Unsupervised Topic Extraction from Privacy Policies

David Sarne
Bar-Ilan University
Ramat Gan, Israel
david.sarne@biu.ac.il

Jonathan Schler
Holon Institute of Technology
Holon, Israel
schler@hit.ac.il

Alon Singer
H-F & Co. Law Offices
Tel-Aviv, Israel
alon@h-f.co

Ayelet Sela
Bar-Ilan University
Ramat Gan, Israel
ayelet.sela@biu.ac.il

Ittai Bar Siman Tov
Bar-Ilan University
Ramat Gan, Israel
Ittai.Bar-Siman-Tov@biu.ac.il

ABSTRACT

This paper suggests the use of automatic topic modeling for large-scale corpora of privacy policies using unsupervised learning techniques. The advantages of using unsupervised learning for this task are numerous. The primary advantages include the ability to analyze any new corpus with a fraction of the effort required by supervised learning, the ability to study changes in topics of interest along time, and the ability to identify finer-grained topics of interest in these privacy policies. Based on general principles of document analysis we synthesize a cohesive framework for privacy policy topic modeling and apply it over a corpus of 4,982 privacy policies of mobile applications crawled from the Google Play Store. The results demonstrate that even with this relatively moderate-size corpus quite comprehensive insights can be attained regarding the focus and scope of current privacy policy documents. The topics extracted, their structure and the applicability of the unsupervised approach for that matter are validated through an extensive comparison to similar findings reported in prior work that uses supervised learning (which heavily depends on manual annotation of experts). The comparison suggests a substantial overlap between the topics found and those reported in prior work, and also unveils some new topics of interest.

CCS CONCEPTS

• **Information systems → Document topic models**; • **Theory of computation → Unsupervised learning and clustering**.

KEYWORDS

Topic modeling, unsuprevised learning, privacy policies

ACM Reference Format:
David Sarne, Jonathan Schler, Alon Singer, Ayelet Sela, and Ittai Bar Siman Tov. 2019. Unsupervised Topic Extraction from Privacy Policies. In *Companion Proceedings of the 2019 World Wide Web Conference (WWW'19 Companion), May 13–17, 2019, San Francisco, CA, USA.* ACM, New York, NY, USA, 6 pages. https://doi.org/10.1145/3308560.3317585

WWW '19 Companion, May 13–17, 2019, San Francisco, CA, USA

ACM ISBN 978-1-4503-6675-5/19/05.
https://doi.org/10.1145/3308560.3317585

1 INTRODUCTION

A Privacy policy is a legal document that details the ways a website or a mobile application processes, collects, stores and shares end users' information, including their personally identifiable information. Typically, it also notes the purposes of these practices and the rights of the end users in that regard. Research on privacy policies has drawn much attention over recent years, for two main reasons. First, despite the importance of these documents as binding legal agreements, they are often not carefully considered by users, primarily due to their length, level of detail and the legal language used, all making them difficult to understand [1, 18]. Notably, the application of simplification techniques to reduce ambiguity and complexity of privacy policies has been found to have little or no effect on consumers' understanding of privacy policies, their willingness to share personal data, and their expectations about their rights [2, 26]. Second, the recent EU General Data Protection Regulation (GDPR) fostered substantial revisions to privacy policies, leading to various debates about the role of regulation in this important area [7, 13]. As such, much research has been carried out with the goal of understanding the structure, content and evolution of these policies.

In line with the growing trend of using AI technologies in LegalTech, there have been many recent reported efforts to use machine learning and text mining methodologies for analyzing privacy policies, aiming primarily at extracting and mapping the specific topics they address [10, 14, 16, 27]. Still, to date, all attempts to use machine learning and other AI-based tools for analyzing privacy policies utilized supervised-learning based techniques, which extensively rely on tagged repositories hence requiring substantial human effort for annotating the data [6, 23, 28].

In this paper, we suggest the analysis of corpora of privacy policies using topic modeling–an unsupervised learning technique. To that end, we use a cohesive framework which includes the construction of a targeted corpus, division into segments of interest, clustering into topics, and grouping. The use of unsupervised learning encapsulates many advantages. First, it enables analyzing any new corpus with a fraction of the effort required by supervised learning, which is highly useful for understanding specific classes of policies (e.g., based on genre, country of origin, culture). Second, it enables comparing privacy policies over time (e.g., before and after GDPR came into effect) to reflect changes in the topics addressed, compliance with the new regulations and evolution trends in general.

We apply the topic modeling framework over a corpus of 4,982 privacy policies crawled from the Google Play Store. We extract the most prevalent 36 topics encompassed in the corpus, enabling a thorough analysis of the structure, scope and focus of current privacy policies. In particular, the distribution of paragraphs over the different topics provides insight into what privacy policy drafters believe matters in terms of user privacy.

Naturally, the use of unsupervised learning calls for strict validation, as it is possible that the list of extracted topics is merely a subset of the actual list, focusing on limited areas of interest that are particularly structured. Therefore, as a means of validation the list of topics extracted is compared to a taxonomy found in Wilson et al [27], listing alternative set of privacy-policy topics, which is of wide use nowadays (e.g., see Harkous et al [10]). The comparison is carried out through manual mapping of the 36 extracted topics to those found in the referenced list. The results reveal a substantial overlap between topics, suggesting that the unsupervised topic modeling according to the proposed framework is highly effective and leads to a division into topics quite similar to that obtained with supervised learning techniques. Furthermore, several topics in our list do not have equivalences in prior work, possibly indicating new topics of interest that have been missed by supervised learning. In addition, several topics map to the topic identified in prior work possibly unfolding different aspects of the topic or suggesting finer-grained sub-topics of a rather general matter.

2 RELATED WORK

In 2017, the Global Privacy Enforcement Network (GPEN) published a report that evaluates the extent to which users control their personal information [8]. Based on a manual assessment of privacy notices, communications and practices of 455 websites and apps from various sectors by 24 data protection regulators from around the world, the report concluded that "there is significant room for improvement in terms of specific details contained in privacy communications". In this context, the development of automated tools for assessing privacy policies is highly desirable, and indeed in very recent years a growing number of studies suggested using machine-learning-based solutions [10, 27, 28], as well as other AI-related methods [15] to address this challenge.

Liu et al [15] suggested the use of Hidden Markov models (HMM) for privacy policy segmentation and for identifying whether a given pair of privacy policy paragraphs discuss the same topic. They found that HMM based segmentation methods performed on par with the lower half of human evaluators.Most others implemented supervised text-mining techniques for analyzing privacy policies. The use of supervised techniques requires a highly qualified annotated dataset for model construction. Such a dataset became available with the publishing of the OPP-115 manually annotated privacy policies corpus in 2016 by Wilson et al [27]. The corpus includes 115 privacy policies, annotated by domain experts using a designated annotation tool that was developed for this purpose. The tool offers a taxonomy based on 22 topics for the annotation purpose. The OPP-115 nurtured several studies that used supervised machine learning techniques for privacy policy analysis, using the data supplied as part of the corpus as a reference for validation. For example, Zimmeck et al [28] analyze thousands of mobile applications privacy policies by applying classifiers based on the OPP-115 corpus, trying to point, inter alia, inconsistencies between a mobile application and its privacy policy. Sathyendra et al [23] make use of OPP-115 corpus for training and evaluation of models used for automatic detection of opt-out provisions in privacy policies. Harkous et al [10] use OPP-115 for introducing an automatic framework for privacy policies (Polisis). The framework is based on a hierarchy of neural network classifiers, trained using the corpus data.

Common to all the above studies is their reliance on supervised learning. As discussed earlier, it requires either tagged repositories or investment of substantial human effort to annotate the data. Furthermore, this effort needs to be reinvested whenever analyzing specific classes of policies or attempting to analyze privacy policies over time in order to identify changes in a topic of interest. To the best of our knowledge, the only attempt to use unsupervised learning based techniques for analyzing privacy policies is the work of Ramanath et al [22]. Still, that work does not aim at topic modeling per se. Instead its focus is on identifying the points of transition between topics in privacy policies. Furthermore, that work evaluates the success of the proposed method using the subjective opinion of a human evaluator rather than by comparison to the findings reported in prior work.

Finally, we note that the GPEN report mentioned above also offers a taxonomy of topics related to privacy policies, generated by experts. These, however, focus primarily on users' control over personal information and include only 7 fairly wide topics (also called indicators in the report). All topics suggested there (except for a topic titled "automatic decision") are included in the OPP-115 list. Hence we base our comparison on the latter.

3 UNSUPERVISED LEARNING & TOPIC MODELING

Topic models are a type of statistical methods at frequent use for the discovery of abstract topics in a corpus [5, 9, 20]. Their underlying assumption is that there are groups of words (with different distributions) in the corpus, such that each represents a topic. A document in the corpus, contains those topics in different distributions. Topic modeling, which is an unsupervised machine learning method, attempts to extract the most probable distribution of words into topics through an iterative generative process which terminates upon convergence. The process does not require prior labeling or annotation of the documents. This makes it useful in complex or large data sets, where it is too cumbersome or complicated to extract topics manually [4].

Early topic models were described by Papadimitriou et al [21], followed by Hofmann's [11] Probabilistic Latent Semtnatic Analysis (PLSA) and Blei et al's [3] Latent Dirichlet Allocation (LDA) methods. The latter, which we use in this paper, is commonly used both for topic modeling in various domains (e.g. politics [9], psychology [20] and law [5]), and as a foundation for new topic model algorithms [12]. LDA assumes a sparse Dirichlet prior distribution over document-topic and topic-word distributions. During its iterative training phase, the algorithm learns and refines these distributions parameters. The resulting distributions are reflective of the assumption that topics usually use a relative small group of words, and a document contains a relatively small number of topics.

4 FRAMEWORK & IMPLEMENTATION

While there are various available "off-the-shelf" tools for topic modeling based on a corpus (e.g., mallet [19], graphlab (turi.com), familia (github.com/baidu/Familia)), the analysis of privacy policies is not trivial and requires several complementary capabilities, primarily related to content acquisition, content cleaning, segmentation and post-processing. In this section we propose a general framework for topic extraction of privacy policies. In addition, we provide the details and choices made in a specific implementation carried out in order to evaluate the effectiveness of unsupervised topic modeling of privacy policies, based on this framework.

Figure 1 provides an overview of our proposed design for topic modeling of privacy policies. While many of the components of our framework are commonly found in work dealing with corpus analysis, the synthesis we suggest is tailored for the specific case of privacy policies.

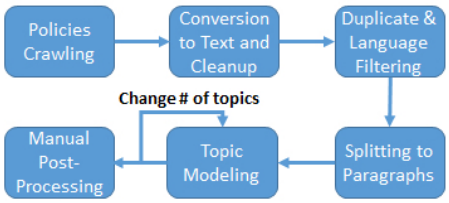

Figure 1: Overview of the topic modeling process.

Policies Crawling. To the best of our knowledge the only publicly available corpuses of privacy policies are the OPP-115 [27] and the ACL/COLING 2014 Dataset [15]. The rather small number of documents they contain (115 policies in OPP-115 and 1010 in the COLING Dataset) and the fact that the additional information they offer (i.e., labeling) is irrelevant for topic modeling using unsupervised learning makes them inappropriate for our purposes. Furthermore, any existing corpus is relevant to the time the policies it contains were collected. It may become obsolete as privacy policies are continuously being revised over time, especially in response to major regulatory changes, as in the case of GDPR [24]. Having the ability to construct new large and timely corpora is thus crucial. Here, crawling can be highly useful, as it offers full control over the number, properties and richness of the policies collected.

Taking the example of privacy policies of mobile apps, a new corpus can be constructed relatively easily, as a link to a relevant privacy policy document is provided as part of the information presented for any app in major app stores (e.g., Google Play, Apple's app store). Since these stores provide advanced search functionality, a corpus of a rather targeted set of policies can be constructed by automatically visiting the corresponding links and downloading the documents (that are stored on the app-developers' servers). Complementary meta-data related to the app itself (e.g., version, date, genre) can be downloaded as part of the process as it is stored in a rather structured manner in the HTML file.

For evaluation purposes we generated a corpus by crawling policy links from the Google Play Store and downloading the policies directly from the developers' servers. For diversity we selected policies that belong to 57 different Google Play categories. Overall, our crawler visited 24,071 app meta data pages during June 2018. Surprisingly, 19% of them did not have a privacy policy URL, leaving us with 19,458 URLs.

Conversion to Text, Cleanup and Filtering. Upon downloading a privacy policy, it should go through some pre-processing before added to the corpus. In many cases the first step is to convert the document into text (as it is formatted as an HTML file), removing tags, scripts and comments. This can be done using various off-the-shelf tools and libraries (e.g., Goose)[1]. Additional pre-processing include filtering of duplicate policies (as apps developed by the same company commonly use the same privacy policy), removal of policies in languages other than the intended one (e.g. by using NLTK or langdetect)[2] and removal of stop words (e.g. by using one of the stop words vocabularies commonly found in text mining tools such as GraphLab).

In our implementation, we removed 5,855 URL duplicates out of the 19,458 URLs crawled, leaving us with 13,603 URLs. We used the 'Goose' Python library to convert the documents to plain text and remove HTML meta-data. Further removal of text duplicates (identical privacy policies that were crawled from different URLs) and filtering out non-English language policies (using 'langdetect' Python library), led to a corpus of 4,982 documents, which served as the basis for our analysis.

Splitting to segments. Topic modeling requires the pre-processed policy to be divided into segments, each potentially dealing with a single topic. Here, one can use automatic tools for identifying points of transition between topics in privacy policies, such as the one suggested by Ramanath et al [22]. Alternatively, a common practice in topic modeling, which is particularly applicable for structured legal texts such as privacy policies, is to rely on the document's division into paragraphs as a means for segmentation [17]. The idea is to get to the smallest segments possible while keeping the essence of the text, bearing in mind that even if a topic is extracted based on paragraphs that are only part of a larger topic then later on it can be joined with topics representing other parts of the larger topic.[3] Furthermore, paragraphs can be easily identified and are generally seen as a self-contained unit of a discourse. Tools for dividing a document into paragraphs are abundant (e.g., NLTK's TextTiling Algorithm).[4]

In our implementation we preferred the division to paragraphs due to the above-mentioned advantages and the fact that the effectiveness of the alternative approach is somehow uncertain due to its reliance on subjective evaluation and the lack of comparison to an established set of topics. The process (carried out using TextTiling) resulted in 45,622 paragraphs. The average paragraph size was found to be 304 words, with standard deviation of 274 words and median of 248 words. From the latter set, we removed stop words using GraphLab's text analytics dictionary.

Topic Modeling and Assignment. Topic modeling requires to predefine the number of topics to be extracted. Trial and error may be required in order to set an effective number of topics, that will yield a set of cohesive and meaningful topics. The completeness and effectiveness of the set extracted can be measured in terms of the stylized statistics related to the division of paragraphs into

[1] github.com/GravityLabs/goose
[2] nltk.org/, github.com/Mimino666/langdetect.
[3] In fact, as demonstrated onwards, having several topics which are part of a larger topic enables finer-grained analysis.
[4] nltk.org/_modules/nltk/tokenize/texttiling.html

topics (e.g., mean, minimum, maximum, standard deviation), the distribution of confidence scores assigned to paragraphs within each topic (provided by the topic modeling tool), and possibly a comparison to labeled/existing data sets to validate the results. We note that while a large set of topics is likely to catch subtler aspects of privacy policies, it requires substantially greater amount of manual work for verification, merging and the construction of a topic hierarchy in order to identify higher level areas of interest in later stages.

In our implementation the topic modeling was carried out using GraphLab, a machine learning platform that supports LDA-based topic modeling. We set the number of topics to 100, as on the one hand it is large enough to represent the diversity of topics in a balanced way and on the other hand small enough to ensure that the manual merging of closely related topics is not overly tedious.

After completing the training phase, which took 600 iterations to converge, the LDA created clusters of words in each topic, as well as word probabilities in relation to each topic. Those probabilities were used to assign the most likely topic to each paragraph, by summing the word weights for all the paragraph words per topic, and choosing the topic that obtained the highest score.

Final Manual Processing. Topics obtained through topic modeling are identified by a collection of words. Typically, it is impossible to identify the substantive topic of each word collection by merely reviewing the words. A domain expert can thus be highly valuable, as she can sample some of the paragraphs assigned to each topic and identify its essence. She can also eliminate a topic if it is not cohesive. Topics that survive the elimination process can then be used for (manually) constructing a topics hierarchy, possibly pointing to higher level topics (areas of interest).

In our implementation we recruited as a domain expert a lawyer with 4-years of experience in privacy policies and commercial law. The expert went over each topic, reviewing a sample of at least 30 paragraphs, where paragraphs sampling from the ranked list (according to the confidence score assigned to each paragraph within the topic) followed the Padovan series [25], which provides a decent tradeoff between relevance (score) and sparsity. Based on the paragraphs sampled for each topic the expert provided a one sentence summary of the topic. 18 topics did not survive this process and were declared non-cohesive, leaving us with 82 topics.

Since the LDA model is designed in a manner that the distribution of the privacy policy paragraphs between the LDA model topics is uniform, some of the LDA models' topics cover legal issues that are identical. For example, in our case seven of the topics referred to personal information tracking technologies, using different forms and phrases or simply emphasizing different aspects of the same topic. Ideally, these paragraphs should be joined as they all apply to the same topic (legal issue) and they are not substantively distinguished. Still, the paragraphs are relatively diverse in terms of the words and structure used, such that even reducing the number of requested topics in the topic modeling process would not have grouped them all in a single topic. Therefore, whenever applicable, an expert should merge topics through manual processing. In our specific implementation the expert's merging process resulted in a total of 36 topics, each distinctly identified as dealing with one major legal issue.

5 ANALYSIS AND TOPICS VALIDATION

Table 1 lists the 36 topics that result from the expert's manual merger (third column). The table also specifies the number of topics that were merged into each topic (out of the original 82 topics) and the number of paragraphs that mapped to each topic (fourth and fifth columns, respectively). From the table we observe a substantial variation in the scope and level of detail of topics. Some topics are the result of merging several different original topics (e.g., the specification of the tracking technologies used for collecting personal information and the definition of what is personal information for that sake, each comprises seven original sub-topics) and have thousands of paragraphs mapped to them, possibly indicating the richness of the different ways in which they are formulated as well as the high level of detail they contain. However, most topics comprise a smaller number of original sub-topics and thus seem to show less variation in their formulation and possibly a narrower scope. Notably, three topics are exceptional in that they comprise only one or two of the original sub-topics and yet the number of paragraphs mapped to them is substantial: privacy policy update notification, contact information regarding personal information matters and user's option to opt-out of the privacy policy. This finding suggest that these are highly structured and standard topics that are commonly used.

Table 1 also maps each of the 36 topics in our list to the topics suggested by Wilson et al, which as discussed earlier were commonly used in prior work studies, thus enabling evaluating the validity of the topics mined with our framework.[5] The taxonomy used in Wilson et al's annotation system suggests (see also Figure 3 in Harkous et al [10]) a hierarchic presentation of topics where each topic in the top level of the hierarchy defines a high-level privacy category (first column in Table 1). The lower level (second column in Table 1) defines a set of privacy topics, with 22 such topics overall.

Notably, there is a substantial overlap in topics between the OPP-115 topics and those modeled in our implementation: topics in our list mapped into 17 out of the 22 topics listed in Wilson et al, which can be considered an important validation of our framework. Several of Wilson et al's topics are mapped by more than one topic from our list (four topics were mapped to two topics each and one to three topics from our list), possibly indicating the extraction of finer-grained topics through topic modeling. For example, within the topic of information type collected by 1st party, our topics distinguish between general information collected and financial personal information collected. As another example, the OPP-115 topic related to audience groups maps in our list into two topics that distinguish between the general population and the specific population of kids with respect to handling of personal information, a distinction that can be particularly important.

Five topics from the Wilson et al list are not present in our list (marked with grey background in the second column). These topics are related to information types collected by 3rd party, scope of access (for reviewing, editing, deleting, etc.), information type of data retained, user choice related to policy change and practices that are not covered in the policy. We note that each of those belongs to a different area and all other sibling topics of that level are adequately

[5] The mapping was carried out manually by the domain expert.

Table 1: Mapping of topics to those suggested by Wilson et al [27].

OPP-115 Category	OPP-115 Topic	Topics in Our Unsupervised-Learning Implementation	Merged	Paragraphs
1st party collection	Collection mode	Personal information tracking technologies	7	3674
		Personal information collection methods	2	890
	Information type	Collected personal information description	3	1237
		Collected financial personal information description	1	655
	Purpose	Personal information collection purposes	4	1706
3rd party collection	Action	Personal information shared with 3rd Parties	6	2915
		Personal information collection directly by 3rd parties through app	2	584
	Information type	-	-	-
	Purpose	Forced disclosure of personal information (e.g. court order)	3	1366
	-	Disclosure of personal information by user directly to 3rd parties	1	571
Access. Edit, Delete	Access scope	-	-	-
	Access rights	Users right to view/update/delete their personal information	5	2610
Data Retention	Retention period	Personal information retention period	1	456
	Retention purpose	Purpose of personal information collection	1	429
	Information type	-	-	-
	-	Collected technical related personal information description	2	1221
		Collected geographical data	1	593
		App does not collect personal information	1	507
		Implication of granting permission to app to collect personal information	1	468
Data Security	Security measure	How personal information is protected	1	291
Int. and Specific Audience	Audience group	Personal information regulations	4	1801
		Geographical transfer of personal information	3	1609
		Personal information of kids	1	885
Do Not Track	Do not track policy	Do Not Track	1	321
Policy Change	Change type	Privacy policy updates	1	322
	User choice	-	-	-
	Notification type	Privacy policy update notifications	1	1071
Other	Introductory	What is personal information	7	3408
	Contact information	Contact information regarding personal information matters	1	1013
	Practice note covered	-	-	-
	-	Paragraphs that are irrelevant to privacy	3	1143
		General legal provisions (e.g. Limitation of liability/governing Law/jurisdiction/disclaimers)	5	2457
		Collected non-personal information	4	1150
		Details about creation of an account and related password	1	440
		App developer representations regarding compliance with personal information regulations	1	335
Choice Control	Choice type	Clarification regarding user's own decision to share their personal information	1	120
		User option to opt-out	2	1585
	Choice scope	Notice that user's acceptance of personal information processing can be withdrawn	1	534
	-	Sharing of personal information by user and related risks	1	667
		Notice about user's explicit grant of right to use their personal information by the app	1	597
		Warnings from unwanted disclosure of personal information	1	443

mapped. Thus, the absence of the five topics from our list does not result in missing a complete area of interest, but rather in lack of specification of a certain aspect that possibly exhibits high variation in the way it is formulated in privacy policies. Interestingly, there is no correlation between the number of paragraphs mapped into a topic (hence possibly its importance or complexity) in our list and its appearance or absence in the other list.

On the other hand, 13 topics from our list do not map to any of the topics in Wilson et al (marked with grey background in the third column). Some of these topics do not directly address privacy issues (e.g., Collected non-personal information, General legal provisions (e.g. Limitation of liability/governing Law/jurisdiction)), yet others certainly unfold new considerations and aspects that have been missed in prior work, e.g., risks related to sharing personal information by the user, privacy policy updates, and disclosure of personal information by the user directly to 3rd parties. Eight of the specific topics we found that do not map to areas of interest in Wilson et al's taxonomy, suggest new topics of interest within the specific area, and five do not map to an existing area, although they do not appear to warrant a new area of interest of their own.

Finally, we note that our topic-modeling-based analysis yields insights related to the importance and weight of the different levels of hierarchies (areas) suggested in prior work as well as to the extent of details they exhibit—the greater the number of topics associated with each category and the number of paragraphs mapped to them, the greater the breadth of aspects the category comprises and the level of detail or complexity it entails. These notions are illustrated in Table 2 through a comparison to the set of expert-derived areas of interest provided in Wilson et al. The table depicts the percentage of paragraphs mapped into each area of interest out of the total number of paragraphs in our corpus (second column), based on the mapping given in Table 1. The third column is the percentage of annotations found in the OPP-115 corpus mapped to each area out of the total number of annotations there (see Table 2 in Wilson et al).

Table 2: OPP-115 [27] topics and the number of topics from our analysis corresponding to each.

OPP-115 Category	% of Paragraphs	% annotations in OPP-115
Other	23.0%	15.4%
1st Party Collection	18.7%	38.5%
3rd Party Collection	13.6%	22.5%
International and Specific Audience	12.4%	4%
Data Retention	10.9%	1.6%
Choice Control	9.9%	7.7%
Access, Edit, Delete	6.5%	3.2%
Policy Change	3.5%	2.4%
Do Not Track	0.8%	0.4%
Data Security	0.7%	4.3%

We learn from the table that even though the distribution of segments into topic areas is somewhat different, in general areas that received greater weight in the manual annotations of OPP-115 were identified as the major areas of interest also by our unsupervised topic modeling. With respect to the most minor areas of interest, the two methods provide almost the same weight assessment. A

substantial difference with the major-weight areas derives from the difference in the "Other" topic, arguably due to the new topics we found in our topic-modeling based analysis, which were naturally assigned into this category. Hence we are able to provide a good estimation of the weight and breadth of different privacy-policies-related categories without the extensive human labor that is required for labeling.

6 DISCUSSION AND CONCLUSIONS

The encouraging results related to our framework validation reported in the previous section suggests that topic modeling based on unsupervised learning can be highly effective in extracting topics of interest of privacy policy corpuses. The implications for regulators are many, as discussed throughout the paper. This method provides a means for monitoring the evolvement of privacy policies over time and for reflecting changes in their content, in terms of the topics being addressed, given a specific set of regulations or following a regulatory change.

To the best of our knowledge, our work is the first to propose, implement and evaluate an unsupervised topic modeling framework for analyzing privacy policies topics of interest. In addition to validating the findings the comparison to previous work provides a somewhat finer-grained list of topics under each high-level area of interest. Additionally, we find a set of topics that were not proposed through previous applications of supervised analysis. One possible explanation for the newly added topics is the recent coming into effect of GDPR. It points to the importance of the proposed unsupervised method and highlights its usefulness and contribution in an ever changing regulatory environment.

While unsupervised learning does not require the human labor necessary for annotating the data in supervised learning, it does require some manual post-processing by a human expert to verify the coherence of the topics mined, summarizing the content of the different topics and merging them whenever applicable. We emphasize that this human effort is several orders of magnitude smaller than the effort required for annotating data for supervised methods.[6]

We hope that the success of the specific implementation reported in the paper and the public availability of the corpus (to which we keep adding policies and newer versions of existing policies for purposes of comparative analysis) will facilitate further insights and contribution. Other directions for future work include the analysis of differences in privacy policies across domains and over time. In addition, we plan to evaluate advanced methods for document segmentation, beyond current paragraph-based approach, that will better reflect changes in the topic discussed.

ACKNOWLEDGMENTS

This research was partially supported by the ISRAEL SCIENCE FOUNDATION grant No. 1162/17.

REFERENCES

[1] Yannis Bakos, Florencia Marotta-Wurgler, and David R Trossen. 2014. Does anyone read the fine print? Consumer attention to standard-form contracts. *The Journal of Legal Studies* 43, 1 (2014), 1–35.

[2] Omri Ben-Shahar and Adam Chilton. 2016. Simplification of privacy disclosures: an experimental test. *The Journal of Legal Studies* 45, S2 (2016), S41–S67.

[3] David M Blei, Andrew Y Ng, and Michael I Jordan. 2003. Latent dirichlet allocation. *Journal of machine Learning research* 3 (2003), 993–1022.

[4] Zhiyuan Chen and Bing Liu. 2014. Topic modeling using topics from many domains, lifelong learning and big data. In *International Conference on Machine Learning*. 703–711.

[5] Hyo Shin Choi, Won Sang Lee, and So Young Sohn. 2017. Analyzing research trends in personal information privacy using topic modeling. *Computers & Security* 67 (2017), 244–253.

[6] Elisa Costante, Yuanhao Sun, Milan Petković, and Jerry den Hartog. 2012. A machine learning solution to assess privacy policy completeness. In *Proc. of WPES*. 91–96.

[7] M. Degeling, C. Utz, C. Lentzsch, H. Hosseini, F. Schaub, and T. Holz. 2018. We Value Your Privacy ... Now Take Some Cookies: Measuring the GDPR's Impact on Web Privacy. *ArXiv e-prints* (2018).

[8] GPEN. 2017. GPEN Sweep 2017 - User Controls over Personal information.

[9] Derek Greene and James P Cross. 2017. Exploring the political agenda of the European parliament using a dynamic topic modeling approach. *Political Analysis* 25, 1 (2017).

[10] Hamza Harkous, Kassem Fawaz, Rémi Lebret, Florian Schaub, Kang G. Shin, and Karl Aberer. 2018. Polisis: Automated Analysis and Presentation of Privacy Policies Using Deep Learning. In *27th USENIX Security Symposium (USENIX Security 18)*. USENIX Association, 531–548. https://www.usenix.org/conference/usenixsecurity18/presentation/harkous

[11] Thomas Hofmann. 1999. Probabilistic latent semantic analysis. In *Proceedings of UAI*. 289–296.

[12] Wei Li and Andrew McCallum. 2006. Pachinko allocation: DAG-structured mixture models of topic correlations. In *Proceedings of the 23rd international conference on Machine learning*. 577–584.

[13] T. Linden, H. Harkous, and K. Fawaz. 2018. The Privacy Policy Landscape After the GDPR. *ArXiv e-prints* (Sept. 2018). arXiv:cs.CR/1809.08396

[14] Fei Liu, Nicole Lee Fella, and Kexin Liao. 2016. Modeling language vagueness in privacy policies using deep neural networks. In *AAAI Fall Symposium on Privacy and Language Technologies*.

[15] Fei Liu, Rohan Ramanath, Norman M. Sadeh, and Noah A. Smith. 2014. A Step Towards Usable Privacy Policy: Automatic Alignment of Privacy Statements. In *COLING*. ACL, 884–894.

[16] Frederick Liu, Shomir Wilson, Peter Story, Sebastian Zimmeck, and Norman Sadeh. 2018. Towards Automatic Classification of Privacy Policy Text. CMU-ISR-17-118R, CMU-LTI-17-010 (June 2018).

[17] Yue Lu and Chengxiang Zhai. 2008. Opinion Integration Through Semi-supervised Topic Modeling. In *Proc. of WWW*. 121–130. https://doi.org/10.1145/1367497.1367514

[18] Florencia Marotta-Wurgler. 2012. Does Contract Disclosure Matter? *JITE* 168, 1 (2012), 94–119.

[19] Andrew Kachites McCallum. 2002. MALLET: A Machine Learning for Language Toolkit. (2002). http://mallet.cs.umass.edu.

[20] Kate Niederhoffer, Jonathan Schler, Patrick Crutchley, Kate Loveys, and Glen Coppersmith. 2017. In your wildest dreams: the language and psychological features of dreams. In *Proc. of CLPsych*. 13–25.

[21] Christos H. Papadimitriou, Hisao Tamaki, Prabhakar Raghavan, and Santosh Vempala. 1998. Latent Semantic Indexing: A Probabilistic Analysis. In *Proc. of PODS*. 159–168. https://doi.org/10.1145/275487.275505

[22] Rohan Ramanath, Fei Liu, Norman M. Sadeh, and Noah A. Smith. 2014. Unsupervised Alignment of Privacy Policies using Hidden Markov Models. In *ACL (2)*. 605–610.

[23] Kanthashree Mysore Sathyendra, Shomir Wilson, Florian Schaub, Sebastian Zimmeck, and Norman Sadeh. 2017. Identifying the Provision of Choices in Privacy Policy Text. In *Proc. of EMNLP*. 2764–2769.

[24] Yan Shvartzshnaider, Noah Apthorpe, Nick Feamster, and Helen Nissenbaum. 2018. Analyzing Privacy Policies Using Contextual Integrity Annotations. *arXiv preprint arXiv:1809.02236* (2018).

[25] I. Stewart. 1996. Tales of a Neglected Number. *Scientific American* 274 (June 1996), 102–103. https://doi.org/10.1038/scientificamerican0696-102

[26] Lior Jacob Strahilevitz and Matthew B Kugler. 2016. Is Privacy Policy Language Irrelevant to Consumers? *The Journal of Legal Studies* 45, S2 (2016), S69–S95.

[27] Shomir Wilson, Florian Schaub, Aswarth Abhilash Dara, Frederick Liu, Sushain Cherivirala, Pedro Giovanni Leon, Mads Schaarup Andersen, Sebastian Zimmeck, Kanthashree Mysore Sathyendra, N. Cameron Russell, Thomas B. Norton, Eduard Hovy, Joel Reidenberg, and Norman Sadeh. 2016. The Creation and Analysis of a Website Privacy Policy Corpus. In *Proc. of ACL*. 1330–1340. https://doi.org/10.18653/v1/P16-1126

[28] Sebastian Zimmeck, Lieyong Zou Ziqi Wang, Bin Liu Roger Iyengar, Florian Schaub, Shomir Wilson, Norman Sadeh, Steven M. Bellovin, and Joel Reidenberg. 2017. Automated analysis of privacy requirements for mobile apps. In *Proc. of NDSS*.

[6] For example, the OPP-115 corpus includes 23,194 annotated data practices, 128,347 annotated attributes and 102,576 annotated text spans 102,576, all performed manually.