631 policies, and that trained participants outperform the algorithms, but at a much higher cost.

1 Introduction

A 2015 Pew Research Centre survey found that 91% of American adults either agree or strongly agree that they have lost control of how their private information is collected and used [1]. The collection of personally identifiable information (PII) by online service providers is often justified with claims of creating a more user-centric web experience. However, PII is sold and shared frequently with third parties that use it to profile users and track them across domains. While users are increasingly concerned about their privacy online [2] they scarcely understand the implications of PII sharing [3].

Privacy policies are the only means of informing users and mitigating their fears over privacy loss, and by law, companies have to disclose the gathering, processing, and sharing of PII in their privacy policies [4, 5, 6, 7]. Unfortunately, most policies are often lengthy, difficult and time-consuming to read, and as a result are infrequently read [8, 9, 10, 2]. The demotivating nature and the difficulty of reading privacy policies amounts to a lack of transparency. Failing to provide usable privacy policies prevents users from making informed decisions and can lead them to accept terms of use jeopardizing their privacy and PII.

Recently, Cranor et al. showed through their analysis of 75 policies that most policies do not provide enough transparency about data collection for the users to make informed privacy decisions [11].

In addition to length and readability, privacy policies also differ from one another by their content of legal and technical jargon, and coverage [9, 8, 12]. While it is true that FIPPs (Fair Information Practice Principles) and OECD (Organization for Economic Co-operation and Development) offer general guidelines for writing privacy policies, they only provide a conceptual framework; a qualitative review of policies reveals that language and structure being used differs between policies and economic zones (E.U., U.S.A, Canada) [9, 13]. There is also an inconsistent amount of jargon used between policies, and policies of organizations in the E.U. tend to have supplementary information that tends to be absent from American and Canadian policies [12, 14, 11]. Boilerplate language is mostly the norm for cookie policies.

To improve readability and semi-automate their evaluation, attempts to introduce a standard structure to privacy policies has met with limited success. For example, the preeminent Platform for Privacy Preferences (P3P) [15, 16] failed due to poor adoption and issues related to validating policies [17]. In response to user concerns, several prominent online service providers started using privacy enhancing technologies (PET), e.g., opt-out mechanisms, anonymisation of personal data, and layered policies [18, 19]. Without these becoming a common standard, opaque and verbose policies are still the norm.

Given the widespread deployment of privacy policies and their importance to users, we propose to semi-automate the evaluation of policies and support their reading through the use of intelligent reasoning. By combining intelligent reasoning with natural language processing (NLP) techniques we hope to reduce the time needed for users to find key information in policies by highlighting sections directly related to users' privacy concerns.

The first step to constructing an intelligent reasoning system would be to conduct contextual analysis of privacy policies, for comprehension, and capture it in a taxonomy, a hierarchical representation of privacy policy concepts. In other words, the taxonomy would capture the vocabulary of privacy policies. There are two main ways to capture vocabularies: manually and automatically. Manually capturing the vocabulary involves reading and knowing domain text which would involve hours of manual labour that can be costly. An easier approach would be to use NLP techniques to automatically extract keywords for taxonomy creation [20, 21]. This requires fewer man hours compared to the manual methods; hence, is cheaper. Presently, no taxonomy exists for the online privacy policy domain. Hence, the ultimate objective of this research is to create and validate a taxonomy with the aid of NLP algorithms and then further by subject matter experts. This paper focuses only on the keyword and keyphrase extraction part of this larger process.

Keyword and keyphrase extraction forms the backbone of topic modelling and information retrieval systems. In this paper, NLP was used to extract keywords and keyphrases from hundreds of policies. In addition, we employed trained participants to extract keywords and phrases from a much smaller subset of policies, with the intent of comparing the manual extraction to that of the NLP. The overall goal of evaluating automatic keyword extraction algorithms was to examine which algorithm performs best against human annotators for the domain of online privacy policies. The best of these would then be chosen to become part of a larger process to enrich domain-expert curated taxonomy which would then be used to construct an ontology.

Our research confirms that whilst automatic keyword and keyphrase extraction remains a difficult task, supervised learning algorithms perform marginally better against unsupervised algorithms. Furthermore, trained annotators, collectively, can cheaply out-perform domain experts; their combined output being further used to improve the training set for the supervised learning algorithm.

The remainder of this paper is structured as follows: Sec. 2 discusses related work; Sec. 3 describes previous work on which the current research is built on; Sec. 4 presents motivation and describes methodology; Sec. 5 presents our investigation with training non-domain experts for the task of keyword and keyphrase extraction; Sec. 6 presents our investigation after working a supervised learning algorithm for the task of keyword and keyphrase extraction; Sec. 7 discusses the implication and avenues for future work; we conclude our work in Sec. 8.

2 Related Work

In a previous study, Wilson et al. (2016) created the OPP-115 corpus, a corpus of 115 manually annotated privacy policies with $23,000$ *data practices*. A *data practice* is roughly defined as a purpose or consequence of collecting, storing, or generating data about a user. In the study, the 10 domain experts (privacy experts, public policy experts, and legal scholars) used a custom designed web-based tool to annotate practices and assign various attributes to them for classification purposes. Each policy took an average of 72 minutes to annotate. Whilst this approach produced a high quality and nuanced data set, it is costly to expand and maintain such a knowledge base both in time and money. Automation and crowdsourcing could reduce the cost of creating and maintaining such a data set, and still maintain a reasonable amount of quality.

In the research conducted by Ramnath et al., the researchers proposed combining machine learning and crowdsourcing (for validation) to semi-automate the extraction of key privacy practices [22]. Through their preliminary study they were able to show that non-domain experts were able to find an answer to their privacy concern relatively quickly ($\sim$ 45s per question) when they were only shown relevant paragraphs that were mostly likely to contain an answer to the question. They also found that answers to privacy concerns were usually concentrated rather than scattered all over the policy. This is an important find because it means that if users were directed to relevant sections in the policy they would be able to address their privacy concerns relatively quickly instead of reading the entire policy. Additionally, Pan and Zinkhan have showed that when users are presented with a short and straightforward policy, they are more inclined to read it [23].

In a more recent user study conducted by Wilson et al. (2016), the quality of crowdsourced answering of privacy concerns was tested against domain experts with particular emphasis on highlighted text. The researchers found that highlighting relevant text had no negative impact on accuracy of answers. They also found out that users tend not to be biased by the highlights and are still likely to read the surrounding text to gain context and answer privacy concerning questions. They also found an 80% agreement rate between the crowdsourced workers and the domain experts for the same questions [24]. Similarly, Mysore Sathyendra et al. showed through their study that it was possible to highlight and extract opt-out practices from privacy policy using keywords and classification algorithms with reasonable accuracy; of the various models tested, best model used a logistic regression classification algorithm with a manually crafted feature set and achieved an F1 score of 59% [25].

The general drawback of crowdsourcing, especially with respect to privacy policies, is that it relies on non-expert users to read policies to provide data. Since most users are not motivated to reading policies to begin with, it would take a long time to crowdsource enough data to be useful. However, what is clear is that highlighting relevant text with appropriate keywords can still provide some feedback to the concerned users that are inclined to read shorter policies.

3 Background

Keyword/keyphrases extraction remains a difficult task; the state-of-the-art performance of keyword extraction algorithms hovers around $20-30\%$ [26]. Keyword (or keyphrase) extraction has been historically used to recognize key topics and concepts in documents. This task involves identifying and ranking candidate keywords based on the relatedness to the document. Keyword extraction algorithms utilize various techniques to perform their task: statistical learning, part-of-speech (POS) tagging, lexical and syntactic feature extraction. Generally, they work in two steps:

1. Identifying candidate keywords/keyphrases from the document using heuristics.
2. Recognizing if the chosen candidate keywords/keyphrases are correct or not using supervised and unsupervised methods.

3.1 Supervised Learning vs. Unsupervised Learning

Machine learning can be divided into two broad types: supervised and unsupervised learning. The majority of the machine learning techniques involve supervised learning algorithms which rely on a tagged corpus for training a model to learn features (keywords) from the text. After sufficient training, the model is then applied on similar corpus to extract keywords. The keyword assignments made over the training data set forms the reference, also known as controlled vocabulary, and treated as classes used in a classification problem. Some examples of supervised learning algorithms include, K-nearest neighbour (k-NN) [27, 28], Naive Bayes (NB) [29], GenEx [30], and Support Vector Machines (SVM) [31].

Since creating a tagged corpus is a very time consuming task, unsupervised learning algorithms are used which do not require any training set for the training of models. They instead rely on linguistic and statistical features of the text. The task is framed as a ranking or clustering problem.

3.2 Per Policy Keyword Extraction Fares Best

Our previous research [32] used five unsupervised learning algorithms to extract keywords for the purpose of identifying key concepts with the goal of generating a taxonomy for the online privacy policy domain. The research was conducted in two experiments. In the first, the algorithms were evaluated over a smaller corpus where a set of manually extracted terms by the researcher was used as the baseline. Researcher's manually extracted terms are also used in Experiment I (Sec. 5) as the baseline. Second, the algorithms were evaluated over a larger corpus where the results from the best performing algorithm from the first experiment was held as the baseline. While the algorithm Term-Frequency Inverse Document Frequency (TF-IDF) achieved an F_1-score of 27% over a small corpus (21 policies), over a large corpora (631 policies) algorithms evaluating single documents individually, such as AlchemyAPI [1] and TextRank [33], performed the best.

Results from both experiments suffered from four major types of errors. *Overgeneration* errors are a type of precision error where the algorithm incorrectly identifies a candidate term as a keyword because one of its sub-words appears frequently in a document or corpus. *Redundancy* errors are a type of precision error where the algorithm correctly identifies a keyword simultaneously identifying another keyword that is semantically similar, e.g. 'account use' and 'account usage'. *Infrequency* errors occur when a candidate term is not selected due to its low frequency of appearance in a document or a corpus. *Evaluation* errors are a type of recall error that occur when a candidate term is not identified as a keyword despite it being semantically similar to a baseline keyword. Since the unsupervised algorithms focus on the task of ranking and/or clustering based on semantic and lexical analysis, these errors are a result of language used in the privacy policies which tends to be inconsistent between policies.

The alternative to unsupervised learning algorithms are supervised learning algorithms in which a model is first trained on a set of manually extracted terms, and thus the task becomes one of classification, i.e., whether a candidate keyword should be classified as a document keyword or not. In case of privacy policies, we expect the results to improve because training a model would reduce various errors that occurred with unsupervised learning. For example, since the keywords are first being extracted manually; if the terms are infrequent, the

[1] `http://www.alchemyapi.com/products/alchemylanguage/keyword-extraction`

trained model would learn this bias reducing infrequency errors. In this paper we employ a supervised learning algorithm to test our hypothesis.

4 Study Design

The primary objective of this experiment was to test whether a supervised learning algorithm could outperform the unsupervised learning algorithms used in our previous study. As such, choosing an effective supervised challenger was key. To compare unsupervised algorithms, a supervised learning algorithm, KEA, was investigated. KEA is an effective supervised learning algorithm utilizing a Naive Bayes algorithm for training learning models [34]. It is a simple and well-known algorithm that is has been often used as a baseline throughout the literature [35, 36, 37, 38, 39, 40].

KEA works in two phases: training and model creation, and extraction. In the training phase, candidate keyphrases are selected both from the training documents and the corpus, features (attributes) are calculated, and keyphrases determined. Candidate keyphrase selection works in three phases: text pre-processing, identifying the candidate keyphrases, and stemming and case-folding. Features that are calculated are, TF-IDF and first occurrence (the first time the term occurs in a document), which are then discretized for the machine learning scheme. Finally, the keyphrase are determined from the discretized values using the Naive Bayes technique [41]. In the extraction phase, the candidate keyphrase selection is repeated on the documents to calculate feature values. Then the Naive Bayes algorithm is used along with the values calculated in the model to determine if a candidate keyphrase is a keyphrase or not.

The study was broadly broken up into two parts: Experiment I examines manual extraction by trained non-domain experts, while Experiment II compares supervised and unsupervised techniques.

5 Experiment I: Manual Keyword Extraction

5.1 Participants

Four participants were selected for this experiment. Since this was a preliminary study conducted to test if non-experts can be trained enough to extract important keywords, we thought that 4 participants was enough. All had a graduate level education in Computer Science with varying knowledge of online privacy and were male with a mean age of 33.65 (range $22 - 60$). None had a research background in privacy and they rarely read website privacy policies.

5.2 Procedure

The participants were briefed on intelligent reasoning systems and taxonomies to ensure they understood the basic concepts and how their keywords would be used. A set of criteria for manual extraction was then provided with examples to help participants select the appropriate keywords and keyphrases as illustrated in Table 1). In order to have some time limit, it was estimated that it took less than 2 hours to read and annotate 5 policies. Hence, the participants were given 2 hours to read 5 privacy policies (a subset of the 21 privacy policies used for the training model; see Table 2), and highlight terms (unigrams, bigrams, trigrams, etc.) they thought were important: concepts, themes, and terms; pertaining to the online privacy domain and as outlined in the criterion. The 5 policies selected were from different industry sectors; intended for a diverse audience; and conformed to the laws of multiple countries; they included: Google, Facebook, UEFA, Royal Bank of Canada, and Wal-Mart (including policy for California). Over a 2 hour period, each participant was presented with the privacy policies in a different order to reduce the possibility of an ordering bias.

Participants used an open source program called 'Skim'[2] for annotation. A Python script was then used to extract all of the highlighted keywords and store the results in a comma separated values (CSV) file for further analysis.

Table 1. Criteria for manually extracting key terms.

Concept	**Examples**
Legal terms	*Online Privacy Protection Act, non-disclosure agreement*
Legal organizations (government, regulatory, commercial, and computing organizations)	*federal trade commission*
Acronyms of legal organizations and acts	*FTC, COPPA*
Legal entities that can be used to define an organization or an individual	*personal information, address, account id, internet protocol address*
Data sharing	*3rd party cookies, aggregate information, google analytics*
Hosting	*backup storage, servers*
Web & tech related terms	*ad data, cookies, analytics, tracking cookies*
Legal actions and legal processes	*tracking, surveillance*
Mobile privacy	*geo-location, device identification*

To ensure consistency of key term extraction across the data sets, the following post-processing steps were taken to normalize the text:

1. All of the terms were first converted to lowercase.
2. Non-printable characters (as defined by the `string.printable` set in Python3) were removed; and the remaining special characters that were not caught by

[2] `https://sourceforge.net/projects/skim-app/`

Table 2. Breakdown of the 21 privacy policy corpus for Experiment I.

Domain	**No. of websites selected**
Healthcare	1
Insurance	2
Banking & Financial	5
E-Commerce	3
File Sharing	1
Search Engines	2
Social Networking	3
EU Specific	3
Cloud Hosting	1
Total	21

previous filters (*@#), as well as other ASCII based characters from the `string.punctuation` set in Python3 were removed.

3. Tokenized numbers were also removed as they do not tend to add value to the taxonomy e.g. '1945'.
4. The standard Porter Stemming Algorithm [42] was used from the NLTK[3] library to consolidate inflected word forms to their root.
5. Finally, duplicates were removed from the resulting sets.

5.3 Results

First, the data collected from all of the participants were compared to the data set generated by the primary researcher. The results are shown in Table 3.

It must be noted that participant 3 only completed 3 of the 5 policies because he found reading some policies quite challenging and hence taking longer to read. He also reported to initially having trouble understanding the task. Despite this, participant 3 was not dropped because we were mostly interested in the quality of the keywords rather than completion of task specifically. Furthermore, our analysis is mostly based on individual work, and despite the third participant failing to complete the task we wanted to highlight that they were still able to achieve results about half as good as the researcher.

In general, participants reported that policies were repetitive and often vaguely described their intent with regard to collecting personal information. When asked to state which privacy policy was most clear and readable, Facebook was described as the most transparent with UEFA being the least. The highest F_1*-score* was 59% with a mean of 51.75%.

In order to test the collective efficacy of the annotations, the researcher's data set was compared with the combined data set of all of the participants. The results are reported in Table 4.

[3] `http://www.nltk.org/`

Table 3. Results from manual keyword extraction by participants.

	Researcher	1	2	3	4
Terms	560	581	650	353	504
Precision		49%	55%	59%	56%
Recall		51%	64%	37%	51%
F$_1$-score		50%	**59%**	45%	53%
JSC		0.67	0.58	0.71	0.64

Table 4. Comparing performance of manual extraction primary researcher vs. combined data set generated by participants.

	Researcher	Participants
Terms	560	1038
Precision		40%
Recall		75%
F$_1$-score		52%
JSC		0.65

Finally, all five data sets were compared to each other by holding one of the data set as the baseline and comparing it with the rest. The results are reported in Table 5. The mean of all of the values was 52.1%, which agreed with the previous analysis in Table 4. This was significantly higher than the $20 - 30\%$ performance of most state-of-the-art keyword extraction algorithms.

Table 5. Comparing F_1*-scores* between participants' and researcher's data sets.

	Researcher	P1	P2	P3	P4
Researcher	-	50%	59%	45%	53%
P1	50%	-	56%	46%	52%
P2	59%	56%	-	50%	58%
P3	45%	46%	50%	-	55%
P4	53%	52%	58%	55%	-

6 Experiment II: Supervised Learning

For this experiment, a corpus of 631 privacy policies as developed by the Data Management and Privacy Governance Lab at the University of Guelph was used to evaluate the supervised learning algorithm- KEA. This is the same corpus that was used to evaluate unsupervised learning algorithms in the previous paper.

6.1 Part I

In the first part of the experiment, a set of 21 policies were used for manual extraction of keywords which in turn were used to train the learning model for KEA. The 21 policies were qualitatively determined to ensure diversity. They were selected based on the their: length, transparency, comprehension (level of difficulty), intended geographic audience (U.S., E.U, Canada), industry sectors (healthcare, e-commerce, etc.), and the most visited websites[4]. A breakdown of these policies is summarized in Table 2. An overview of Experiment II Part I is shown in Fig. 1.

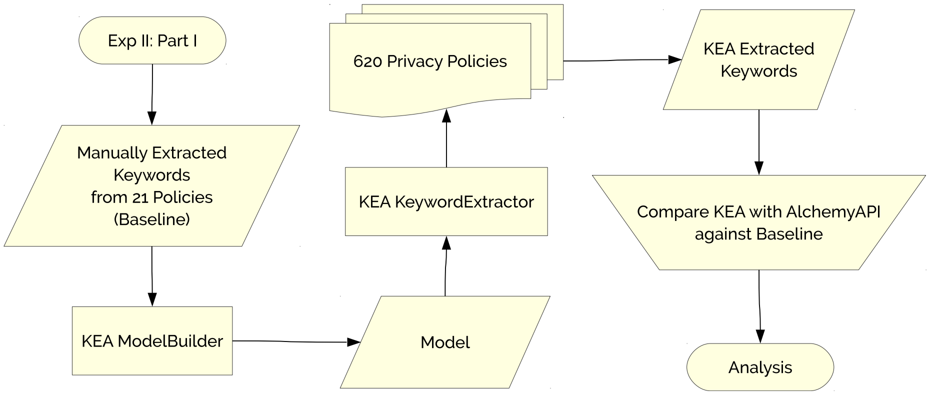

Fig. 1. Experiment II Part I: Testing KEA against manual extraction and AlchemyAPI.

Results. Once a model was trained over the manually extracted set of keywords, KEA was then run over the entire corpus. Results from the algorithm were then compared with the results from unsupervised algorithms (see Table 6). Initial results showed that the supervised algorithm performed better than the unsupervised ones but not significantly.

6.2 Part II

To demonstrate that KEA performs well over smaller training sets, and better when results across multiple annotators are combined, in the second part of the experiment, the model was trained on the annotation for only the 5 policies chosen for Experiment I (Sec. 5). The trained model was then tasked with extracting keywords and keyphrases from the rest of the corpora (626 policies). This was done for all the participants and the results were compared against each other; reported in Table 7.

[4] As listed under: `https://en.wikipedia.org/wiki/List_of_most_popular_websites`

Table 6. Comparing performance of unsupervised vs. supervised learning algorithms for keyword extraction

	AlchemyAPI	**KEA**
Terms	12635	10798
Precision	3%	4%
Recall	44%	47%
F_1-score	5%	7%
JSC	0.97	0.97

Table 7. Comparing F_1*-scores* between participants' and researcher's generated data sets from KEA.

	Researcher	**P1**	**P2**	**P3**	**P4**
Researcher	-	62%	70%	64%	72%
P1	62%	-	71%	75%	72%
P2	70%	71%	-	67%	79%
P3	64%	75%	67%	-	67%
P4	72%	72%	79%	67%	-

Since the training set only contained 5 privacy policies with roughly half of the terms present amongst all of the participants, the training set did not contain enough variance to train distinctly different models. The obvious exception being participant 3 who only annotated 3 policies.

Once again, participants' data sets were combined and compared against the researcher's data set of generated keywords to see if the accuracy rose with more number of annotators; the results are presented in Table 8. The combined data set achieves a score of F_1*-score* of 71%. One reason why the scores are different for the combined data sets (Table 4) is due to the smaller training set for the third participant. With less labelled data to train with, the training algorithm could overgeneralize what is not considered a keyword and hence ignore a large number of candidate keywords.

Table 8. Comparing performance of keyword extraction from KEA based on primary researcher vs. combined data set generated by participants.

	Researcher	**Participants**
Terms	30261	49377
Precision		57%
Recall		93%
F_1-score		71%
JSC		0.45

7 Discussion & Future Work

The mean F_1-*score* of 52% in Experiment I (Table 3) demonstrates an important aspect of keyword extraction with privacy policies- keyword extraction with privacy policies is a hard task. While 52% is a score better than the $20-30\%$ performance of most state-of-the-art keyword extraction algorithms, a qualitative analysis of the annotated keywords/keyphrases by the participants suggests that despite having a concrete set of rules, examples, and training, participant's understanding of the technical terms and the text can still result in a diverse and non-overlapping set of terms. What one participant considers an important concept is not shared by a peer. This was mostly true for the more ambiguous and less technical parts of the policies, while technical details were easily picked up by all participants. Since all of the participants had a background in Computer Science, technical details would be easy to comprehend and more transparent to them. However, this might not be the case if the participant had no technical background; then, everything would have been equally less transparent. When written ambiguously privacy policies are difficult to comprehend.

In Part I of Experiment II it was found that the supervised learning algorithm improved the F_1-*score* of keywords being extracted. This is important for two primary reasons: quality and cost. One of the most time consuming tasks involved in generating a taxonomy is capturing major important themes and concepts of the target domain. This is where keyword extraction plays an important role in reducing the time taken to capture all of the themes and concepts. In this case, instead of reading a large number of privacy policies to identify all of the important themes and concepts, supervised learning promises to be a viable alternative. The generated keywords act as candidate terms that can be used to enrich a taxonomy, thus reducing the cost and time of reading a large number of privacy policies as well as improving the quality of the taxonomy by including terms that might have been covered in text that might not have been read due to resource restrictions.

Furthermore, the diversity of keywords/keyphrases found, in Experiment I, between participants can be used to improve the training data for supervised learning. The training data in Experiment II Part I was generated by a single researcher, it might be useful for another researcher or domain expert to read a set of non-overlapping policies and generate another set of training data. This would not only validate the current training data but also create more labelled data to train the model on. It would also prove helpful to review the candidate terms generated by participants in the first experiment and enrich the present training data, i.e., carefully merge all of the data sets to create a more comprehensive data set that captures all relevant keywords/keyphrases including the ones that might have been missed by individuals within the group.

Part II of the second experiment showed that if non-domain experts are given sufficient training, a supervised learning algorithm trained with their labelled data set could result in a training model that is able to extract most of the keywords and keyphrases that are being extracted with the help of training model, and built with labelled data from domain experts. Hence, it is possible to reduce cost and train non-experts and extract keywords and keyphrases with reasonable success.

Currently, privacy policies are heterogeneous as there are no laws/guidelines that mandate a certain structure, concepts, or terminology. This makes finding, identifying and understanding relevant information a time consuming task. By utilizing an intelligent reasoning system and mapping important concepts, ideas, and themes our work helps to identify important sections in an unstructured privacy policy; thus, resulting in less time needed to find important information in policies improving transparency and making polices more usable. It could further be used to introduce structure in future policies.

8 Conclusion

In this paper, we extended our previous work on keyword and keyphrase extraction over the domain of online privacy policies. In Experiment I, the difficulty of extracting keywords from privacy policies was demonstrated and the challenges associated with this task was discussed. In Experiment II we applied a supervised algorithm for keyword extraction, and demonstrated its superior performance over unsupervised algorithms applied to the same corpus of 631 online privacy policies. Our results confirm that using natural language processing techniques for keyword and keyphrase retrieval from privacy policies remains a challenging task.

Our preliminary results will guide further research in the field of online privacy and machine learning, and making policies more transparent and usable. We intend to improve our training set by having other domain experts (researchers and lawyers) identify key concepts, ideas, and themes in a non-overlapping set of privacy policies. In addition, it will be useful to compare how trained supervised algorithms perform against domain experts. Our research forms the first step in creating a context aware system for real-time privacy policy evaluation.

References

[1] M. Madden and L. Rainie, "Americans Attitudes About Privacy, Security and Surveillance," May 2015, accessed June 10, 2016. [Online]. Available: http://www.pewinternet.org/2015/05/20/americans-attitudes-about-privacy-security-and-surveillance/

[2] C. Jensen and C. Potts, "Privacy Policies as Decision-Making Tools: An Evaluation of Online Privacy Notices," *2004 Conference on Human Factors in Computing Systems*, vol. 6, no. 1, pp. 471–478, 2004.

[3] S. Winkler and S. Zeadally, "Privacy Policy Analysis of Popular Web Platforms," *IEEE Technology and Society Magazine*, vol. 35, no. 2, pp. 75–85, 2016.

[4] "Privacy Online: Fair Information Practices in the Electronic Marketplace," May 2000, accessed June 10, 2016. [Online]. Available: http://1.usa.gov/1XeBiuY

[5] "Council Regulation 2016/679 of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection

Regulation) [2016] OJ L119/I ," April 2016, accessed June 10, 2016. [Online]. Available: http://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:32016R0679&from=EN

[6] "Personal Information Protection and Electronic Documents Act," April 2000, accessed June 10, 2016. [Online]. Available: http://laws-lois.justice.gc.ca/eng/acts/P-8.6/index.html

[7] "Digital Privacy Act," June 2015, accessed June 10, 2016. [Online]. Available: http://laws-lois.justice.gc.ca/eng/annualstatutes/2015_32/page-1.html

[8] A. M. McDonald and L. F. Cranor, "Cost of reading privacy policies, the," *ISJLP*, vol. 4, p. 543, 2008.

[9] A. M. Mcdonald, R. W. Reeder, P. G. Kelley, and L. F. Cranor, *A comparative study of online privacy policies and formats*. Springer Berlin Heidelberg, 2009, pp. 37–55.

[10] G. R. Milne and M. J. Culnan, "Strategies for reducing online privacy risks: Why consumers read (or don't read) online privacy notices," *Journal of Interactive Marketing*, vol. 18, no. 3, pp. 15–29, 2004.

[11] L. F. Cranor, C. Hoke, P. G. Leon, and A. Au, "Are They Worth Reading-An In-Depth Analysis of Online Trackers' Privacy Policies," *I/S: A Journal of Law and Policy for the Information Society*, vol. 11, p. 325, 2015.

[12] S. Wilson, F. Schaub, A. A. Dara, F. Liu, S. Cherivirala, P. G. Leon, M. S. Andersen, S. Zimmeck, K. M. Sathyendra, N. C. Russell, T. B. Norton, E. Hovy, J. Reidenberg, and N. Sadeh, "The creation and analysis of a website privacy policy corpus," in *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, vol. 1, 2016, pp. 1330–1340.

[13] N. Sadeh, R. Acquisti, T. D. Breaux, L. F. Cranor, A. M. Mcdonalda, J. R. Reidenbergb, N. A. Smith, F. Liu, N. C. Russellb, F. Schaub, and et al., "The usable privacy policy project: Combining crowdsourcing, machine learning and natural language processing to semi-automatically answer those privacy questions users care about," 2013.

[14] Y. Sun, "Automatic Evaluation of Privacy Policy," 2012.

[15] L. F. Cranor, M. Langheinrich, and M. Marchiori, "A P3P preference exchange language 1.0 (APPEL1. 0)," *W3C working draft*, vol. 15, 2002.

[16] L. F. Cranor, "P3P: Making privacy policies more useful," *IEEE Security and Privacy*, vol. 1, no. 6, pp. 50–55, 2003.

[17] R. Lämmel and E. Pek, "Understanding privacy policies," *Empirical Software Engineering*, vol. 18, no. 2, pp. 310–374, April 2013.

[18] "Ten steps to develop a multilayered privacy notice," February 2006, accessed June 10, 2016. [Online]. Available: https://www.huntonprivacyblog.com/wp-content/uploads/sites/18/2012/07/Centre-10-Steps-to-Multilayered-Privacy-Notice.pdf

[19] M. Munur, S. Branam, and M. Mrkobrad, "Best Practices in Drafting Plain-Language and Layered Privacy Policies," September 2012, accessed June 10, 2016. [Online]. Available: https://iapp.org/news/a/2012-09-13-best-practices-in-drafting-plain-language-and-layered-privacy/

[20] P. Cimiano and J. Völker, "Text2Onto," in *Natural language processing and information systems*. Springer, 2005, pp. 227–238.

[21] P. Velardi, P. Fabriani, and M. Missikoff, "Using text processing techniques to automatically enrich a domain ontology," in *Proceedings of the international conference on Formal Ontology in Information Systems-Volume 2001*. ACM, 2001, pp. 270–284.

[22] R. Ramanath, F. Schaub, S. Wilson, F. Liu, N. Sadeh, and N. A. Smith, "Identifying relevant text fragments to help crowdsource privacy policy annotations," in *Second AAAI Conference on Human Computation and Crowdsourcing*, 2014.

[23] Y. Pan and G. M. Zinkhan, "Exploring the impact of online privacy disclosures on consumer trust," *Journal of Retailing*, vol. 82, no. 4, pp. 331–338, 2006.

[24] S. Wilson, F. Schaub, R. Ramanath, N. Sadeh, F. Liu, N. A. Smith, and F. Liu, "Crowdsourcing Annotations for Websites' Privacy Policies: Can It Really Work?" in *Proceedings of the 25th International Conference on World Wide Web*. International World Wide Web Conferences Steering Committee, 2016, pp. 133–143.

[25] K. M. Sathyendra, F. Schaub, S. Wilson, and N. Sadeh, "Automatic Extraction of Opt-Out Choices from Privacy Policies," in *2016 AAAI Fall Symposium Series*, 2016.

[26] S. N. Kim, O. Medelyan, M.-Y. Kan, and T. Baldwin, "SemEval-2010 Task 5: Automatic Keyphrase Extraction from Scientific Articles," in *Proceedings of the 5th International Workshop on Semantic Evaluation*, ser. SemEval '10. Stroudsburg, PA, USA: Association for Computational Linguistics, 2010, pp. 21–26.

[27] C.-m. A. Yeung, N. Gibbins, and N. Shadbolt, "A k-Nearest-Neighbour Method for Classifying Web Search Results with Data in Folksonomies," in *Proceedings of the 2008 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology - Volume 01*, ser. WI-IAT '08. Washington, DC, USA: IEEE Computer Society, 2008, pp. 70–76.

[28] S. Tan, "Neighbor-weighted K-nearest neighbor for unbalanced text corpus," *Expert Systems with Applications*, vol. 28, no. 4, pp. 667–671, 2005.

[29] E. Frank, G. W. Paynter, I. H. Witten, C. Gutwin, and C. G. Nevill-Manning, "Domain-specific keyphrase extraction," in *16th International Joint Conference on Artificial Intelligence (IJCAI 99)*, vol. 2. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 1999, pp. 668–673.

[30] P. D. Turney, "Learning algorithms for keyphrase extraction," *Information retrieval*, vol. 2, no. 4, pp. 303–336, 2000.

[31] K. Zhang, H. Xu, J. Tang, and J. Li, "Keyword extraction using support vector machine," in *Proceedings of the 7th International Conference on Advances in Web-Age Information Management*, ser. WAIM '06. Berlin, Heidelberg: Springer-Verlag, 2006, pp. 85–96.

[32] D. A. Audich, R. Dara, and B. Nonnecke, "Extracting keyword and keyphrase from online privacy policies," in *2016 Eleventh International Conference on Digital Information Management (ICDIM)*, Sept 2016, pp. 127–132.

[33] R. Mihalcea and P. Tarau, "TextRank: Bringing order into texts," in *Proceedings of the 2004 conference on empirical methods in natural language processing*. Association for Computational Linguistics, 2004.

[34] I. H. Witten, G. W. Paynter, E. Frank, C. Gutwin, and C. G. Nevill-Manning, "KEA: Practical automatic keyphrase extraction," in *Proceedings of the fourth ACM conference on Digital libraries*. ACM, 1999, pp. 254–255.

[35] S. N. Kim and M.-Y. Kan, "Re-examining automatic keyphrase extraction approaches in scientific articles," in *Proceedings of the workshop on multiword expressions: Identification, interpretation, disambiguation and applications*. Association for Computational Linguistics, 2009, pp. 9–16.

[36] S. Siddiqi and A. Sharan, "Keyword and Keyphrase Extraction Techniques: A Literature Review," *International Journal of Computer Applications*, vol. 109, no. 2, 2015.

[37] S. N. Kim and T. Baldwin, “Extracting keywords from multi-party live chats,” in *Proceedings of the 26th Pacific Asia Conference on Language, Information, and Computation*, 2012, pp. 199–208.

[38] M. Grineva, M. Grinev, and D. Lizorkin, “Extracting key terms from noisy and multitheme documents,” in *Proceedings of the 18th international conference on World wide web*. ACM, 2009, pp. 661–670.

[39] W.-t. Yih, J. Goodman, and V. R. Carvalho, “Finding advertising keywords on web pages,” in *Proceedings of the 15th international conference on World Wide Web*. ACM, 2006, pp. 213–222.

[40] K. S. Hasan and V. Ng, “Automatic Keyphrase Extraction: A Survey of the State of the Art.” in *ACL (1)*, 2014, pp. 1262–1273.

[41] D. D. Lewis, “Naive (Bayes) at forty: The independence assumption in information retrieval,” in *European conference on machine learning*. Springer, 1998, pp. 4–15.

[42] M. F. Porter, “An algorithm for suffix stripping,” *Program*, vol. 14, no. 3, pp. 130–137, 1980.