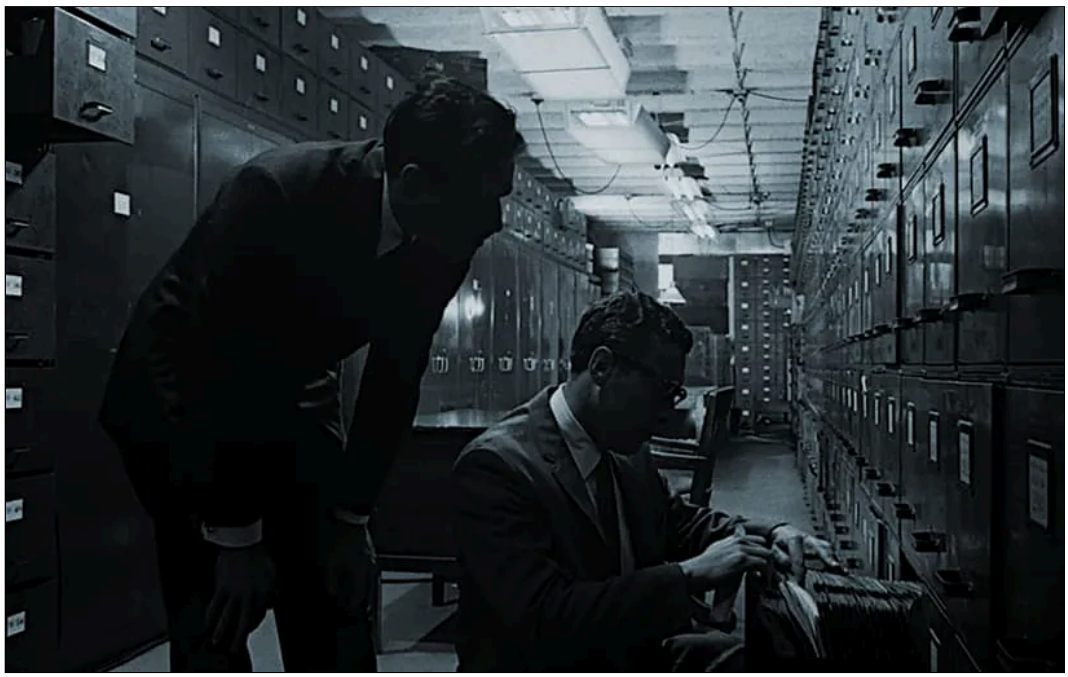
Analyzing 25 Years of Privacy Policies with Machine Learning

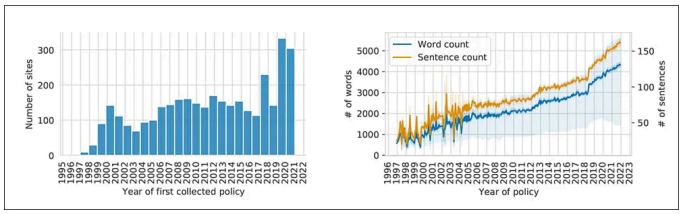
Originally published at https://www.unite.ai/analyzing-25-years-of-privacy-policies-with-machine-learning/



A recent study has used machine learning analysis techniques to chart the readability, usefulness, length and complexity of more than 50,000 privacy policies on popular websites in a period covering 25 years from 1996 to 2021. The research concludes that the average reader would need to devote 400 hours of 'annual reading time' (more than an hour a day) in order to penetrate the growing word counts, obfuscating language and vague language use that characterize the modern privacy policies of some of the most-frequented websites.

The report states:

> *'The average policy length has almost doubled in the last ten years, with 2159 words in March 2011 and 4191 words in March 2021, and almost quadrupled since 2000 (1146 words).'*

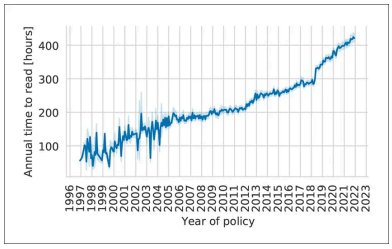


The mean word count and sentence count among the corpus studied, over a 25 year period. Source: https://arxiv.org/pdf/2201.08739.pdf

Though the rate of increase in length spiked when the GDPR and the California Consumer Privacy Act (CCPA) protections came into force, the paper discounts these variations as 'small effect sizes' which appear to be insignificant against the broader long-term trend. However, GDPR is identified as a possible cause of growing 'vague' language in policies (see below).

Assuming a reading speed of 250 words per minute, the paper contends that the average privacy policy now takes 17 minutes to read, while more popular policies (i.e. policies associated with a high number of users) take 23 minutes to complete.

The longest policy in the dataset, from Microsoft, requires 152 minutes to consume, according to the research, which leveraged a number of variants on Google's BERT language model.



Growth in the rate of annual hours needed to read modern privacy policies, assuming that the reader visits 1462 unique websites per year.

Much of the recent increase in verbosity and ambiguity in privacy policies is attributed by the paper as a reaction to attempts over the last two decades to impose regulations, but also to the disingenuous use of regulatory compliance requirements as an excuse to stealthily increase the scope and opacity of privacy policies.

> *'Overall, our results show that recent privacy regulations have not substantially improved the privacy of users online, but rather led to more bloated privacy policies that describe more and more invasive data practices.'*

Though a number of Natural Language Processing (NLP) papers have addressed the readability and other aspects of privacy policies in recent years, the author believes that this is the first project of its type to provide such a broad overview of policy development in recent decades.

The paper is titled *Privacy Policies Across the Ages: Content and Readability of Privacy Policies 1996–2021*, and comes from Isabel Wagner at the Cyber Technology Institute of De Montfort University in the UK.

Elliptical Language

The report also suggests that the average number of 'obfuscating words' (i.e. *acceptable*, *significant*, *mainly*, and other words that do not provide definitive meaning) in privacy policies increased steadily up to 2018, but then shot up from a median of 227 around March of 2018 to 304 in June of 2020.

The author contends that this rise is attributable to the effects of GDPR, and the paper finds that over two thirds (72%) of sentences in the privacy policies studied contained a minimum of one obfuscating word.

Readability

Across three common measures of reading difficulty, the study found that *'privacy polices have become increasingly hard to read over the years'*. The authors estimate that 41% of current-applicable policies available in 2021 had a median Flesch Reading Ease (FRE, higher is better) of just 31.8, with the author observing *'This score indicates a very difficult text that is best understood by university graduates'*.

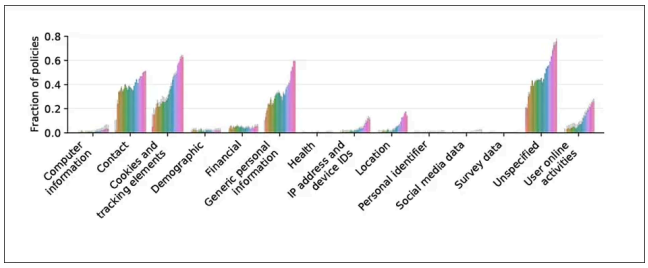
At the same time, only *6.7%* of the policies achieved an FRE score above 45 (which, the report notes is the reading standard required for insurance policies in the state of Florida).

Policy Change Awareness

The work also addresses the extent to which privacy policies include details about how the potential consenter will eventually be notified in the event of subsequent updates, which may affect the user's willingness to maintain the agreement.

The author observes:

> *'In 2021, 73% of policies include a statement about policy change. Of these, 34% state that changes will be announced by a notice in the privacy policy, 37% will post a notice on the website, and 22% will send a personal notice (the remaining policies leave the notification type unspecified).*
>
> *'As a result, most users are unlikely to become aware of changes in privacy policies.*
>
> *'In addition, users are offered almost no meaningful choice when policies change. Of the policies that notify the user of changes, only 12% offer a new opt-in, whereas 34% give no choice and 54% leave it unspecified.'*



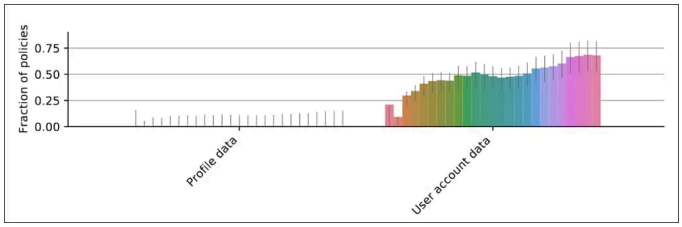
The paper's findings on the described methods for notifying users about policy changes.

Limited Choice Regarding Tracking

According to the study, a far greater range of mechanisms are offered in privacy policies for accessing user-account information than for accessing user profile data. Profile data can be created and updated through automated and non-obvious mechanisms, whereas user account data is not only explicitly granted by the user, but also obliged to be editable under regulations of various jurisdictions.

Consumer choice over cookie consent in privacy policies (a topic that has attracted heated debate since the advent of GDPR promulgated hundreds of thousands of cookie consent popups for EU instances of international and European websites) is generally addressed in the policies, but hides a more important layer of less accessible data*:

> *'[The] choices regarding cookies are insufficient to protect users from all tracking because choice or control mechanisms are rarely offered for* computer information*,* device identifiers*, and* personal identifiers*, which allow tracking users via fingerprinting.'*



A stark contrast in the available level of control granted by privacy policies between profile data (which may be obtained by implicit or stealthy means) and user account data (where some measure of control is frequently mandated by GDPR, the California Consumer Privacy Act (CCPA), and similar national and regional mechanisms).

Data

To obtain the data for the study, the author crawled websites for links to their privacy policies, frequently finding it necessary to widen the scope beyond the initial result, due to the number of non-integral policies that link out to further policies (each of which has potential to change either in tandem with or independent of the parent or related policy).

The Wayback Machine was used to obtain historical policies, though it was necessary when considering results to account for policies which had been blocked from crawling or archiving via a robots.txt configuration file (a small text file containing instructions to web-crawling indexing agents regarding pages and other entities that they should not include in a public index).

One snapshot per month was obtained from the Wayback Machine by its CDX API for each identifiable and continuous applicable policy, using Firefox under Selenium. Performing optical character recognition on policies only available in PDF format was not considered for the project, which limited itself to the (far greater) number of available HTML policies.

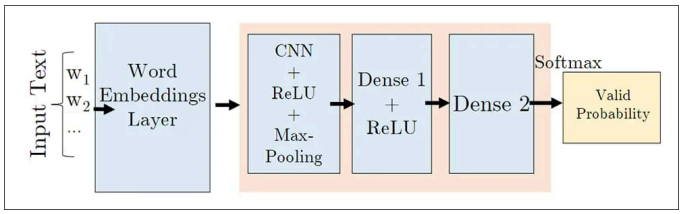
One interesting result from the project is that the clarity and readability of pornographic websites has actually improved over the studied interval – possibly in anticipation of growing calls for increased regulation and clarity. In order to gather these documents, it was necessary to obtain them with additional crawls from residential IP addresses, due to the university's content-blocking protocols.

Initially 1,068,683 documents were obtained, equaling 120,265 unique documents containing an average of 39.1 policy articles or clauses and 4.4 unique policy texts for each link.

English Only

As is common in similar recent studies, the project was not able to address non-English privacy policies, which were discarded during the data-cleaning stage using the PYCLD2 package.

To distinguish privacy policies from other types of material, the project used a classifier developed in 2019 as a joint initiative from the University of Wisconsin and the École Polytechnique Fédérale de Lausanne.



Architecture of the IS-POLICY classifier. Source: https://arxiv.org/pdf/1809.08396.pdf

Though the IS-POLICY classifier was trained on the same 1,000-document corpus as in the originating paper, the author had to obtain new non-policy documents for training, since the original sources were not available.

After filtering, the data was reduced to 56,416 unique privacy policies.

* *The paper's inline citation is converted to a hyperlink here, italic toggling is from the paper.*

First published 31st January 2022.

Writer on machine learning, domain specialist in human image synthesis. Former head of research content at Metaphysic.ai, until its dissolution into DNEG's Brahma.ai.
Portfolio site: martinanderson.ai
Contact: martin@martinanderson.ai

Discover More