# Semi-Automated Analysis of Large Privacy Policy Corpora

Alden Dima
National Institute of Standards and Technology
alden.dima@nist.gov

Aaron Massey
University of Maryland Baltimore County
akmassey@umbc.edu

## Abstract

*Regulators, policy makers, and consumers are interested in proactively identifying services with acceptable or compliant data use policies, privacy policies, and terms of service. Academic requirements engineering researchers and legal scholars have developed qualitative, manual approaches to conducting requirements analysis of policy documents to identify concerns and compare services against preferences or standards. In this research, we develop and present an approach to conducting large-scale, qualitative, prospective analyses of policy documents with respect to the wide-variety of normative concerns found in policy documents. Our approach uses techniques from natural language processing, including topic modeling and summarization. We evaluate our approach in an exploratory case study that attempts to replicate a manual legal analysis of roughly 200 privacy policies from seven domains in a semi-automated fashion at a larger scale. Our findings suggest that this approach is promising for some concerns.*

## 1. Introduction

Privacy policies support commerce by informing potential customers of business practices involving their data. The U.S. Federal Trade Commission (FTC) investigates incongruities between stated and actual business practices. Mishandling of privacy can be costly as two recent cases highlight. In July 2019, Equifax agreed to pay at least $650 million for a data breach that affected some 150 million people [1]. That month, Facebook was fined $5 billion for violating customer privacy and was forced to revamp its privacy protection practices [2].

These cases highlight the reactive nature of investigations, which typically begin only after a complaint is received. There is a need for rapid analysis of many privacy policy documents from websites across multiple industries [3]. Unfortunately, proactively evaluating privacy practices is tedious and challenging. These analyses often require significant manual effort. For example, in 2016, Marotta-Wurgler (Section 3) published a study of 261 privacy policies whose goal was to measure online site compliance with self-regulatory guidelines [4]. A team of nine investigators hand coded the presence of 49 recommended practices [4]. With over one billion web servers [5], manual analyses cannot scale to the Web.

We propose a semi-automated method based on the natural language processing of a large privacy policy corpus that relies on a manual analysis of a smaller subset. We wish to determine the extent that we can automate large-scale analyses using a sentence similarity baseline method. Using a corpus of two thousand documents, we will focus on the notice-related concerns from the Marotta-Wurgler study. Our research questions are:

**RQ1:** Which of the Marotta-Wurgler notice-related privacy concerns can be identified from our corpus using a sentence similarity-based approach?

**RQ2:** How well does our similarity-based approach identify privacy concerns in our policy corpus?

Our results indicate that semi-automated analysis can indeed allow for future work to target much larger corpora. We were able to identify relevant sentences containing keywords for 9 of the 21 notice-related concerns identified by Marotta-Wurgler with an average sensitivity and specificity of 93.8 % and 98.3 %, respectively. When combined with other concerns whose sentences were identified with higher average specificity (97.8 %.) but lower average sensitivity (35.6 %), sentences from 17 of the 21 notice-related privacy concerns were identified. This suggests that, while not all relevant sentences can be found, most of those found will be.

Our work proceeds as follows. In Section 2, we survey concepts and techniques necessary for this work. Section 3 describes two prior analyses of privacy policy corpora. In Sections 4, and 5, respectively, we describe our methodology and evaluation process, and present the results of evaluating our method with a corpus of 2061 policies. We continue with a discussion our results and the limitations of our approach in Sections 6 and 7, respectively, and conclude with a summary of our findings in Section 8.

URI: https://hdl.handle.net/10125/71180
978-0-9981331-4-0

## 2. Background

In this section, we describe techniques and concepts that are related to this study. We begin by discussing extractive text summarization, followed by vector space models. We then discuss semantic similarities and conclude with the total recall problem.

**Extractive Text Summarization** Our approach depends on the identification of sentences similar to manually-selected exemplar sentences relevant to a privacy concern. We use extractive text summarization to find candidate exemplar sentences likely to match the greatest number of relevant sentences with the same keywords. A few of the highest-ranking sentences can serve as a summary. We chose LexRank, an extractive text summarization algorithm that uses similarity graphs to connect sentences whose cosine similarities exceed a threshold [6].

**Vector Space Models** In order to identify similar sentences, we must first find a suitable computer representation to encode their features. Text inherently lends itself to sparse representations requiring large amounts of computer memory and computing time [7, 8]. For some of our work, we addressed this via a dimensionality reduction technique, latent semantic indexing (LSI), which uses a singular value decomposition to project the words into a smaller concept space [9, 10, 11, 12]. As an added benefit, Words that appear in similar contexts are projected into the same concepts so that documents can be close to each other without sharing common words [11].

Newer techniques such as word2vec [13], GloVe [14], Elmo [15], and BERT [16] use neural networks to achieve dense representations. These techniques are more sensitive to the surrounding word context and give additional semantic information. We chose a preexisting BERT embedding for a portion of this work and compared the results obtained with those from using LSI.

**Semantic Similarity** Once we identified the exemplar sentences and calculated the sentence features, we needed a means to identify sentences that are similar to the exemplars. We used cosine similarity, which treats the word features as vectors and uses the angle between two vectors; the greater the angle, the less similar the words [17].

**The Total Recall Problem** The semi-automated analysis of privacy policies shares aspects with the Total Recall Problem (TPC), an information retrieval problem where only a small fraction of a population is relevant and each member of that population can be inspected [18, 19]. The goal is to reduce the cost of high recall with a human in the loop [18, 19]. The best strategy for language-based TPC is to apply active learning with vector-based features to a manually-labeled subset and then use the trained system to prioritize the remaining items [19]. A challenge is knowing when to stop the iterative process of manually labeling positive examples and use the trained model to predict the unlabeled ones [19].

Our baseline method is similar to the active learning approach. For example, we use vector-based features and identify positive examples (the exemplar sentences). However, our method does not share the active learning's iterative nature. We evaluate the resulting similarity-based classifiers on a subset of the data, but do not iteratively train them. The active learning approach to the Total Recall Problem may serve as a useful strategy for situations where we are unable to identify sentences with high sensitivity using our method.

## 3. Related Work

In this section, we will survey four prior studies as well a deep-learning based analysis framework.

**RE 2013 Study and Corpus** To allow us to generalize our results, our exploration required a sufficiently large and diverse corpus. We chose a corpus of 2061 privacy policy and related documents drawn from prior privacy document analysis work, the Google top 1000 websites, and the 2012 Fortune 500 companies [3]. It was created to determine if automated text mining can help requirements engineers determine whether a policy document contains requirements expressed as either privacy protections or vulnerabilities. Massey et al. used topic modeling to identify documents addressing concerns expressed via goals-based requirements engineering and demonstrated the ability to limit searches of the entire corpus to less than 100 documents for certain goal keywords [3].

**Marotta-Wurgler Study** We benefited from the privacy concerns identified in Marotta-Wurgler's 2016 study [4] of 261 privacy policies taken from seven online markets [4]. She found that the average policy only complies with 39 % of the 2012 FTC guidelines [4]. The policies were often silent in required or important areas. This causes problems because there are often no default rules to address these areas. Many companies collect information and consumers have no way of knowing how much is collected, how it is used, how long it is kept, or with whom it is shared. The guidelines do not seem to be driving policy development and the idea of companies "competing on privacy" appears to be flawed [4].

**Polisis** The challenges of analyzing privacy policies at scale led Harkous et al. to create Polisys, deep learning based policy query answering framework that uses convolutional neural networks classifiers, a privacy taxonomy, and a custom language model created from 130,000 privacy policies for state-of-the-art results for structured and free-form querying of privacy policies [20]. Polisys relies on an embedded privacy taxonomy, which can limit its use. Harkous et al. suggest that this can be mitigated via additional training using another annotated data set.

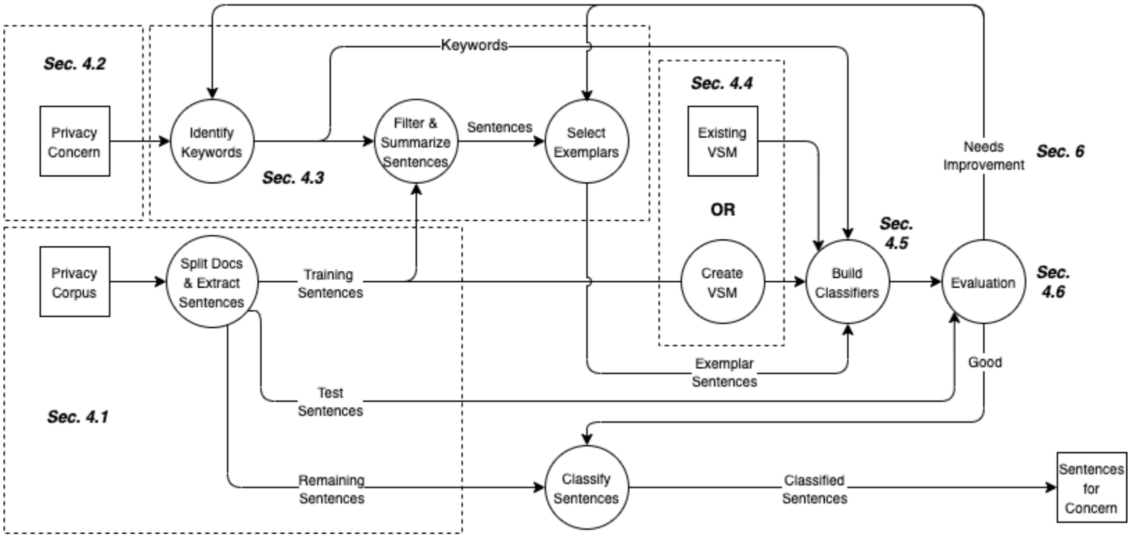

**Figure 1:** Data flow diagram of the semi-automated methodology for identifying sentences related to a privacy concern described in Section 4. Portions of the diagram are labeled with the section or subsection which describes them.

We believe that the method described in Section 4 can be used as part of a process to create such a data set.

**GDPR Impact Studies** Two recent studies attempted to ascertain the impact of the European Union's (EU) General Data Protection Regulation (GDPR) on privacy policies. The first by Degeling et al., used a combination of automated and manual methods to routinely inspect some 6,579 websites across the EU for evidence of changes due to the enactment of the GDPR [21]. They identified and downloaded some 112,041 privacy policies into a database, extracted the individual sentences, identified changes with hash functions and analyzed sentences for occurrences of GDPR-related phrases. They found that most EU websites had made changes to address the GDPR, though not all new requirements were being met.

Linden et al. also addressed the impact of the GDPR on privacy policies. They collected 6,278 pairs of pre- and post-GDPR versions of English-language privacy policies and examined them for changes in five areas: visual presentation, syntactic text features, category-based coverage, compliance to a small set of requirements, and the specificity of certain privacy practices [22]. They identified the English-language privacy policies using a convolutional neural network and made use of Polisis [20] to automatically label text segments. They conclude that the GDPR has driven many recent changes in privacy policies, particularly in the EU [22].

## 4. Methodology

Our ultimate goal is to enable the future analysis of large Web-based privacy policy corpora to determine the extent to which specific privacy concerns are being addressed. While previous studies have used manual methods on smaller corpora, they will be intractable for our future intended corpora. We plan to leverage a semi-automated analysis of a small subset of a large corpus to perform an automated analysis of the rest. Our approach requires dividing a large corpus into two portions: a smaller subset of the privacy policies amenable to combined manual and automated analyses and the much larger remainder which can only be analyzed using automated methods. Our goal for this work was to explore whether this approach is feasible and to identify potential issues for future research. The methodology described below was intended to represent a minimally viable technique and serves to establish a baseline for more refined approaches. We used the following criteria to answer our research questions:

**RQ1 Measures:** We determined which of the notice-related privacy concerns could be identified using our similarity-based approach while

(a) *Excluding* privacy concerns that address external language or entities outside of a privacy policy.

(b) *Excluding* privacy concerns that have no identifiable keywords and require human interpretation to determine whether they have been addressed.

(c) *Excluding* privacy concerns which do not have both exemplar sentences containing relevant keywords and higher sensitivity and specificity similarity-based classification results with our test corpus.

**RQ2 Measures:** We determined how well our similarity-based approach identified notice-related concerns by clustering the classification results and identifying the different cases as described in Section 4.6.

**Table 1:** Notice-related concerns from Marotta-Wurgler's study. Privacy concerns with a † are those for which we are able to identify sentences using our method (Section 5).

| Concern | Description | Concern | Description |
|---|---|---|---|
| N1 | Policy is accessible through direct link from the homepage | N12 | PII used internally for business purposes |
| N2 | Users asked for consent when signing up via clickwrap | N13 | PII used for stated, context-specific purposes |
| N3 | Layered or short notice is collected and stored | N14 † | Profile, picture, or other data may be used in ads |
| N4 | Contact data is collected and stored | N15 † | Third party may place ads that track user behavior |
| N5 † | Computer data is collected and stored | N16 | Recipients of shared or sold data are identified |
| N6 † | Interactive data is collected and stored | N17 | Words such as "affiliates" are defined, if used |
| N7 † | Financial information is collected and stored | N18 † | Company alerts user to material changes in policy |
| N8 | Content is collected and stored | N19 † | User must explicitly assent to material changes |
| N9 † | Sensitive information is collected and stored | N20 † | Material changes are retroactive |
| N10 | Geolocation information is collected and stored | N21 | Describes data procedures if company is sold or closes |
| N11 | Cookies used | | |

We quantified the overall extent of these cases with the fraction of the test corpus sentences containing relevant keywords associated with the results for each case.

Our methodology, which is illustrated in Fig. 1 and described below in Sections 4.1 through 4.6, began with the selection of the small corpus and its division into training and test sets (Section 4.1). We then chose the privacy concerns to study (Section 4.2), identified their exemplar sentences from the training set (Section 4.3), created a vector space model from the training set sentences (Section 4.4), and built similarity-based sentence classifiers (Section 4.5). We ended with the evaluation of the classifiers against the test set in Section 4.6.

## 4.1. Creation of Training and Test Sets

We began with a corpus consisting of several thousand privacy policies, the RE 2013 corpus discussed in Section 3. For the purposes of this work, this corpus will serve in the place of a smaller subset of a much larger corpus. It has been the subject of prior analyses and is well understood. We randomly divided this corpus into training and test sets using a 66/33 % split. We continued by segmenting each document of the training and test sets into individual sentences. We treated these sentences as individual documents when we searched the corpus for similar sentences to build our classifiers. These steps are depicted in the lower left portion of Fig. 1.

## 4.2. Identification of Privacy Concerns

The upper left corner of Fig. 1 depicts privacy concern identification. We focused on the set of 21 notice-related concerns from Marotta-Wurgler [4] listed in Table 1. The first three describe the online prominence of the policy (N1, N2, and N3). Subsequent concerns focus on the user data collected and stored (N4 through N10), the use of cookies (N11), how personal and personally identifiable information is used (N12 through N14), third parties (N15 through N17), the handling of material changes (N18 through N20), and data procedures if the site is sold or ceases to exist (N21).

Notice is central to the "notice and choice" privacy model espoused by the FTC which encourages the creation of privacy policies focusing on data collection [4]. These concerns also form the largest subset of the 49 in Marotta-Wurgler's study and we believed that their concrete nature would make them amenable to our approach.

## 4.3. Identification of Exemplar Sentences

In this and the following two subsections (Sections 4.4 and 4.5), we will use the privacy concern N19 ("User must explicitly assent to material changes") as the basis of a running example to explain our methodology.

We began by identifying keywords for each privacy concern (Fig. 1, upper right). For privacy concern N19, we chose "you", "opt-in", and "change" as one set of possible keywords. These seemed reasonable based on the results of text searches of the privacy corpus for "material changes."

We then filtered the training set using the stemmed keywords. The N19 keywords and sentences in the training set were stemmed. For example, the sentence describing N19:

> *You will be given the choice at that time to "opt-in" for any additional uses or disclosures of your personally identifying information and/or health-related personal information that you made available to us prior to the change in the Privacy Policy.*

became

> *you will be given the choic at that time to opt in for ani addit use or disclosur of your person identifi inform and/or health relat person inform that you made avail to us prior to the chang in the privaci polici*

Stemming simplifies text searches by normalizing words so that their different forms are made identical.

Because many sentences matched the keywords, we summarized the filtered sentences using LexRank (Section 2) to identify a few of the most relevant ones. We then selected those, which in our judgement are indicative of the privacy concern to be our exemplar sentences and excluded those that are not. For example, the previous sentence was chosen as an exemplar. Despite containing

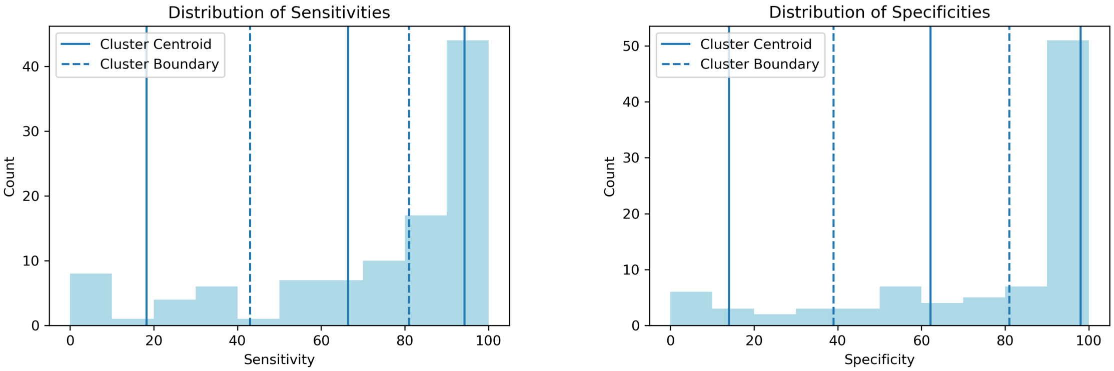

**Figure 2:** The distribution of sensitivities and specificities for the classifiers trained and evaluated using the process described in Sections 4.1 through 4.6. Undefined values are not shown. The solid lines denote 3-means cluster centroids and the dashed lines represent the boundaries between the clusters.

the keywords, the following sentence is not concerned with material changes and was not chosen:

> *You may change your interests at any time and may opt-in or opt-out of any marketing / promotional / newsletters mailings.*

## 4.4. Creation of Vector Space Models

We evaluated two approaches for creating vector-space models (VSM) of our privacy corpus. The first approach used LSI and the second a pre-existing BERT embedding (Fig 1, middle right).

For the LSI-based approach, we first created a traditional vector-space model (VSM) of the training set sentences by stemming their words. Unlike typical practice, we did not remove stop words because they help determine if a sentence should be selected. We then created bag of words (BOW) for each sentence and applied a term frequency-inverse document frequency (TFIDF) transformation followed by latent semantic indexing (LSI) using 400 dimensions, the value determined empirically by Bradford as giving the best results for large corpora [23].

For the BERT-based approach, we used a preexisting embedding which directly embeds each sentence into a 1024-dimension space [16]. Unlike LSI, we did not first stem the words of the sentences when using this preexisting embedding as it was not created using stemmed words.

## 4.5. Building Sentence Classifiers

Using the exemplar sentences and VSM from Sections 4.3 and 4.4, we created both LSI- and BERT-based sentence classifiers for each concern (Fig 1, right).

These classifiers begin by first selecting sentences from the test data that contain the predetermined keywords associated with each privacy concern and then selected those whose cosine similarity exceeds a selection threshold relative to at least one of their exemplar sentences. We chose a threshold value of 0.5 based on the intuition that for values above this threshold, the two vectors representing the sentences are more generally aligned with each other.

For example, consider a classifier that uses the exemplar sentence described in Section 4.3. It begins by using the keywords associated with that exemplar sentence, “you”, “opt-in”, “change” to filter test sentences. Let us assume that the following sentence is to be classified:

> *You will be notified if any of the material changes that affect the use of your personal information and asked to opt-in to the new use of your personal information.*

This sentence contains the keywords so the classifier will then calculate its embedding. This value is compared with that of the exemplar and the test sentence is rejected as being below the selection threshold. It may yet be classified as relevant if the classifier has another exemplar for which the sentence exceeds the similarity threshold.

For the BERT-based version of this classifier, the similarity between the exemplar exceeds the selection threshold, and the sentence is classified as relevant. The differences between the LSI- and BERT-based classifiers are presented in Section 5 and discussed in Section 6.

We chose this classification scheme to address class imbalances in the training set, to simplify classifier training, and to minimize the need for human-annotated data. Most of the sentences are not relevant to a given privacy concern and a random sample will not contain enough positive examples to train a sensitive classifier. The combination of keywords and exemplar sentences helped improve the classification sensitivity to the small number of relevant sentences. State-of-the-art methods, such as those based on Deep Learning, offer high accuracy with additional complexity and the need for large amounts of annotated data. We instead decided to pursue a simple approach that can still produce useful results.

### 4.6. Evaluating the Sentence Classifiers

We decided to evaluate the performance of the classifiers described in Sections 4.1 through 4.5 (Fig. 1, lower right) in terms of sensitivity and specificity instead of precision and recall as is typically done for information retrieval. This is a consequence of us being more interested in minimizing false negatives than minimizing false positives. With privacy policies, the former represent obligations missed by the search which cannot easily be found otherwise while the latter can be eliminated via manual inspection. False negatives can represent legal liabilities especially for a new product or service.

Sensitivity and recall are both synonyms for the true positive rate which is the fraction of correctly classified positive items [24, 25]. Specificity is defined as the true negative rate (the fraction of correctly classified negative items) [24, 25]. Classifier evaluation in terms of precision is sensitive to class imbalances because changes in the test data class distribution will change the classifier's apparent performance [24, 26]. By only considering a single class of a binary classification, specificity and sensitivity do not suffer from this "class skew" [24, 26]. This separate focus on the performance for each class also allows for the consideration of "cost skew", the different cost associated with errors for each class [25].

As mentioned above, we are more concerned with false negatives than false positives; for privacy policy analysis, we believe that the cost associated with misclassifying sentences as not relevant is larger than that of misclassifying them as relevant. Using sensitivity and specificity will allow us to separate these two situations.

The similarity-based classifiers were evaluated using sentences drawn from the test data. For each classifier, sentences that contained the keywords were used to create a test set. The sentences selected and rejected by each classifier were manually examined to identify false positives and false negatives; we randomly sampled classifier results when the output was too large for human analysis.

We calculated the sensitivity and specificity for each classifier and used $k$-means clustering ($k = 3$) to partition these values (sensitivity clusters: 0, 1, and 2 and specificity clusters: 0, 1, and 2). We chose $k = 3$ to roughly separate the sensitivity and specificity values into three groups denoting "low," "medium," and "high" to facilitate their analysis. Further investigation is needed to determine whether this grouping of values is optimal but we believe that it is a good starting point. These clusters were then used to group the results and divide them into five cases:

(i) **Higher sensitivity and specificity** (sensitivity cluster 2 and specificity cluster 2).

(ii) **Higher specificity** (cluster 2 or "high") **with lower sensitivity** (cluster 0 "low" and cluster 1 or "medium").

(iii) **Higher sensitivity** (cluster 2 or "high") **with lower specificity** (cluster 0 or "low" and cluster 1 or "medium").

(iv) **Mixed lower sensitivity and specificity values** (all clusters except sensitivity cluster 2 and specificity clusters 2)

(v) **Undefined sensitivity or specificity**

## 5. Evaluation Results

We will now present the results of evaluating the method described Sections 4.1 through 4.5 using the process described in Section 4.6 for the use case of searching the RE 2013 privacy policy corpus for privacy concerns N4 through N20 described in the Notice section of the Marotta-Wurgler study. Concerns N1, N2, and N3 were deemed as being addressed outside of the posted privacy policies. Concerns N12 and N13 were combined because they represented concerns that were too similar to be resolved individually. Concern N21 was omitted because initial searches of the corpus with command-line tools did not find any candidate sentences for this privacy concern.

We began by dividing corpus into training and test data using a 66 %/33 % split as described in Section 4.1. This resulted in training data consisting of 1373 policy documents and test data with 687 policy documents. We then used spaCy to segment the 140 000 training sentences and 68 400 test sentences.

Exemplar sentences were drawn from the training data; we identified 56 sets of keywords and exemplar sentences using the process described in Section 4.3. We then used gensim to create a 400-concept VSM of the training data sentences using LSI (Section 4.4).

For the BERT-based approach, we used the large embedding provided by the Flair NLP framework [16, 27] which embeds text into a 1024-dimension space.

The keywords, exemplar sentences, and the vector space models were then used to create 112 classifiers using the Python-based gensim package [28] as described in Section 4.5, one for each combination of embedding approach, set of keywords, and associated exemplar sentences. An input sentence is selected if contains the keywords and if its cosine similarity with an exemplar sentence exceeds the preset selection threshold.

The classifiers were evaluated using the test data (Section 4.6). Of the 68 400 sentences in the test data, 5695 matched the keywords used by our classifiers and were used for evaluation. We calculated classifier sensitivity and specificity; their distributions are shown as histograms in Fig. 2. We used $k$-means clustering to partition the sensitivities and specificities into three clusters: "low," "medium," and "high"; their centroids and boundaries appear as vertical lines in Fig. 2. These clusters

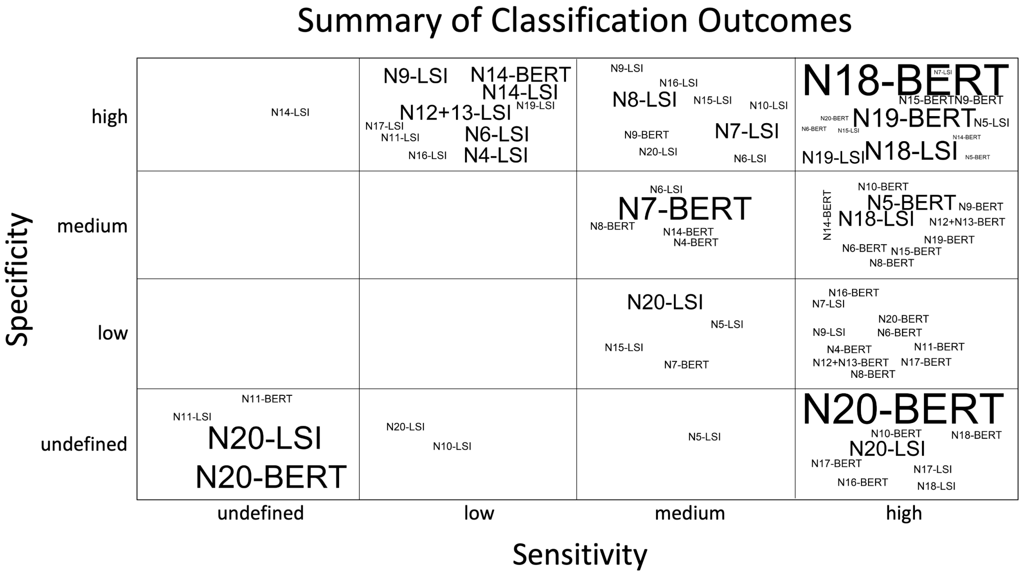

**Figure 3:** Results partitioned into tag clouds by their sensitivity and specificities. Each cell represents a separate tag cloud. The upper right cell contains results with the highest sensitivities and specificities, whereas the lower left cell has undefined sensitivities and specificities. Tag height is proportional to the number of associated classifiers. Tag position is not significant.

were used to group the results in Fig. 3 which were then divided into the five cases described above in Section 4.6 and in the rows of Table 2.

**Case 1: Combined Higher Specificity and Sensitivity** In addition to the results shown in Table 2, Table 3 gives details for the Case 1 classifiers which belonged to both sensitivity cluster 2 ("high sensitivity") and specificity cluster 2 ("high specificity"). This case contains 9 of the 17 concerns for which we applied our method and consisted of mostly relevant sentences with few missing.

**Case 2: Higher Specificity with Lower Sensitivity** These results consisted of mostly relevant sentences but the lower sensitivities meant some were missing. When combined with Case 1 above, 97 % of the test sentences were classified with high specificity; sentences identified as relevant to a privacy concern were likely to be relevant. These high specificity results corresponded to all of the concerns (N4 through N20) we considered.

**Case 3: Lower Specificity with Higher Sensitivity** Here, most of the relevant sentences were selected but many non-relevant ones were included as well. When combined with Case 1, 93 % of the test sentences contain both cases' keywords and be found with high sensitivity. These high sensitivity classifications also corresponded to all of the concerns (N4 through N20) we considered.

**Cases 4 and 5** The last two rows of Table 2 show the results for *Case 4 (Lower Specificity and Lower Sensitivity)* and *Case 5 (Undefined Sensitivity or Specificity)*. The former were a mixture of relevant and non-relevant sentences. The latter's had undefined specificities or sensitivities which occurred when the test set lacked either positive or negative examples.

We can now address our research questions:

**RQ1:** Which of the Marotta-Wurgler notice-related privacy concerns can be identified using our simple sentence similarity-based approach?

**Answer:** Based on the measures that we've chosen for **RQ1** in Section 4, we are able to identify sentences from 9 of the 20 one notice-related privacy concerns given in the Marotta-Wurgler study. These privacy concerns are marked with a dagger in Table 1.

**RQ2:** How well does our similarity-based approach to identify privacy concerns in a privacy policy corpus?

**Answer:** The 9 notice-related concerns from Case 1 above (Section 5) were classified with an average sensitivity of 93.8 % and an average specificity of 98.3 %. When combined with the results of Case 2 whose sentences were identified with an average specificity of 97.8 % and an average sensitivity of 35.6 %, then sentences from 17 of the 21 notice-related privacy concerns can be identified with specificities of at least 81 %.

The implications of these results for large scale analyses of privacy policies are discussed in Section 6.

## 6. Discussion

Our goal is a means to analyze large privacy policy corpora with reduced manual effort by leveraging the semi-

**Table 2:** Privacy Policy Sentence Classification Results

| Case | Specificity Level | Sensitivity Level | Candidate Sentences | Candidate % of Total | Found with LSI | Found with BERT | Specificity, % Min./Avg./Max. | Sensitivity, % Min./Avg./Max. |
|---|---|---|---|---|---|---|---|---|
| 1 | Higher | Higher | 777 | 13.6 | 188 | 209 | 83.3 / 98.3 / 100 | 81.3 / 93.8 / 100 |
| 2 | Higher | Lower | 5458 | 95.8 | 1253 | 555 | 82.2 / 97.8 / 100 | 0.00 / 35.6 / 71.4 |
| 3 | Lower | Higher | 5000 | 87.8 | 113 | 4379 | 0.00 / 41.6 / 80.0 | 87.3 / 92.6 / 100 |
| 4 | Lower | Lower | 837 | 14.7 | 109 | 427 | 0.00 / 44.7 / 80.0 | 50.0 / 68.2 / 80.0 |
| 5 | Undefined † | | 263 | 4.6 | 162 | 129 | NA | NA |

† Contains results with either unspecified sensitivities or specificities.

automated analysis of a small random sample that to develop classifiers that can be used to analyze the remainder. These classifiers are grounded in the corpus; we don't expect them to work with other corpora. Rather, we see the overall methodology as being the key contribution to the analysis large privacy policy corpora.

We used privacy concerns identified by Marotta-Wurgler to explore if we can combine keyword searches with manual selection of exemplar sentences to develop classifiers that find test set sentences similar to the exemplar sentences from the training set. If successful, this could allow for tagging of documents in a large corpus with a manual effort similar to that needed for a sample.

Our classification results fell into the five cases given in Section 5 that represent the different conditions under which sentences for the privacy concerns are identified. We will now discuss the results for these cases.

**Case 1: Combined Higher Sensitivity and Specificity** Table 3 reveals that for some of the privacy concerns, the use of keywords and exemplar sentences combined with the similarity-based classifiers allowed for test set sentence identification with few false negatives and false positives. This case represents an ideal situation that offers automation with little manual effort to extend Marotta-Wurgler style analyses to large privacy policy corpora.

**Case 2: Higher Specificity with Lower Sensitivity** This case mostly contains LSI-based results (Fig. 2, middle of top row). Here we identified relevant sentences, but not all of them. The average specificity of 97.8 % and the large fraction of test sentences with keywords suggest that the best improvements will come by increasing classifier sensitivity. This can come by increasing the default number (20) of exemplars suggested by LexRank. For example, there were 1348 test sentences for (N9) "collect + personal + information"; more exemplar sentences could have increased the matches and the sensitivity.

When combined with Case 1, 97 % of the Case 2 test sentences contain both cases' keywords and can be found with high specificity; sentences identified as relevant are likely to be relevant. These combined cases also contain all the concerns (N4 through N20) that we considered.

A more sophisticated classifier could improve sentence identification by training on a large manually curated data set. Alternatively, an active learning that prioritizes sentences for manual inspection, such as the one described by Yu and Menzies may be useful [19].

**Case 3: Lower Specificity with Higher Sensitivity** This case's average sensitivity of 92.6 % suggests that our approach may be useful for small numbers of sentences amenable to manual removal of false positives, especially for the higher specificity values (80 %).

A means of improving the specificity when there are many sentences, is to train better classifiers using a manually annotated data drawn from initial results. These new classifiers, could then be used with the similarity-based one to identify sentences with a higher specificity.

These results are dominated by the BERT-based classifiers implying that BERT-based features may be more advantageous than LSI-based ones by more consistently allowing for the use of secondary classifiers when the initial classification produces higher sensitivity results.

**Case 4: Lower Specificity and Lower Sensitivity** These results may be enough for the automated comparative analyses described above. For example, N6 ("pages + viewed") with LSI classifiers, and N7 ("collect+financial+information") with BERT classifiers had specificities of 77.9 % and 80.0 % and sensitivities of 79.8 % and 73.1 %, respectively. As in Case 2, better classifiers and active learning may improve the results.

**Case 5: Undefined Sensitivity or Specificity** Undefined sensitivities meant that there were no relevant test data items. For example, though the N14 keywords ("use + photograph + advertising − agree") matched three test sentences, all were correctly identified as not relevant. Undefined specificities meant that there were only relevant test set items; all sentences matching the keywords should be selected and our classifiers provided no benefit.

**Application to Privacy Policy Analyses** As discussed above, we rely on relevant keywords for each privacy concern; not all keywords lead to high sensitivity results. If a concern is best captured by multiple sets of keywords, each leading to classifications with differing sensitivities, then we will likely miss relevant sentences. However, we obtained high-specificities for the majority of the notice-related privacy concerns. While we cannot guarantee finding all relevant sentences, we can achieve situations

**Table 3:** Sentence Classification Results for Case 1

| concern | keywords | VSM | exemplars | test set | sens | spec |
|---|---|---|---|---|---|---|
| N5 | collect+browser +type | LSI | 20 | 65 | 88.9 | 100 |
| N5 | collect +operating +system | LSI | 16 | 65 | 90.4 | 100 |
| N5 | collect+operating +system | BERT | 16 | 65 | 86.8 | 83.3 |
| N6 | your+browsing +history | BERT | 20 | 9 | 87.5 | 100 |
| N7 | include+credit +history | LSI | 20 | 16 | 83.3 | 100 |
| N9 | collect+medical +information | BERT | 16 | 12 | 83.3 | 83.3 |
| N9 | collect+sensitive +information | BERT | 20 | 16 | 92.9 | 100 |
| N14 | use+information +advertising | BERT | 20 | 428 | 94.6 | 88.5 |
| N15 | advertisements +third+usage | BERT | 11 | 12 | 90.9 | 100 |
| N15 | advertisements+third +visit+collect-agree | LSI | 15 | 11 | 100 | 100 |
| N15 | advertisements+third +visit+collect-agree | BERT | 15 | 11 | 90.0 | 100 |
| N18 | material+change+prior | LSI | 20 | 22 | 89.5 | 100 |
| N18 | material+change+prior | BERT | 20 | 22 | 100 | 100 |
| N18 | notice+change+notify | BERT | 20 | 58 | 100 | 100 |
| N18 | notice+change+prior | BERT | 20 | 68 | 100 | 100 |
| N18 | policy+change+notify | LSI | 20 | 58 | 82.1 | 100 |
| N18 | policy+change+notify | BERT | 20 | 58 | 96.4 | 100 |
| N18 | policy+change+prior | LSI | 20 | 17 | 100 | 100 |
| N19 | material+change +accept | LSI | 8 | 3 | 100 | 100 |
| N19 | material+change +accept | BERT | 8 | 3 | 100 | 100 |
| N19 | notice+change +accept | LSI | 16 | 20 | 81.3 | 100 |
| N19 | notice+change +accept | BERT | 16 | 20 | 100 | 100 |
| N19 | policy+change +accept | LSI | 20 | 28 | 100 | 100 |
| N19 | you+opt-in+change | BERT | 1 | 2 | 100 | 100 |
| N20 | material+change +previous | BERT | 3 | 6 | 100 | 100 |

where most of those identified are and their relative frequencies and distributions across a corpus can form the basis of large-scale analyses that yield valuable insights, such as differences across market segments, in a manner similar to Marotta-Wurgler's analysis.

We envisage iterative analyses, starting with keyword selection using text searches to verify that they lead to relevant sentences. Subsequent classification using exemplar sentences may uncover anomalies such as the absence of a privacy concern in a segment of the remaining corpus. Further investigation could lead to new keywords and exemplar sentences. Because its substantial qualitative nature, the concept of saturation should be used to end the iterations; it is reasonable to stop when no new information is found despite systematic attempts [29].

## 7. Limitations

The methodology we present is a starting point for further investigation of large scale semi-automated policy analyses. Despite our encouraging results, we have not solved the problem of semi-automated privacy policy analysis.

Our use of keywords is intended to mitigate the class imbalances that occur in a random samples of privacy policy sentences. We do not offer guidance on keyword selection or evaluation. We believe we have made reasonable choices. Further research could better determine their selection or eliminate them altogether.

We rely on similarity to exemplar sentences to reduce the need for human-annotated ground truth and to keep our baseline method simple while producing useful results. We have necessarily forsaken the performance of state-of-the-art methods. Additional studies could explore the tradeoffs of using more complex methods.

We assumed sentences are at the appropriate level of granularity for privacy policy analyses. This may not hold for all privacy concerns, but this was both convenient for our analysis and worked reasonably well for those described by Marotta-Wurgler's study.

For simplicity, we used a threshold for the cosine similarity of 0.5 with the intuition that above this value two vectors are more mutually aligned than not. In practice, classifier receiver operating characteristic (ROC) curves could be used to select the thresholds.

The clustering of sensitivities and specificities into "low," "medium," and "high" (Fig. 3) and their grouping into "lower" and "higher" (see Table 2 and Section 5) result from our choice of clustering algorithm and our need to partition our results for analysis and discussion. This is a qualitative interpretation similar to inter-rater comparisons [30]. Further work is necessary to determine optimal result clustering and labels in practice.

## 8. Summary and Future Work

We presented a semi-automated methodology for extending Marotta-Wurgler style analyses to large privacy policy corpora and provided an initial feasibility evaluation with an existing privacy policy corpus. Our methodology begins with a keyword-based search followed by extractive summarization to find the most representative sentences in a policy corpus. They are then manually inspected to find exemplars for similarity-based classifiers that identify other relevant sentences. The classifiers identified sentences addressing concerns with varying results. For those identifiable with higher sensitivity and specificity, the manual effort of analyzing a small subset of a large privacy policy corpus could lead to automated analysis of the rest. This condition accounted for 9 of the 17 privacy concerns for which we used our methodology.

When sentences were identified with higher sensitivity but lower specificity, the results were dominated by the BERT-based similarity classifiers. The possibility of improving the specificity suggests an advantage of BERT over LSI. Here the non-relevant sentences could be removed manually or by using another classifier trained on a subset of the corpus. When combined with the higher sensitivity and specificity results, we were able to identify sentences from all seventeen privacy concerns with which we evaluated our methodology.

Our future efforts will focus on increasing the number of privacy concerns identified while maintaining the aesthetic of simplicity and minimal training data requirements. We will investigate using our method to create training sets for additional downstream classifiers as well as active learning approaches. We will also evaluate its use on a much larger privacy policy corpus.

## References

[1] S. Cowley, "Equifax to pay at least $650 million in Largest-Ever data breach settlement," *The New York Times*, July 2019.

[2] E. C. Baig, "Facebook fined $5 billion by FTC, must update and adopt new privacy, security measures," *USA Today*, July 2019.

[3] A. K. Massey, J. Eisenstein, A. I. Anton, and P. P. Swire, "Automated text mining for requirements analysis of policy documents," *2013 21st IEEE Intl Requirements Eng. Conf., RE 2013 - Proc.*, pp. 4–13, 2013.

[4] F. Marotta-Wurgler, "Understanding privacy policies: Content, Self-Regulation, and Markets," Tech. Rep. 16-18, New York University School of Law, 2016.

[5] L. Netcraft, "January 2019 web server survey | netcraft." `https://news.netcraft.com/archives/2019/01/24/january-2019-web-server-survey.html`, 2019. Accessed: 2019-8-19.

[6] G. Erkan and D. R. Radev, "LexRank: Graph-based lexical centrality as salience in text summarization," *1*, vol. 22, pp. 457–479, Dec. 2004.

[7] R. Feldman, "Text mining," in *Handbook of Data Mining and Knowledge Discovery* (W. Klosgen and J. M. Zytkow, eds.), pp. 749–757, Oxford University Press, 2002.

[8] J. Lin and D. Gunopulos, "Dimensionality reduction by random projection and latent semantic indexing," *Proc. Text Mining Workshop, at the 3rd SIAM Intl. Conf. Data Mining*, 2003.

[9] C. D. Manning and H. Schutze, *Foundations of Statistical Natural Language Processing*. Cambridge, MA: The MIT Press, 1999.

[10] B. Liu, *Web Data Mining*. Berlin: Springer-Verlag, second edi ed., 2011.

[11] S. Deerwester, S. T. Dumais, and T. K. Landauer, "Indexing by latent semantic analysis," *J American Soc. for Inform. Sci.*, vol. 41, no. 6, pp. 391–407, 1990.

[12] T. K. Landauer, P. W. Foltz, and D. Laham, "An introduction to latent semantic analysis," *Discourse Process.*, vol. 25, pp. 259–284, Jan. 1998.

[13] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean, "Distributed representations of words and phrases and their compositionality," in *Adv. Neural Inform. Process. Sys. 26* (C. J. C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K. Q. Weinberger, eds.), pp. 3111–3119, Curran Associates, Inc., 2013.

[14] J. Pennington, R. Socher, and C. Manning, "Glove: Global vectors for word representation," in *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pp. 1532–1543, 2014.

[15] M. E. Peters, M. Neumann, M. Iyyer, M. Gardner, C. Clark, K. Lee, and L. Zettlemoyer, "Deep contextualized word representations," Feb. 2018.

[16] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," Oct. 2018.

[17] S. Harispe, S. Ranwez, S. Janaqi, and J. Montmain, *Semantic Similarity from Natural Language and Ontology Analysis*, vol. 8 of *Synthesis Lectures Human Language Technologies*. Morgan & Claypool Publishers, 2015.

[18] M. R. Grossman, G. V. Cormack, and A. Roegiest, "TREC 2016 total recall track overview," in *TREC*, 2016.

[19] Z. Yu and T. Menzies, "Total recall, language processing, and software engineering," Aug. 2018.

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

[21] M. Degeling, C. Utz, C. Lentzsch, H. Hosseini, F. Schaub, and T. Holz, "We value your privacy ... now take some cookies: Measuring the GDPR's impact on web privacy," in *Proceedings 2019 Network and Distributed System Security Symposium*, (Reston, VA), Internet Society, 2019.

[22] T. Linden, R. Khandelwal, H. Harkous, and K. Fawaz, "The privacy policy landscape after the GDPR," *Proc. on Privacy Enhancing Tech.*, vol. 2020, pp. 47–64, Jan. 2020.

[23] R. B. Bradford, "An empirical study of required dimensionality for large-scale latent semantic indexing applications," in *17th ACM conference on Information and knowledge mining*, (Napa Valley, California), pp. 153–162, ACM Press, 2008.

[24] T. Fawcett, "ROC graphs: Notes and practical considerations for researchers," Tech. Rep. HPL-2003-4, HP Laboratories, Jan. 2003.

[25] J. Korst, V. Pronk, M. Barbieri, and S. Consoli, "Introduction to classification algorithms and their performance analysis using medical examples," in *Data Science for Healthcare: Methodologies and Applications* (S. Consoli, D. Reforgiato Recupero, and M. Petković, eds.), pp. 39–73, Cham: Springer International Publishing, 2019.

[26] F. M. Harper, D. Moy, and J. A. Konstan, "Facts or friends?: distinguishing informational and conversational questions in social Q&A sites," in *CHI '09: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, Assoc. for Computing Machinery, 2009.

[27] A. Akbik, T. Bergmann, D. Blythe, K. Rasul, S. Schweter, and R. Vollgraf, "FLAIR: An Easy-to-Use framework for State-of-the-Art NLP," in *Proc. 2019 Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations)*, pp. 54–59, 2019.

[28] R. Řehůřek and P. Sojka, "Software framework for topic modelling with large corpora," in *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*, (Valletta), pp. 45–50, ELRA, May 2010.

[29] S. B. Merriam and E. J. Tisdell, *Qualitative Research*. John Wiley & Sons, Inc., 4th editio ed., 2016.

[30] K. A. Hallgren, "Computing Inter-Rater reliability for observational data: An overview and tutorial," *Tutor. Quant. Methods Psychol.*, vol. 8, no. 1, pp. 23–34, 2012.