Analyzing Privacy Policies as Data

A privacy policy is a written statement that covers how an enterprise collects, handles and processes personal data. It ensures that the enterprise complies with regulatory standards for specific industries and publicly communicates to users how their data are collected, stored and shared. Taking privacy policies in isolation provides the views and perspectives of only one organization. But what if an organization's privacy policy could be improved by comparing it to other organizations' privacy policies and requirements? Moreover, how can different privacy policies be consistently compared? A possible explanation is found in text mining.

Text mining is a method of transforming data from a text file to numerical values, making it possible to identify the frequency, context and similarities of key terms used in documents such as privacy policies. Information retrieval is a technique used in text mining that searches for specific terms and provides an output based on the results from one or multiple documents. Grouping or centralizing multiple documents in the same data set allows comparability, increasing the scale of the analysis and extending it across different documents. In other words, based on queries, a group of privacy policies can be searched to identify and compare the most relevant terms they contain or to highlight whether a relevant term or definition is not present in that group of policies.

The practical application of this technique allows enterprises to improve controls related to privacy policies by considering the proper use of privacy terms, developing a common vocabulary for a particular industry or sector, and observing the use of relevant terms prescribed by regulators. This granular control of key and mandatory terms can increase accuracy and align the privacy policy's content with user and business needs.

Data Analysis

To demonstrate the use of information retrieval, the current privacy policies of Meta,[1] Google[2] and WhatsApp[3] were selected, as these three are among the largest social media platforms.[4] A CRAN Project package called Quanteda was used to tokenize the words used in the documents.[5,6] The tokenization process converts each word from text to values that can be used for data analysis. This results in multiple tokenized privacy policies that form a document-feature matrix. The text can be segmented into sentences and words. Once the main functions are defined, the various policies can be processed to assess word frequency, context and similarities. During data preparation, commands were introduced to ignore punctuation, digits and URLs. In this analysis, the queries ranked the outputs using an ordered list of relevant documents, referring to each privacy policy token identified (**figure 1**).[7]

A word cloud shows the 30 tokens with the highest frequency in the data set of the three privacy policies (**figure 2**). The words "information," "services" and "use" occurred most frequently.

FIGURE 1
Privacy Policy Tokens

Text	Enterprise	Year	Types	Tokens	Sentences
Text1	Meta	2022	1,264	5,861	230
Text2	Google	2022	1,268	5,603	190
Text3	WhatsApp	2021	1,022	4,453	171

THIAGO DE OLIVEIRA TEODORO | CISA, CDPSE

Is a consultant focusing on governance, risk and compliance (GRC). He has 15 years of professional experience in auditing and internal controls in the public and private sectors.

FIGURE 2

Highest Frequency Tokens Word Cloud

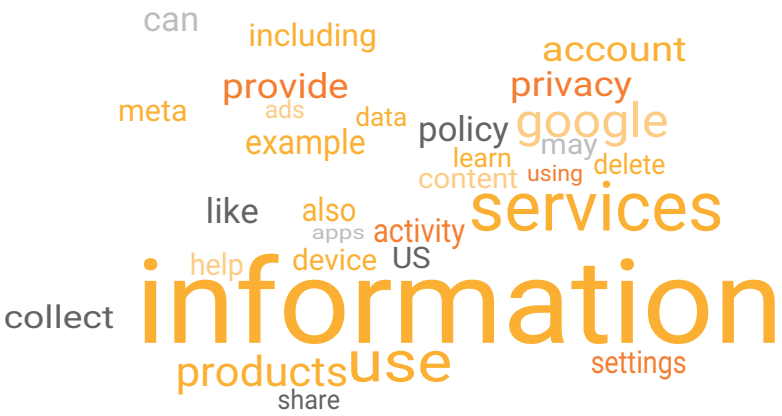

FIGURE 3

TF-IDF Results

Item	Feature	Frequency	Rank	Document Frequency	Group
1	Google	20.074404	1	2	All
2	WhatsApp	13.836516	2	1	All
3	Search	10.496668	3	1	All
4	Meta	10.037202	4	2	All
5	Personal	9.542425	5	1	All
6	Sites	8.111061	6	1	All
7	YouTube	6.202576	7	1	All
8	Saved	4.771213	8	1	All
9	Facebook	4.754464	9	2	All
10	Them	4.578373	10	2	All

FIGURE 4

KWIC Results: Use of "Data" in US Privacy Act of 1974

Location	Context Before	Word Searched	Context After
Text1, 9	when implementing the open	data	policy agencies shall incorporate
Text1, 81	implementation of the open	data	policy to facilitate effective
Text1, 91	implementation of the open	data	policy I direct the
Text1, 107	issuance of the open	data	policy the [chief information officer] CIO and
Text1, 131	integration of the open	data	policy into their operations
Text1, 165	adoption of the open	data	practices b within 90

Methods

The Quanteda package allows an analysis and a granular review of key terms using the information retrieval technique. Three methods were used to explore the data set of the three privacy policies:

1. Term frequency-inverse document frequency (TF-IDF)
2. Key words in context (KWIC)
3. Cosine similarity

TF-IDF computes scores and weights the importance of a word by searching for it in the different documents. This numerical statistic is higher if a word is unique to one specific document and lower if the word appears in many documents. **Figure 3** shows the top 10 results of this analysis. It is not surprising that all three social media platforms' names have a high frequency. However, the terms "search" and "personal" occur more frequently in, or are unique to, the Meta privacy policy. In the Meta policy, the words "search" and "personal" are used in different contexts, such as referring to many types of "research" and "personalized" services. If the scope of the analysis includes verifying how enterprises address "personal" information, variations in the use of the term can be highlighted and benchmarked. In the social media industry, "personalized" is a business term connecting people and content. In this case, the term "personalized" is used more often in the Meta policy than in the other two, moving it up in rank. Interestingly, the Meta privacy policy does not expressly mention "personal" information.

KWIC can help explain the intended meaning of a given term by analyzing the words found before and after the searched term. Most policies include a glossary of terms and their definitions, but how the term is actually used within the policy can provide a better understanding of its context.

One possible application of this method is to compare an existing regulation and current privacy policies. Using the search term "data," partial results are shown in **figures 4** and **5** comparing the US Privacy Act of 1974, as amended[8] and the privacy policies of the three social media platforms (**figure 5**). There is a distinct difference. Whereas the Privacy Act emphasizes "open data" and the government's role, the social media platforms emphasize appropriate disclosure of the use of collected, stored and shared

personal data. Ideally, when considering compliance with regulations, it is important to use a common vocabulary, with specific terms and definitions to increase accuracy.

The functions discussed can help identify the differences between privacy policies, benchmark best practices, and reveal opportunities for improvement by querying specific and more granular terms.

Another application of the KWIC method is to compare privacy policies in the same industry, such as benchmarking best practices. "Consent" is a key term used in privacy policies, but its scope and application can vary among jurisdictions. In Canada, for example, "informed consent" or "meaningful consent" is required. This implies that consent must relate to the purpose for which the information is required and must consider the sensitivity of the personal information. In addition, ordinary users must understand what consent means and that they are allowed to withdraw such consent.[9] It is a good practice to refer to these aspects of consent and consider how the enterprise or the industry aligns with regulatory requirements. Consent is required to disclose personal information; however, such consent may not be required for specific cases treated as exceptions. For instance, the US Privacy Act lists 12 exceptions.[10] Performing a granular analysis for the term "consent" results in only six items from the three social media privacy policies. The term occurs more frequently in Google's privacy policy (text2) than in WhatsApp's (text3) and is not used in the Meta (text1) privacy policy at all (**figure 6**).

Cosine similarity measures the frequency of a particular word or sentence in a document by searching (querying) and creating a rank.[11] Based on term frequency vectors, the function measures how similar two documents are (the angle between the vectors). A comparison of the three privacy policies shows that Google's policy is slightly more similar to Meta's (89.54 percent) than to WhatsApp's (87.47 percent), even though Meta owns WhatsApp. Preparing a data set of two privacy policies permits the level of similarity between them to be measured. Benchmarking two policies can identify whether there is a significant change between a new policy and an old one or how one enterprise is positioned relative to another enterprise. Comparing the policies of Google and Snap Inc. (Snapchat), for instance, reveals an 82.61 percent similarity. Thus, the cosine similarity can provide a baseline to compare privacy policies.

Conclusion

Analyzing privacy policy as data is a suitable approach for comparative studies that support research in different jurisdictions to identify unique topics. It is also a valuable tool for exploring areas of particular interest among different privacy policies.

FIGURE 5

KWIC Results: Use of "Data" in Social Media Privacy Policies

Location	Context Before	Word Searched	Context After
Text1, 1124	you visit and cookie	data	like through social plugins
Text1, 2629	and similar technologies including	data	that we store on
Text1, 2959	and advertising vendors and	data	providers who have the
Text2, 948	G.P.S. and other sensor	data	from your device IP
Text2, 990	The types of location	data	that we collect and
Text2, 1205	web storage or application	data	caches databases and server

FIGURE 6

Use of "Consent" in Social Media Privacy Policies

Location	Context Before	Word Searched	Context After
Text2, 2002	we will ask for your	consent	before using your information
Text2, 2937	following cases with your	consent	We will share personal information
Text2, 2949	when we have your	consent	For example if you
Text2, 3005	ask for your explicit	consent	to share any sensitive
Text2, 4171	privacy policy without your explicit	consent	We always indicate the
Text3, 3292	want to revoke your	consent	to our use of

The functions discussed can help identify the differences between privacy policies, benchmark best practices, and reveal opportunities for improvement by querying specific and more granular terms. This approach is equivalent to an inductive inference that begins with a specific observation and results in a general explanation. For better results, it is essential to review the actual privacy policy for validation and consistency. It is important to use a common vocabulary when referring to privacy principles and frameworks, keeping in mind that some policies may use a synonym or refer to terms in a specific way that is not captured by the function's output result.

LOOKING FOR MORE?

- Read *Privacy in Practice Survey 2023. www.isaca.org/resources/reports/privacy-in-practice-2023-report*
- Learn more about, discuss and collaborate on privacy in ISACA's Online Forums. *https://engage.isaca.org/onlineforums*

Endnotes

1 Meta, "Privacy Policy," 26 July 2022, *https://mbasic.facebook.com/privacy/policy/printable/*
2 Google, "Privacy Policy," 4 October 2022, *https://www.gstatic.com/policies/privacy/pdf/20221004/a2n853ky/google_privacy_policy_en-GB.pdf*
3 WhatsApp, "Privacy Policy," 4 January 2021, *https://www.whatsapp.com/legal/privacy-policy*
4 Statista, "Most Popular Social Networks Worldwide as of January 2022, Ranked by Number of Monthly Active Users," *https://www.statista.com/statistics/272014/global-social-networks-ranked-by-number-of-users/*
5 Benoit, K. *et al*.; *Quantitative Analysis of Textual Data*, CRAN, 8 December 2022, *https://cran.r-project.org/web/packages/quanteda/quanteda.pdf*
6 Quanteda, "Quantitative Analysis of Textual Data," *https://quanteda.io/*
7 Teodoro, T.; "Privacy Policy as Data," Github, 2022, *https://github.com/thiago-teodoro/Privacy-Policies-as-Data*
8 US Government, Privacy Act of 1974, as amended, 5 U.S.C. §552a, USA, 1974, *https://www.govinfo.gov/content/pkg/USCODE-2018-title5/pdf/USCODE-2018-title5-partI-chap5-subchapII-sec552a.pdf*
9 Office of the Privacy Commissioner of Canada, The Personal Information Protection and Electronic Documents Act (PIPEDA): Fair Information Principle 3—Consent, Canada, August 2020, *https://www.priv.gc.ca/en/privacy-topics/privacy-laws-in-canada/the-personal-information-protection-and-electronic-documents-act-pipeda/p_principle/principles/p_consent/*
10 Department of Justice, *Overview of the Privacy Act: 2020 Edition*, USA, *https://www.justice.gov/opcl/overview-privacy-act-1974-2020-edition/disclosures-third-parties#exceptions*
11 Han, J.; M. Kamber; J. Pei; "Getting to Know Your Data," *Data Mining: Concepts and Techniques, 3rd Edition*, Morgan Kaufmann, USA, 2012, *https://doi.org/10.1016/B978-0-12-381479-1.00002-2*