A Large-Scale Exploration of Terms of Service Documents on the Web

Soundarya Nurani Sundareswara, Mukund Srinath, Shomir Wilson, C. Lee Giles
Pennsylvania State University
State College, Pennsylvania, USA
{sxn5310,mus824,shomir,clg20}@psu.edu

ABSTRACT

Terms of service documents are a common feature of organizations' websites. Although there is no blanket requirement for organizations to provide these documents, their provision often serves essential legal purposes. Users of a website are expected to agree with the contents of a terms of service document, but users tend to ignore these documents as they are often lengthy and difficult to comprehend. As a step towards understanding the landscape of these documents at a large scale, we present a first-of-its-kind terms of service corpus containing 247,212 English language terms of service documents obtained from company websites sampled from Free Company Dataset. We examine the URLs and contents of the documents and find that some websites that purport to post terms of service actually do not provide them. We analyze reasons for unavailability and determine the overall availability of terms of service in a given set of website domains. We also identify that some websites provide an agreement that combines terms of service with a privacy policy, which is often an obligatory separate document. Using topic modeling, we analyze the themes in these combined documents by comparing them with themes found in separate terms of service and privacy policies. Results suggest that such single-page agreements miss some of the most prevalent topics available in typical privacy policies and terms of service documents and that many disproportionately cover privacy policy topics as compared to terms of service topics.

CCS CONCEPTS

• **Information systems** → *Web mining*.

KEYWORDS

terms of service, corpus, availability

ACM Reference Format:
Soundarya Nurani Sundareswara, Mukund Srinath, Shomir Wilson, C. Lee Giles. 2021. A Large-Scale Exploration of Terms of Service Documents on the Web. In *ACM Symposium on Document Engineering 2021 (DocEng '21), August 24–27, 2021, Limerick, Ireland.* ACM, New York, NY, USA, 4 pages. https://doi.org/10.1145/3469096.3474940

DocEng '21, August 24–27, 2021, Limerick, Ireland

ACM ISBN 978-1-4503-8596-1/21/08…$15.00
https://doi.org/10.1145/3469096.3474940

1 INTRODUCTION

Terms of service are legal agreements that describe a set of conditions or requirements to which users must agree in order to use the services offered by an organization. Terms of service (ToS) are variously referred to as *terms of use*, *terms and conditions* or *user agreement*. While there are no laws that require websites to post ToS, several laws mandate these disclosures in certain cases. For example, the Uniform Commercial Code in the USA contains requirements for making written disclaimers of warranties in cases where websites conduct commercial transactions. According to the Digital Millennium Copyright Act (DMCA), if an online service provider wishes to be protected from copyright infringement suits for content posted on websites that it is hosting, then it needs to disclose contact information to receive claims of copyright infringement. It is therefore expected that users read and understand the ToS provided by an organisation before using the services offered. Users are often faced with major challenges while reading ToS as they tend to be lengthy and confusing. A study revealed that when users were presented with a ToS document, they spent an average of 51 seconds reading the document when it should have taken them 15-17 minutes to read the document completely [7], and that information overload was a significant factor that impacted reading behavior. These documents are rarely read and seldom help in consumer decision making.

While ToS are usually posted on a website along with the website privacy policy, to the best of our knowledge, there is no work aimed at automatically analyzing and clarifying ToS documents. ToS;DR is a solitary initiative towards manually annotating ToS and privacy policies. ToS;DR aims to help web users understand the data practices and ToS agreements of websites. However, such manual approaches are hard to scale and annotations exist for a very few documents. Prior efforts in automatically analyzing legal agreements have been mainly concerned with privacy policy documents with several corpora [9, 10] and techniques such as question answering [8], vagueness detection [4] and detection of compliance issues [11] used to simplify privacy policies.

We therefore analyze ToS documents using topic modelling and build a large corpus to aid the application of natural language processing techniques in automatically simplifying ToS documents. Additionally, we estimate the availability of ToS documents on the web. Often, ToS documents are unavailable on websites. They might be presented as a link on the website, purported to lead to a page containing the document, but fail to fulfil the request. This presents a basic obstacle to notice and choice, with possible legal implications for the website operator. Understanding the types of unavailable documents, their frequencies , can help regulators and

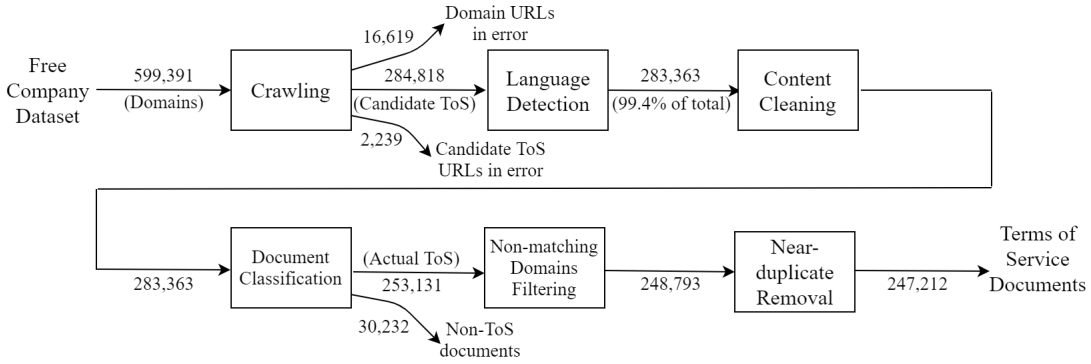

Figure 1: Document collection and classification pipeline for ToS

legal experts address the problem of unavailability and help users to make informed decisions.

We make the following contributions:

- We build the first large scale ToS corpus, containing 247,212 English language documents.[1]
- We estimate the overall availability of ToS documents.
- We conduct topic modelling and identify the various topics in ToS documents.

2 DOCUMENT COLLECTION

We used a list of company domains from the Free Company Dataset[2] provided by People Data Labs. The dataset is a collection of over 7 million global companies and their information. Our initial investigation of ToS indicated that not many company websites had the document. However, the ToS is necessary since the document delineates a set of conditions to which website users have to agree in order to use the services offered. We observed that company domains with a privacy policy were more likely to have ToS documents. In a random sample of 50 domains with at least one privacy policy in the data set, 50% (25) had the ToS document. Since the websites were obtained from a corpus of companies, and since privacy policies are used to educate users about companies' data practices, we assume that those websites that have privacy policies also provide goods or services to consumers. It would therefore be reasonable to expect those websites that have a privacy policy to also have a ToS document. In order to build the ToS corpus, we only selected company domains with a privacy policy.

Figure 1 shows the pipeline for extracting ToS documents from company domains that have at least one privacy policy. This pipeline was inspired from the work of Srinath et al. [9] to gather privacy policies from the web. 599,391 unique domains from Free Company Dataset were selected that provided at least one valid English language privacy policy. Manual examination of 100 company homepages revealed that a majority of ToS documents contained a hyperlink with the keyword *terms*. Therefore, we used Scrapy[3] to crawl the domain URLs to obtain *candidate* ToS documents by searching for HTML *href* attributes containing the word *terms* after case folding the characters. A total of 284,818 URLs were successfully crawled from 261,885 domains. We refer to these as *candidate ToS* as it is not determined if they are actually ToS documents. The remainder either did not have a ToS hyperlink on the landing page, did not qualify the URL selection criteria, had errors from crawling domain URLs (16,619) or had errors from candidate ToS URLs (2,239). As errors from candidate ToS URLs indicate a potential failure mode, we captured them for further analysis.

We detected the language of 284,818 successfully crawled candidate ToS documents using LangID [5], a language identification tool available as a Python library. It accepts a segment of text and returns the identified language of the text. It is capable of detecting 97 languages across different domains. Language detection step resulted in 99.4% (283,363) of documents in English.

We extracted the text of crawled English language candidate ToS web pages using Boilerpipe[4] [3] which is an open source tool for boilerplate removal. To separate non-ToS, we performed document classification for which we manually labelled 1,000 randomly selected documents with train-test sets split in the ratio 4:1. We trained a random forest classifier with TF-IDF features extracted from document text after tokenization and removing stop words. The precision, recall and F1 scores for the model are 0.935, 0.940 and 0.935 respectively. Documents with probability greater than or equal to 0.5 were considered as ToS.

In the next step, we removed ToS documents whose domains did not match the set of unique domains initially considered. This was done to accurately evaluate the availability of ToS as ToS hyperlinks of third-party websites embedded on domain web pages were also crawled during the crawling phase. Finally, we filtered near-duplicates from 248,793 ToS documents obtained from the previous step. We first grouped ToS belonging to the same domain. We then applied the simhash [2] algorithm on shingles [1] of size 3 for each document. Hamming distance between each pair of ToS within the same domain was calculated using the simhashes and near-duplicates were removed by choosing a hamming distance threshold [6] of 7.

[1] https://privaseer.ist.psu.edu/data
[2] https://docs.peopledatalabs.com/docs/free-company-dataset
[3] https://scrapy.org/
[4] https://github.com/misja/python-boilerpipe

Table 1: Keywords, %paragraphs (%P) and %documents (%D) corresponding an interpreted topic in ToS as a separate agreement

Interpreted Topic	Keywords	%P	%D
Use License	content, right, user, license, create_derivative, submission, grant, distribute	4%	39%
Intellectual Property	site, right, website, content, material, property, trademark, copyright	9%	67%
Payment Terms	order, product, credit_card, time, credit, price, card, purchase	7%	40%
Introduction	term, agreement, condition, service, provision, website, site, right	15%	75%
Indemnification	party, third, claim, including, service, agree, officer_director, directly_indirectly	6%	49%
Disclaimer of Warranties	site, website, service, warranty, content, information, material, link	12%	76%
Termination	right, material, otherwise, content, comment, website, discretion, reserve	4%	35%
Circumstances beyond Reasonable Control	customer, good, shall, seller, delivery, buyer, password, control	7%	35%
Limitation of Liability	damage, loss, liability, limitation, including, direct_indirect, liable, service	5%	62%
Copyright Infringement Claims	notice, email, address, communication, copyright, mail, message, mobile	3%	28%
User Eligibility	service, information, user, data, access, account, website, provide	20%	71%
Dispute Resolution	shall, client, arbitration, payment, court, subscription, dispute, agreement	9%	46%

3 AVAILABILITY OF TERMS OF SERVICE

Dead Links to Candidate ToS. While crawling for ToS, we recorded the URLs that led to error pages. Here the candidate ToS URLs that appear to link to the ToS document produces an error page. 2,239 URLs led to error pages, out of which 2,054 had error types indicating unavailability of ToS. 94.4% (1,940) had *HTTP Errors* and 404 Not Found was the most dominant error code. There were 88 instances of *DNS Lookup Error*. 14 URLs referencing localhost returned *Connection Refused Error*. 10 links had *Value Error* that comprised of incorrect or misspelled URLs on the website landing page and 2 URLs returned *File Not Found Error*.

Non-ToS Documents. 10.7% (30,232) of the candidate ToS documents were classified as non-ToS documents. On analyzing their contents, we found that they belonged to three main categories: pages containing glossary of terms, ToS document web pages with no text and ToS pages that contained only the heading with empty content in the body. This shows the existence of websites which provide a link that navigates to a valid ToS web page without any text, thereby implying unavailability of the document.

4 TOPIC MODELING

We obtained 247,212 unique English language ToS from a set of 582,700 domains containing at least one privacy policy. We identified that in a small percentage of websites, the links to both the privacy policy page and the ToS page led to the same web page, implying that those websites had the privacy policy and ToS on a single page. Based on URL comparison, we found 8,218 domains that contained a single page containing both the documents. We explored the content of such documents to verify if all topics typically present in a typical privacy policy or a ToS were available.

We used Latent Dirichlet Allocation (LDA), an unsupervised approach to topic modeling that extracts the most probable distribution of words into topics through an iterative mechanism. We considered two classes of documents: (1) Documents having privacy policy and ToS on a single page i.e. single page agreements and (2) Documents having a separate ToS page. We took 3000 documents of each kind and split them into paragraphs. The paragraphs were preprocessed using a utility function in Gensim[5]. Additionally, stop words were removed and bigram and trigram phrases were added to the list of tokens. The LDA model was then applied to the preprocessed paragraph texts. The highest coherence score was obtained when the number of topics was set to 12 for both document types.

Table 1 shows the keywords and interpreted topics for documents having ToS content only. The third column shows the percentage of paragraphs (%P) out of 49,801 that associate to each of the topics, and the fourth shows the percentage of documents (%D) out of 3,000 ToS that have a particular clause as given in the first column. It can be inferred that the most prevalent topic is *User Eligibility* (%P = 20%) followed by *Introduction* (%P = 15%) and *Disclaimer of Warranties* (%P = 12%). At the document level, more than 70% cover *User Eligibility*, *Disclaimer of Warranties* and *Introduction* confirming the prevalence of these topics across ToS documents.

Table 2 shows the keywords and interpreted topics for documents having the privacy policy and ToS on a single page. The third column in table 2 shows the percentage of paragraphs (%P) out of 59,921 that associate to each of the interpreted topics, and the fourth shows the percentage of documents (%D) out of 3,000 single page agreements that have a particular clause as given in the first column. The rows highlighted in gray are the topics available in a typical privacy policy. Regarding paragraphs, we observe that *Privacy Policy Introduction* (29%), *Changes* (11%) and *Cookies* (11%) have the highest percentages and all of them are actually sections included in a privacy policy document. This indicates the prevalence of privacy policy topics as compared to ToS topics on single page agreements thereby suggesting the importance of privacy policy over ToS. On analyzing at the document level, we observe that 88% and 77% of documents have the topics *Privacy Policy Introduction* and *Changes* respectively, implying that a large percentage of single page agreements cover privacy policy as compared to ToS topics.

5 DISCUSSION

Availability Analysis. ToS documents were crawled from websites with at least one privacy policy and hence for our calculations, we make the assumption that websites with a privacy policy provide some form of service to users. Hence all such websites should be expected to have a ToS. After the crawling stage in the document collection and classification pipeline, 261,885 domains generated at least one successful candidate ToS document. 1,960 domains generated at least one candidate ToS with dead links. Since 94.4% of dead links to candidate ToS document returned HTTP errors, we assume

[5]https://pypi.org/project/gensim/

Table 2: Keywords, %paragraphs (%P) and %documents (%D) corresponding to an interpreted topic in single page agreements

Interpreted Topic	Keywords	%P	%D
Gathering Personally-Identifying Information	information, personally, user, identifiable, identifying, website, submission, comment	3%	28%
Third-Party Links and Resources	party, third, website, site, content, link, service, responsible	7%	67%
Disclaimer of Warranties	site, warranty, service, content, information, material, website, merchantability_fitness	8%	64%
Termination	service, right, notice, resolve, arbitration, sole, discretion, reserve	4%	33%
Cookies	information, website, cooky, site, service, browser, user, collect	11%	65%
Privacy Policy Introduction	information, personal, data, service, provide, email, purpose, privacy	29%	88%
Intellectual Property	right, content, property, material, trademark, service, otherwise, copyright	7%	55%
Changes	term, service, website, site, policy, change, condition, privacy	11%	77%
Limitation of Liability	shall, damage, agreement, loss, provision, liability, term, limitation	8%	61%
Indemnification	attorney_fee, party, claim, officer_director, employee, hard_drive, agent, including	3%	33%
Use License & Accuracy of Materials	medium, material, social_medium, social, digital, network, licensee, site	2%	27%
Payment Terms	credit_card, order, payment, customer, seller, buyer, product, credit	7%	38%

that a large percentage of candidate ToS with dead links would have actually pointed to a valid document if not for the HTTP error on the document web page. Hence, considering websites in the data set that have ToS hyperlink on the landing page with hyperlink containing the keyword *terms*, we estimate that **0.74%** of websites contain dead links to ToS. At the end of the document collection and classification pipeline, 234,755 domains in the data set produced at least one valid ToS document. After subtracting the number of domains in error (16,619) from the total number of unique domains initially considered (599,391), we obtain the number of successful domains crawled as 582,700. Therefore, we infer that only **40.3%** of domains in the data set have a valid ToS.

Topic Modeling Analysis. On comparing single page and ToS as separate agreements, we find the interpreted topics *Termination*, *Use License & Accuracy of Materials*, *Intellectual Property*, *Payment Terms*, *Indemnification*, *Disclaimer of Warranties* and *Limitation of Liability* common to both. We observe that the most prevalent topics in ToS documents such as *User Eligibility* and *Introduction* are not covered in single page agreements. However, there is a significant overlap in the topics between the two classes of documents. We also observe that topics covered by more than 60% of ToS documents, such as *Disclaimer of Warranties*, *Limitation of Liability* and *Intellectual Property* fall under the overlap. This means that they are also available in single page agreements and results show that close to 60% of such agreements contain these topics. These findings suggest that single page agreements serve to provide most of the information except for a handful of prevalent topics.

We compared the interpreted topics of single page agreement and privacy policy. The topics in a privacy policy were obtained from the work of Srinath et al. [9] in which LDA topic modeling is used to explore the distribution of privacy practices in their corpus. In this case, we observe that though single page agreements cover prevalent topics such as *Changes* and *Cookies*, the topic overlap is small consisting of only three topics: *Gathering Personally-Identifying Information*, *Cookies* and *Changes*. Single page agreements miss some of the most prevalent topics or sometimes mandatory disclosures of privacy policy such as *First Party Collection*, *Third Party Collection*, *Data Security & Contact* and *European Audience*. The absence of the topic *European Audience* indicates that templates of single page agreements currently in use are not provide a GDPR-compliant. Thus, with respect to privacy policy, single page agreements do not provide sufficient information and are not very useful.

6 CONCLUSION

We built a first of its kind ToS corpus containing 247,212 English language documents by crawling a set of company websites and processing them through a data collection and classification pipeline. We analyzed the unavailability of ToS based on the results obtained at different stages of document collection and classification pipeline. We saw failure modes such as broken links and empty content in the ToS document web page. We estimate that more than 50% of the websites do not provide a ToS document. We identified that some websites navigate to the same page when users attempt to access either the privacy policy or the ToS document. Topic modeling results suggest that a large percentage of single page agreements miss some of the most prevalent topics in ToS. Future work could include fine grained content analysis of documents based on geography, sector of commerce etc. since disclosures in ToS heavily rely on the jurisdiction and the type of organization.

REFERENCES

[1] Andrei Z Broder, Steven C Glassman, Mark S Manasse, and Geoffrey Zweig. 1997. Syntactic clustering of the web. *Computer networks and ISDN systems* 29, 8-13 (1997), 1157–1166.

[2] Moses S Charikar. 2002. Similarity estimation techniques from rounding algorithms. In *Proc. STOC*. ACM, 380–388.

[3] Christian Kohlschütter, Peter Fankhauser, and Wolfgang Nejdl. 2010. Boilerplate detection using shallow text features. In *Proc. WSDM*. 441–450.

[4] Logan Lebanoff and Fei Liu. 2018. Automatic Detection of Vague Words and Sentences in Privacy Policies. In *Proc. EMNLP*. 3508–3517.

[5] Marco Lui and Timothy Baldwin. 2012. langid.py: An Off-the-shelf Language Identification Tool. In *Proc. ACL System Demonstrations*. Jeju Island, Korea, 25–30.

[6] Gurmeet Singh Manku, Arvind Jain, and Anish Das Sarma. 2007. Detecting near-duplicates for web crawling. In *Proc. Web Conference*. ACM, 141–150.

[7] Jonathan A. Obar and Anne Oeldorf-Hirsch. 2020. The biggest lie on the Internet: ignoring the privacy policies and terms of service policies of social networking services. *Information, Communication & Society* 23, 1 (2020), 128–147.

[8] Abhilasha Ravichander, Alan W Black, Shomir Wilson, Thomas Norton, and Norman Sadeh. 2019. Question Answering for Privacy Policies: Combining Computational and Legal Perspectives. In *Proc. EMNLP*. 4949–4959.

[9] Mukund Srinath, Shomir Wilson, and C. Lee Giles. 2020. Privacy at Scale: Introducing the PrivaSeer Corpus of Web Privacy Policies. *arXiv preprint arXiv:2004.11131* (2020).

[10] 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 Norton, Eduard Hovy, Joel Reidenberg, and Norman Sadeh. 2016. The Creation and Analysis of a Website Privacy Policy Corpus. In *Proc. ACL*. 1330–1340.

[11] Sebastian Zimmeck, Peter Story, Daniel Smullen, Abhilasha Ravichander, Ziqi Wang, Joel Reidenberg, N Cameron Russell, and Norman Sadeh. 2019. Maps: Scaling privacy compliance analysis to a million apps. *In Proc. PETs* 2019, 3 (2019), 66–86.