COS424 Final Project: Predicting Compliance in Privacy Policies

Jordan Holland
Princeton University
jordanah@princeton.edu

Ben Kaiser
Princeton University
bkaiser@princeton.edu

Kevin Lee
Princeton University
kl26@princeton.edu

Elena Lucherini
Princeton University
elucherini@princeton.edu

Abstract

There exist several laws that mandate what clauses commercial websites must have in their privacy policies for compliance. The number of extant privacy policies today makes compliance enforcement through manual examination infeasible. In this work we apply machine learning methods to automatically check the compliance of privacy policies against 8 separate outcomes. We find varying success in doing so, with 5-fold cross validation scores of over 95% for Do Not Track and contact detail compliance, while under 40% for determining if the subject consents to third-party tracking.

1 Introduction

Several laws mandate that commercial websites and online servers provide privacy policies. These include the European Union's General Data Protection Regulation (EU GDPR), California's 2004 Online Privacy Protection Act (CalOPPA), and California's 2018 Consumer Privacy Act (CCPA). Each of these laws specifies the clauses that a privacy policy must contain for compliance. For example, CalOPPA requires privacy policies to include clauses that list the categories of personal information that will be collected, a list of third parties with whom it will be shared, and descriptions of processes by which consumers can review and modify their personal data that is held by the company. GDPR's requirements are even more extensive than CalOPPA.

There are millions of extant privacy policies on the Internet, making compliance enforcement through manual examination infeasible. Our project attempts to alleviate this issue by applying machine learning to automatically check the compliance of privacy policies against checklists of GDPR, CalOPPA, and CCPA requirements. Using supervised learning, we develop models to predict 8 unique outcomes that are required by the aforementioned laws. We find varying success in predicting each outcome, with high levels of success predicting outcomes such as the presence of contact details and Do Not Track compliance, but much lower success on outcomes such as the list of third parties who information will be shared with. Ultimately, we believe that our initial results show promise into automatically determining if a website's privacy policy is *not in compliance* by checking against some of the outcomes the classifiers had more success predicting.

Related Work Prior research has considered synthesis and analysis of natural language privacy policies using a combination of automated and manual techniques.

Using contextual integrity – a rigorous framework that defines privacy in terms of the contextual appropriateness of information flows – Shvartzshnaider et al. demonstrated that crowdworkers can produce precise annotations of privacy policy excerpts [5]. These annotations describe the senders,

recipients, and subjects of information, the attributes of that information, and the conditions under which it may be transferred or collected. This approach allows them to quantify the number of information flows expressed in a policy (which is indicative of its complexity), detect incompletely described flows, and identify vague or ambiguous language.

Carnegie Mellon University's Usable Privacy Policy Project has demonstrated strong results combining crowdsourcing and natural language processing to semi-automatically extract key features from policies [4, 8]. In their approach, crowdsourced annotations feed into NLP models, which identify linguistic features that researchers then translate into descriptive logic statements that express specific data practices.

These approaches involve manual processing of policies and in the case of CMU, further manual processing of output. The only prior work we know of that is fully automated, like our approach, is Polisis, a fully automated framework for analyzing privacy policies developed by Harkous et al [2]. The authors developed a privacy-centric language model from a corpus of 130K policies, then applied it to build a neural network classifier that operates on short policy segments, applying fine-grained annotations describing data practices. They achieved strong results, but predictions by neural networks are difficult to understand or explain. Our modeling approach allows for investigation of classifications that can feed back into feature extraction and training to improve results.

Project Overview In this work we attempt to automatically predict the compliance of privacy policies. We leverage a dataset of over one million privacy policies scraped from the web by Princeton Center for Information and Technology Policy (CITP) researchers. We additionally make use of Carnegie Mellon University's OPP-115 corpus, which is a collection of 115 website privacy policies that graduate students in law annotated with tags indicating specific data practices in the text [7]. Specifically, our project consists of extracting features and outcomes from OPP-115, training supervised learning models on this data, evaluating performance, and applying the models to the larger CITP dataset.

Section 2 explores the datasets and provides summary statistics and visualizations. Section 3 describes how we extract outcomes and classification features from the data. Section 4 discusses our modeling approach and classification results. Finally, Section 5 discusses the implications of our findings and suggests directions for future work.

2 Data

2.1 Dataset Structure

CITP policies CITP researchers collected over 1 million privacy policies from online services using Common Crawl, a free repository of web crawl data. By first extracting policy links from crawl metadata, the researchers were able to batch download all of these policies in April 2018. For our project, we sample 10,000 of these policies at random as long as the policies are longer than 500 words and contain the words "privacy", "policy", and "cookie". The policies are stored in .txt files and are cleaned to remove erroneous entries, vestigial HTML tags and entities, and other artifacts of scraping.

OPP-115 The OPP-115 corpus, released in 2016, contains 115 privacy policies for online services and 23,000 annotations specifying data practices in the text. Each policy was read and annotated independently by three graduate students in law; annotations were then consolidated. The policies were extracted from websites sampled from Google search results for top queries in May 2015 and distributed equally across categories (e.g., Arts, Shopping, Business, News).

Annotations begin with one of ten categories, such as *First Party Collection/Use*, *Third Party Sharing/Collection*, *User Access, Edit, & Deletion*, or *Data Security*. Within a category, an annotation specifies a set of key-value attributes; for example, a data practice within the *First Party Collection/Use* category would include attributes describing the means of collection, type of information, purpose of collection, and a number of other features. Finally, each annotation contains pointers to the span of text in the policy to which it applies.

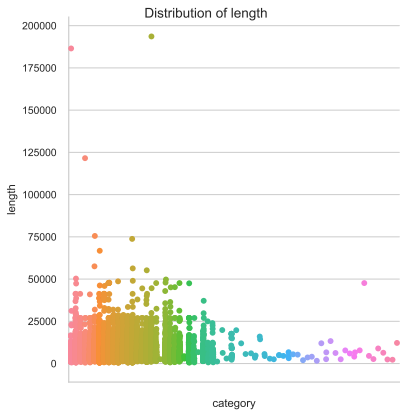


Figure 1: Distribution of length of privacy policies across categories.

2.2 Exploratory Analysis

We begin with exploratory analysis of the two datasets. We group policies by the category of the website and examine features which represent *complexity*: the difficulty that an average user would have in understanding the policy and exercising the rights that it grants them. The two main features we examine are the length of the policy and the number of user choices or decisions it contains.

2.2.1 Exploring the CITP Dataset

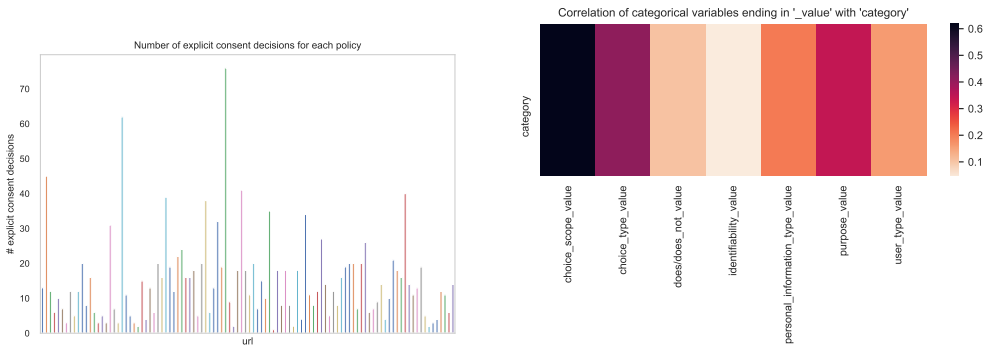
To break the CITP dataset into more managable chunks, we separate websites by their category using the Webshrinker API, which maps a website to a category based on the Interactive Advertising Bureau's standardized website categories [6, 3]. These include Technology & Computing (which comprises over 10% of our dataset), Business, Health & Fitness, and Hobbies & Interests.

Figure 1 shows the distribution of length of the privacy policies organized by category. We observe that the majority of the privacy policies are shorter than 25,000 words; only four policies are longer than 75,000 words.

2.2.2 Exploring the OPP-115 Dataset

Next, we explore Carnegie Mellon University's OPP-115 dataset. First, we find the number of user decisions contained in each privacy policy. We obtain these numbers by filtering the dataset to only include entries that belong to the "User Choice/Control" category, which contains information about the control options available to users in the privacy policies. Then, for each policy we count the number of occurrences of the "Choice Type" attribute, which is used for user choices that are offered explicitly. We exclude instances of the attribute where the value is "Don't use service/feature", as it indicates that the way to opt out of a given clause is to not use the service, only leaving values such as "Opt-in" and "Opt-out". Figure 2a shows the results of this analysis. The maximum number of explicit user choices is 76, found in latinpost.com's privacy policy; the minimum number is given by ocregister.com's policy, with only one explicit consent decision left to the user.

Second, we find correlations between variables in the dataset. We only take into account categorical variables, as the numerical variables only specify the segments of the policy text where the given clause was found. We calculate the correlation with Cramer's V statistic for categorical-categorical association, with Bergsma's bias correction [1]. According to Cramer's V, which ranges from 0 to 1, attributes "Choice Type", "Choice Scope", and "Collection Mode" are correlated. "Choice Scope" indicates the scope of user choices and its values can be "Collection", "Use", "Both", or "Unspecified". "Collection Mode" denotes whether the data collection needs user's explicit consent. Similarly, "User Type" —"User without account", or "User with account"— is correlated with these three attributes and "Identifiability" —whether the data collection is identifiable or anonymized.



(a) Number of explicit consent decisions for each policy.

(b) Correlation between "Category" and other categorical variables in the dataset.

Figure 2

Finally, we look at the correlation of variable "Category" and other categorical variables in the dataset. Figures 2b shows the variables correlated to "Category" according to Cramer's V (entries for variables with no correlation were omitted in the figure). We observe that "Choice Scope" is highly correlated to "Category", followed by "Choice Type" and "Purpose" — denoting the purpose of collecting or using user information, with values such as "Basic service", "Advertising", and "Marketing".

3 Features

The first step in modeling the policies is to specify outcome variables. We read all three regulations – GDPR, CalOPPA, and CCPA – and extract clauses that refer to concrete requirements for provisions. We then disambiguate the clauses to generate a set of potential outcomes.

Next, we determine which of these outcomes our data will allow us to evaluate. For each outcome, we generate Boolean expressions over OPP-115 tags that correspond to whether the outcome was satisfied by the policy or not. The most straightforward example is the outcome which records whether a policy explains how it responds to Do Not Track signals, which are sent by browsers to indicate that a user is requesting that the web application disable individual-level tracking. If the policy contains an OPP-115 tag in the "Do Not Track" category with the attribute "Honored", "Not honored", or "Other", we mark the policy as meeting the outcome. If the attribute is "Not mentioned" or "Mentioned, but unclear if honored", we mark the policy as failing to meet the outcome. Other outcomes involve similar, but generally more complex, logic. Table 1 lists all outcomes along with the distribution of policies in the dataset that comply with the given outcome.

Based on these distributions, it is clear that some of the outcomes will be impossible to predict. For example, 114 out of 115 labeled policies provide acknowledgment if third parties collect personally identifiable information (PII) on their website, which renders the outcome useless in terms of classification. We select 8 outcomes (highlighted in grey in Table 1) with more useful distributions for prediction.

Next, for each outcome, we extract *relevant* n-grams (from $n = 1$ to 4). To determine which spans of text in the policy are pertinent to each outcome, we rely on the `selectedText` attribute of the OPP-115 tags which contains the text to which the tag applies. When evaluating policy outcomes, we save the text for every tag that matches one of the tag clauses in our evaluation expressions. Our intuition is that these tags contain the text that determines whether or not the outcome is met, and the text spans are often short (1-10 words), so they should be highly predictive.

We tokenize all of the saved text, remove stop words, lemmatize using `nltk.WordNetLemmatizer`, and stem using `nltk.PorterStemmer`. We then generate n-grams for each n and generate n-gram frequency tables. Figure 3 shows the resulting vocabulary size for each outcome and value of n; our largest vocabulary was nearly 20,000 items in size.

Outcome	CCPA	GDPR	CalOPPA	Comply	Don't Comply	Unknown
Disclose categories of PII collected	Y	N	Y	111	2	2
Identify third parties with whom PII may be shared	Y	Y	Y	107	2	6
Process to review and request changes to PII	N	N	Y	0	6	109
Process to notify consumers of changes to policy	N	N	Y	53	13	49
Response to Do Not Track	N	N	Y	28	87	0
Other parties collect PII from first party site	N	N	Y	114	1	0
Disclose specific pieces of collected PII upon request	Y	Y	N	41	74	0
Right to request erasure of data	Y	Y	N	37	78	9
Categories of sources from which PII is collected	Y	N	N	113	2	0
Purpose for collecting PII	Y	Y	N	113	2	0
Contact details	N	Y	N	102	13	0
Retention period	N	Y	N	30	85	0
Subject consents to first-party processing	N	Y	N	25	89	1
Subject consents to third-party processing	N	Y	N	15	94	6

Table 1: Outcomes and distributions in OPP-15 dataset

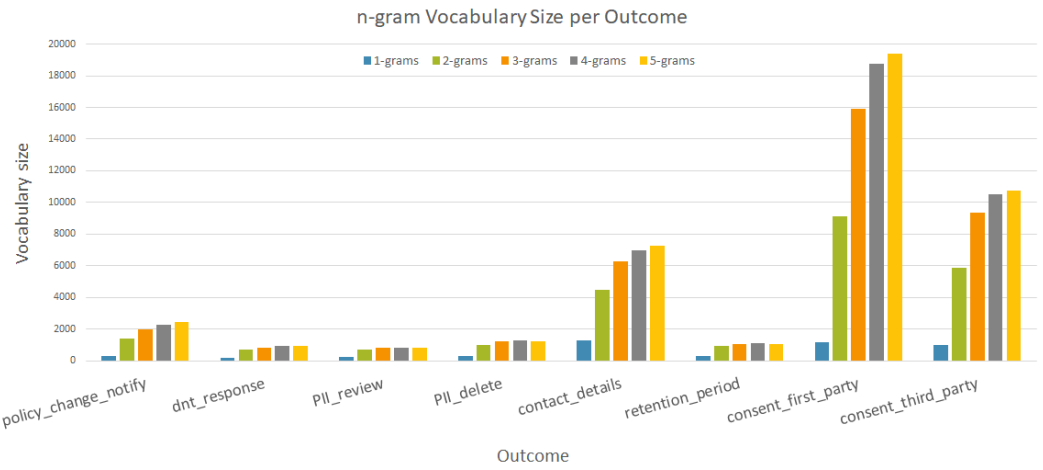


Figure 3: The vocabulary sizes for each n-gram size and outcome value

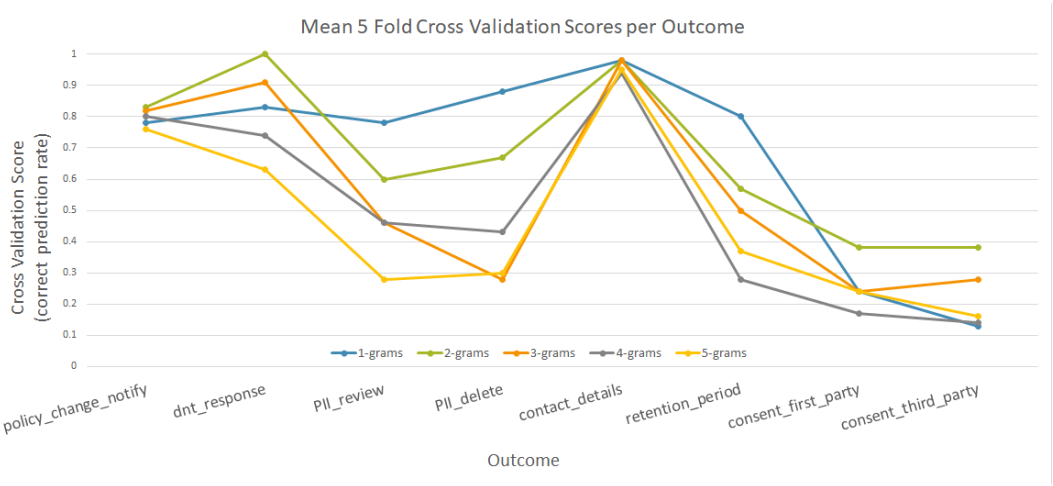


Figure 4: Cross validation scores for our classifiers on the OPP-115 dataset, 80/20 test/train split

Latent structure analysis We attempted to apply latent Dirichlet analysis (LDA) to derive topics that we could use as predictive features or for clustering policies by outcome compliance. We split the text into individual clauses and generated various numbers of topics (intervals between 20 and 500) for each outcome based on these corpora. Our intuition was that while entire policies would likely not cluster based on topic, individual clauses might, and because some outcomes are decided by the presence or absence of a single clause, those clusters might map to our outcomes. Unfortunately, none of the generated topics appeared to correspond to any outcomes. Furthermore, the top associated words with generated topics had little variance between topics. Ultimately, we discarded these features.

4 Results

We attempt to predict each outcome independently, training separate classifiers for each. We use the default implementation of `RandomForestClassifier` in Scikit-learn to predict each outcome. We also ran classification using `DecisionTreeClassifier` from Scikit-learn, producing similar results. Ultimately, the results presented in this section are from the `RandomForestClassifier`. We split the OPP-155 dataset into a training set consisting of 80% of the policies and a testing set containing the other 20%. We present results in three formats:

- Figure 4 shows 5-fold cross-validation scores when our classifiers are applied to the 20% of policies withheld for testing
- Figure 5 shows average precision scores across each outcome.
- Table 2 shows the result of applying our classifiers to the CITP dataset and compares predicted compliance rates in that dataset to the ground truth compliance rates of the OPP-115 dataset.

Cross-validation shows strong performance for many outcomes, particularly from models trained on unigrams and bigrams. The presence of contact details in the policy achieves near-perfect prediction with all five models. This is one of the simpler outcomes to evaluate, as it only relies on the presence of a single attribute tag. Table 3 shows top unigrams for each outcome; “contact”, “number”, and “zip” are all very useful features for this outcome.

The classifier also performs well on the Do Not Track (DNT) feature , achieving a perfect score based on bigrams and a score of 90% based on trigrams. This is also one of the easier features to compute compliance for, as it relies on only two attribute tags. Examining the top features for this outcome (Tables 4 and 5), we see that the top features are the presence of the n-grams “do not” and “do not track”, respectively. “your browser” and “your browser settings” are also highly ranked features, which makes sense as DNT is a browser setting.

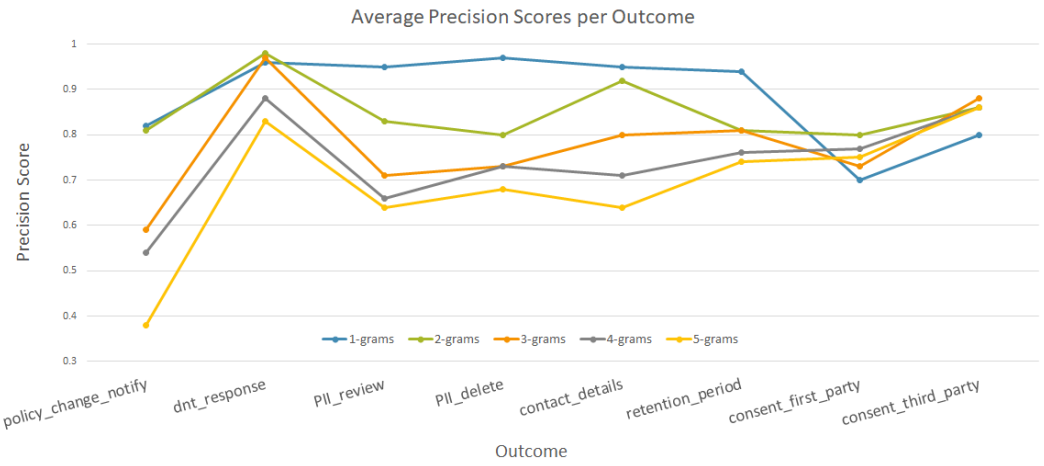


Figure 5: Average precision scores for each classifier. Note that the y-axis begins at 0.3 (to conserve space)

Outcome	**CITP policies in compliance**	**CITP policies not in compliance**	**Pct. CITP policies in compliance**	**Pct. OPP-115 policies in compliance**
policy_change_notify	3637	6363	36%	46%
dnt_response	5241	4759	52%	24%
PII_review	9881	119	99%	36%
PII_delete	3508	6492	35%	32%
contact_details	9701	299	97%	89%
retention_period	9420	580	94%	26%
consent_first_party	4682	5318	47%	22%
consent_third_party	5763	4237	58%	13%

Table 2: Outcome compliance metrics for both datasets. OPP-115 is ground truth; CITP policies are predictions.

The next most successful feature in cross-validation is notification of policy changes. Our models achieve between 75% and 85% correct classification rates, with bigrams and trigrams again performing best. Top trigrams (Table 5) are especially easy to understand as predictive features for this outcome: "publish our changes", "amend this privacy", and "will be notified" all clearly refer to mechanisms by which the policy may be updated and consumers may be notified.

On the other hand, the classifier is particularly unsuccessful at classifying the first- and third- party consent outcomes across all n-grams. We identify two reasons this may be the case. First, the vocabularies are much larger for these outcomes, as shown in Figure 3. The top n-grams describe consumer choices ("opt-in", "you may", "you have"), but nothing specific to data collection, suggesting that key words that may differentiate this outcome such as "collection" or even "first party" and "third party" are not frequent enough to register. Second, evaluating these outcomes involves computing complex expressions over five tags, each with multiple possible satisfactory attribute values. This increases the likelihood that labeling errors, textual ambiguities, or lack of complete coverage in our outcome evaluation expressions could interfere in both training and classification.

Live classification We conclude our classification by applying our models trained on unigrams to the CITP dataset. Results are shown in Table 2. For our unsuccessful outcomes, these results are likely inaccurate, but metrics for our most successful outcomes (contact_details and dnt_response) warrant examination. While 97% of policies comply with the contact details provision, we found that nearly 50% of policies in the CITP corpus failed to comply with Do Not Track requirements.

5 Conclusion

In this work, we attempt to apply machine learning methods to automate the compliance check of privacy policies. After a preliminary exploratory analysis and data processing, we determine 8 different outcomes to evaluate. We obtained scores of over 95% for Do Not Track and contact detail compliance with 5-fold cross-validation. On the other hand, we obtained under 40% accuracy for third-party tracking consent.

This work has several limitations that were highlighted in Section 4. To improve the performance of the classifier, we suggest several avenues for future work. First, collaborating with domain experts (i.e., law students or professionals) would help us more precisely map tags to outcomes or even develop an outcome-specfic annotation language if needed.

For feature generation, we believe that word embeddings may offer greater predictive power than n-grams for some outcomes. This approach would take into account words' context to help disambiguate different phrasings of the same concept (i.e., "users can opt out" vs "users may opt out").

policy_change_notify	dnt_response	PII_review	PII_delete	contact_details	retention_period	consent_first_party	consent_third_party
without	signal	inform	delet	com	inform	e	mail
circumst	track	access	remov	contact	email	ask	consent
indic	cooki	review	updat	number	legal	base	right
amend	yet	process	membership	nc	transact	parti	releas
yahoo	would	year	youd	zip	expir	whether	program
www	world	write	yahoo	zack	yahoo	subscript	note
would	work	would	write	yourcaliforniaprivacyright	without	tell	softwar
wireless	wide	within	would	your	wish	gather	zone
whether	whether	wish	without	york	window	zone	zip
weve	websit	web	within	year	well	zip	zack

Table 3: Top 10 features for 1 grams

policy_change_notify	dnt_response	PII_review	PII_delete	contact_details	retention_period	consent_first_party	consent_third_party
so_pleas	do_not	person_inform	if_you	if_you	the_inform	opt_in	opt_in
updat_date	about_and	may_review	delet_your	send_an	e_mail	we_will	you_have
notic_an	thi_time	inform_in	from_our	privaci_polici	the_email	u_you	you_may
without_notic	includ_do	to_review	of_certain	to_dcccd	your_account	u_submit	can_see
your_yahoo	your_web	you_may	no_longer	u_at	one_time	for_one	by_open
your_user	your_person	not_access	have_inform	zip_code	these_servic	they_wish	but_onli
your_right	your_onlin	flash_lso	set_on	zack_com	transact_is	system_analyz	zone_countri
your_privaci	your_interest	of_view	remov_of	yourself_from	for_busi	thi_web	zip_code
your_personallyidentifi	your_experi	you_can	you_will	yourcaliforniaprivacyright_enthusiastnetwork	destruct_raw	zone_your	zack_web
your_person	your_browser	identifi_inform	not_access	your_websit	90_day	zip_code	yourself_from

Table 4: Top 10 Features for bigrams

policy_change_notify	dnt_response	PII_review	PII_delete	contact_details	retention_period	consent_first_party	consent_third_party
date_at_the	do_not_track	the_abil_to	your_person_inform	if_you_have	e_mail_address	opt_in_we	in_to_receiv
prior_notic_an	consum_opt_out	person_identifi_inform	delet_your_account	contact_u_at	may_not_be	about_your_activ	third_parti_inform
publish_our_chang	includ_do_not	you_may_review	in_our_archiv	or_remov_request	doe_not_store	whether_you_have	on_facebook_your
promin_notic_of	about_and_opt	your_person_inform	delet_of_certain	our_privaci_polici	close_these_cooki	we_collect_and	student_permiss_the
amend_thi_privaci	websit_and_the	to_request_and	updat_or_delet	you_would_like	origin_sourc_inform	submit_inform_name	mail_or_other
privaci_notic_may	your_web_browser	lso_current_on	you_have_the	not_deliv_you	in_our_databas	we_automat_collect	in_counsel_or
signific_appropri_notic	your_person_inform	wish_to_review	have_inform_delet	request_pleas_send	one_time_email	submit_it_to	zone_countri_refer
part_of_our	your_onlin_activ	limit_purpos_of	well_remov_your	www_bbb_org	you_may_wish	refer_to_a	zip_code_associ
will_be_notifi	your_experi_on	to_the_person	remov_from_gamestop	use_the_follow	time_account_inform	number_or_a	zack_web_site
polici_will_be	your_browser_set	password_to_access	custom_servic_at	5th_floor_oakland	audit_purpos_your	when_you_interact	yourself_from_these

Table 5: Top 10 features for trigrams

References

[1] Wicher Bergsma. A bias-correction for Cramer's V and Tschuprow's T. *Journal of the Korean Statistical Society*, 42, 09 2013.

[2] Hamza Harkous, Kassem Fawaz, Rémi Lebret, Florian Schaub, Kang G. Shin, and Karl Aberer. Polisis: Automated analysis and presentation of privacy policies using deep learning. In *27th USENIX Security Symposium (USENIX Security 18)*, pages 531–548, Baltimore, MD, 2018. USENIX Association.

[3] Interactive Advertising Bureau. IAB Tech Lab Context Taxonomy.

[4] Frederick Liu, Shomir Wilson, Peter Story, Sebastian Zimmeck, and Norman Sadeh. Towards automatic classification of privacy policy text. Technical Report CMU-ISR-17-118R, Carnegie Mellon University, June 2018.

[5] Yan Shvartzshnaider, Noah Apthorpe, Nick Feamster, and Helen Nissenbaum. Analyzing privacy policies using contextual integrity annotations. *CoRR*, abs/1809.02236, 2018.

[6] Webshrinker. APIs - Webshrinker.

[7] 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. 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)*, pages 1330–1340, Berlin, Germany, August 2016. Association for Computational Linguistics.

[8] Shomir Wilson, Florian Schaub, Frederick Liu, Kanthashree Mysore Sathyendra, Daniel Smullen, Sebastian Zimmeck, Rohan Ramanath, Peter Story, Fei Liu, Norman Sadeh, and Noah A. Smith. Analyzing privacy policies at scale: From crowdsourcing to automated annotations. *ACM Trans. Web*, 13(1):1:1–1:29, December 2018.