Analyzing Data Privacy Statements Using Transformers

Michael Sneberger
Laela Olsen
Ruiyu Huang
Lavany Deepak Jadhav

ABSTRACT

In this study, we conduct a systematic investigation into the completeness of Data Privacy Statements (DPS) that are complaint with the California Consumer Privacy Act (CCPA). We employed a multi-stage methodology, combining automated web scraping techniques with rigorous manual verification to compile a dataset of 392 DPS pairs from highly trafficked websites. This methodical data collection resulted in the delineation of 230 distinct pairs of DPS categorized by their adherence to CCPA guidelines. Our preliminary analyses were performed using OpenAI's GPT-4, showcasing the practical application of advanced transformers in interpreting legal texts. The study further delves into heuristic evaluations based on transformer-extracted data, examining the influence of CCPA mandates on the granularity of information provided to consumers within DPS. Enhanced analytical methods were also applied, utilizing BERT and GPT embeddings and kmeans and DBSCAN clustering algorithms, to distill significant disclosure patterns. This research not only paves the way for future explorations within computational data privacy analytics but also demonstrates the effectiveness of natural language processing tools in extracting meaningful insights from legal documentation, particularly in the realm of completeness DPS.

1 INTRODUCTION

Data Privacy Statements (DPS aka Policies or Notices[1]) are often required by law (*e.g.* GDPR[2], CCPA). Since January 1, 2020 the California Privacy Protection Act[3] (CCPA) and it associated regulations[4] have required that a DPS containing prescribed information be presented to website visitors. The goal of the research presented in this paper is to determine if these regulations have improved the amount of information presented to website visitors by way of DPS. Prior researchers have noted that DPS are "written by lawyers for lawyers" and often long and difficult to parse thus making it difficult to do research on the contents of DPS at scale. [5] The goal of the project is to use natural language processing and/or transformers to compare non-California and California DPS that are presented simultaneously by the same website in order to gain insights about the contents and other properties of these DPS, and if the prescriptive requirements of CCPA have led website operators to provide additional information to site visitors.

2 RELATED WORK

Significant, high-quality research has been done in the area of DPS, both before and after GDPR and CCPA changed the landscape for DPS. Early work focused on making DPS easier to read by mechanized methods:

- In 2009 Ghazinour, *et al.* built upon (now abandoned) strategies for machine-readable DPS[5] and proposed their own Privacy Policy visualization Modeling (PPVM) to facilitate a graphical visualization of DPS. [2]
- in 2013 Sadeh, *et al.* used natural language processing n conjunction with crowdsourcing in an attempt to create an interface based on user preferences that would automate he reading of DPS. [3]
- In 2018 Harkous, *et al.* again proposed a tool (called Polisis for Policy Analysis) but now in the form or a search function providing users to answers to questions they might have about a given DPS. [4]
- Also in 2018 Wilson, *et al.* again mixed crowdsourcing to facilitate a classification system for DPS content. [5]

whereas later work - well after the adoption of GDPR in the European Union - leaned more heavily into natural language processing, but still focusing on facilitation of the reading of DPS:

- in 2019 Ravichander, *et al.* used both convolutional neural networks and bidirectional encoder representations from transformers as well as experts to answer a compilation of over 1,000 questions about DPS. [6] They found that human experts performed better than machine learning when generating answers. [6]
- In 2020 Shvartzshanider, *et al.* developed a method of extracting specific privacy parameters from DPS using natural language processing and through a contextual integrity lens. [7]

finally, the most recent work leverages developments in data science to analyze DPS:

- In 2022 Sajidur Rahman, *et al.* used logistic regression, FastText, and BERT to develop a machine learning system they dubbed PermPress to be used to analyze permission requests related to mobile application DPS. [8]
- in 2023 Wagner surveyed the development of DPS from 1996 to 2021 using machine learning and natural language processing techniques. [9] The primary observation of the work was the increasing complexity and length of DPS.

[1] While the most popular term for these public-facing statements of data handling is privacy "policy," and CCPA itself uses the term "privacy policy," we agree with Counselor Hintze that a privacy policy is a inward facing document that instructs a data controller's employees on how to handle personal information, thus privacy statement is a more appropriate name for a publicly-facing statement posted on a website. [1]

[2] Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation) [2016] OJ L 119/1

[3] Cal. Civ. Code § 1798.100 – 1798.199

[4] Cal. Code Regs. tit. 11 § 7000 *et seq.*

[5] Platform for Privacy Preferences Project (P3P, and Enterprise Privacy Authorization Language (EPAL).

It should be noted that this prior work, though somewhat diverse, focuses on the contents of *individual* DPS. This differs from the focus of the instant research where we focus on whether the prescriptive requirements set forth in CCPA result in website users who are residents of California being better informed of the privacy practices of the operator of a website they visit: in other words, we compare the CCPA's effect on DPS aimed at residents of California as opposed to general DPS in order to gauge whether the prescriptive rules of the CCPA regarding DPS have lead to California website users being better informed regarding the collection and use of their personal data.

3 DATA

3.1 Web Scraping

In order to compare non-California DPS to purportedly California-compliant DPS we needed to develop a corpus of DPS featuring pairs of DPS presented on the same website: one designed to be CCPA-compliant, and one aimed at visitors who do not reside in California.[6] In order to scrape the text of such targeted pairs of DPS we took a series of steps:

- Using the Tranco library for Python gathered the base URLs for the top 100,000 websites.[7]
- In order to eliminate websites not potentially subject to CCPA made the decision to only move forward with the URLs on the raw list of 100,000 that ended in the .com and .net extensions,[8] resulting in 58,633 URLs to scan. To facilitate offline filtering, these results were saved offline in a .csv file.
- Using the BeautifulSoup library for Python we explored the .com and .net websites searching for the string "privacy" and if found, saving the deep-URL to save the location of privacy statements. If no such results were found a value of FALSE was returned and if it was found a value of TRUE was returned. This resulted in identifying 20,206 deep links that returned a TRUE value and appeared to potentially locate a DPS.
- A series of color-coded filtering was then performed offline on the TRUE results in order to identify potential pairs of non-CCPA and CCPA DPS:
  - The deep URLs were filtered for three strings: ccpa, cpra, and cali with results for each being color coded by filling the cell of the .csv holding the URL with a color
  - As the websites were scanned based on a traffic hierarchy from Tranco, the result rows were sorted alphabetically based URL column of the .csv in order to match repeated hits on the same base URL.
  - The Conditional Formatting/Highlight Cell Rules/Duplicate Values function was used in Excel to identify duplicate URLs in the alphabetized column of base URLs.
  - Finally, tuples of base URLs that matched up with a color-coded CCPA DPS were identified and if the base URL appeared to have both a CCPA DPS a non-CCPA DPS it was kept for further processing.
- These steps resulted in 392 websites being identified that appeared to have both a CCPA DPS and an non-CCPA DPS.
- The deep URLs for these purported pairs of non-CCPA and CCPA websites were then fed into a Python script - again using the BeautifulSoup library - in an attempt to scrape the text of the privacy policies and save them offline in a .csv file for later processing. Unfortunately, random checks of the text scraped showed that that while using BeautifulSoup to identify DPS was effective, it was not effective in cleanly extracting the text of the DPS and/or checking if the deep URL was indeed the correct locator for a DPS.

While there surely are more effective solutions to scraping the text of the seemingly located DPS,[9] given the limited number of targets, and the desire to have pristine copies of the text available for future natural language processing, it was decided to verify the scraping of DPS text by hand. This hand work proved fruitful for multiple reasons:

- While some of the deep URLs links were adjusted for improved correctness, 14 proved to be dead ends that did not lead to a DPS and were unfruitful thus eliminating 12 potential pairs.
- Within the remaining 380 potential pairings, the following was found:
  - Three of the pairs presented the exact same DPS for each half of the pair
  - 147 of the purported pairs blended their non-CCPA and CCPA DPS together, *i.e.*, the CCPA content was inside of a larger DPS (we called this the mixed category).
  - 230 of the pairs were found to have separate non-CCPA and CPPA DPS residing under different deep URLs (we called this the separate category).

It is these 230 pristine "separate" pairs where a single website presents to a visitor separate DPS: one aimed at residents of California, and one aimed at everyone else that will be the primary target of our further investigation.

4 METHODS

We began the first phase of our research by using a transformer/large language model to parse the target DPS pairs. The second phase used the output of the first phase and applied other Natural Language Processing (NLP) tools. The following is an outline of the processing techniques used to analyse the scraped data:

4.1 Using a Cutting Edge Transformer to Parse the DPS

To start analyzing the scraped data, and to become familiar with available cutting edge transformers, an OpenAI account was created

[6] By defining a "consumer" as "natural person who is a California resident" CCPA extends it protections only to residents of California. Since CCPA was enacted several states have enacted somewhat similar laws and many DPS that include CCPA language also call out other states for special treatment.

[7] All code used in the project can be found at: https://github.com/Sneberger/CSE572DMDvilsProject

[8] *E.g.*, non-profits are exempted from complying with CCPA. Cal. Civ. Code § 1798.140(d) which limits the definition of a "business" subject to CCPA as being "organized or operated for the profit or financial benefit of its shareholders."

[9] *See generally* [8]

and an API key obtained. In order to test the API a pair of DPS were chosen based on the CCPA version having a significant table structure presenting information. The idea was to see how the latest iteration of OpenAI's GPT would perform when simply being asked to separately summarize both the CCPA and non-CCPA DPS presented by the test website's DPS pairs.[10] As expected, GPT-4 Turbo did a fine job of summarizing the test DPS, including the data that was embedded in table format.

To fully use the capabilities of the GPT-4 Turbo transformer for parsing the scraped DPS pairs, by way of a Python script we role-prompted the tool with "[y]ou are an expert lawyer specializing in data privacy." We then developed the following series of questions to be asked of the input DPS:

(1) Please describe how data is being collected below?
(2) Please describe the data being collected below?
(3) Please list the data being provided to third parties below?
(4) Please describe how long data is retained below?
(5) Please describe what are users' rights related to the data items found in your response to question 2?

In order to provide output organized in such a manner as to facilitate the secondary analysis set forth below, a specific organization example was provided and we instructed GPT-4 Turbo to "[o]nly include the response in the above format without ANY additional information" and to "[o]nly include one piece of information per line. Do NOT include multiple pieces of information on the same line." This last instruction proved to be critical as during testing it was shown that without it, GPT-4 Turbo would truncate lists of items and end them with "etc." The output was saved offline in a .csv file which provided all scraped data as well as the GPT-4 output for each DPS thus providing not only a corpus of data for the analysis outlined below, but the entire data back-story of the of the project up to this point.[11]

4.2 A Heuristic Analysis of the GPT-4 Output

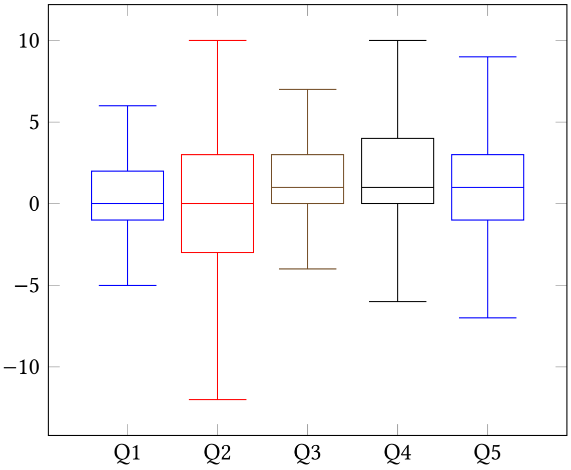
Since we asked GPT to return results in a formally numbered format we could easily parse the number of items returned in response to each question.[12] This set up an simplistic heuristic for examining if California DPS provide more information to website visitors. We wrote a Python script that parsed the GPT results for the non-California and California DPS of a pair separately and counted the number of items returned in response to each of the five questions separately. Results were returned such that if the counts for each DPS of the pair for the question matched, 0 was returned, and if the counts for one of the DPS exceeded the other an integer count of the excess number of items was reported: negative if the non-California DPS had more items, and positive if the California DPS had more items.

4.2.1 Heuristic Results. This data was analyzed with simple statistical measures. Here the "Net" column represents the total number of pairs (230) less the number of pairs returning a zero value indicating GPT returned the same number of items for each DPS of the pair.

230 Sep. Pairs	0	CA	Net	Mean	Median	Std.Dev.
Question 1	54	104	176	0.317	0.000	4.300
Question 2	29	107	201	0.322	0.000	4.968
Question 3	58	121	172	1.391	1.000	3.100
Question 4	76	119	154	1.678	1.000	3.368
Question 5	44	120	186	0.796	1.000	3.594

This removal of the zero values allows us to divide the CA column by the Net column for each question and see that of the pairs of DPS not reporting equal items, a higher percentage of California DPS presented more items, *e.g.*, 59, 53, 70, 74, and 64 percent for Questions 1-5, respectively, showing the California DPS of the pair provided more information than the non-California DPS. The mean, median, and standard deviation columns are based on all DPS of the pairs, including those where an equal number of items were found on each DPS of the pair.

Recalling that a positive values for the pair shows that the GPT result for the California DPS presented more items, note that all mean and median values are positive, showing in the aggregate that more items of information were mined by GPT for California DPS than for non-California DPS in individual pairs measured.[13] This is easier to see by how the quartiles above the median in a box plot have more area than below the median for four of the five questions.



While very simplistic, the strength of this heuristic analysis is in the strength of the GPT's parsing of the text of *individual* pairs of DPS. The heuristic relying on GPT's stellar job of mining individual items of information shows that we can get good results, and come to true conclusion with one, high-performance transformer and

[10] For an example separate pair of DPS see: https://www.compass.com/legal/privacy-policy/ and https://www.compass.com/legal/california-privacy-notice/, respectively

[11] The cost for the GPT-4 services used in this first phase totaled less than $55.00 including testing.

[12] In addition to the prompt language outlined above we instructed GPT-4 to "Please return in the following format:
1)
1.1 Item 1 how collected
1.2 Item 2 how collected
1.3 Items 3 how collected"

[13] When the mixed pairs of DPS were analyzed in the same way, the median for Questions 1 and 2 are negative showing different results that for the separate pairs. This could be because the California portion of a mixed pair DPS often relies on the non-California portion of the DPS and serves as an addendum for California. *See* Appendix 8.4.

some very basis statistical measures - no further natural language processing needed.

4.3 Analysis Using Clustering

Given the task was to identify if California DPS provided more information to website visitors, and GPT-4 had parsed the text of the separate pairs of DPS and listed items presented in the DPS, to preform a formalized NLP analysis of the GPT results we determined that clustering would be an appropriate method of analysis.

4.3.1 BERT Embeddings and Kmeans Clustering. Based on familiarity due the group using the tools in class labs and howework, we decided to try creating embeddings of the GPT output using Bidirectional Encoder Representations from Transformers (BERT), then using those embeddings as input to the kmeans clustering algorithm. The first difficulty that arose was choosing the number of clusters to ask for from kmeans. Using the elbow method was unsatisfactory as the plot resulting from testing a range of cluster counts presented a smooth arc, providing no helpful elbow. Generating silhouette scores was similarly unhelpful.

Ultimately, based on the maximum number of items returned by GPT when parsing the pairs of DPS (17), we decided to ask kmeans for 20 clusters from the GPT output for each of the five questions outlined in Section 4.1 above. A manual review of the 100 kmeans cluster showed that they were unhelpful across the board, grouping items by not-necessarily-significant terms. The BERT and kmeans methodology proved unfruitful.

4.3.2 GPT Embeddings and DBSCAN Clustering. In light or the perceived quality of GPT's output when asked to parse the DPS pairs, we switched to a different scheme. We asked GPT to produce embeddings of the previous produced GPT DPS parsing output. We then used these embeddings in a Density-Based Spatial Clustering of Applications with Noise (DBSCAN) algorithm. An immediate benefit to the new methodology was that DBSCAN decides an appropriate amount of clusters to create from the given data. Again clustering by each of the five questions individually, DBSCAN created 23, 71, 38, 42, and 20 clusters from questions one through five, respectively.[14]

A manual review of the DBSCAN-produced clusters showed they were *vast* improvements over the kmeans clusters, and provided the level of output that led to the use of clustering: specific, tight, unique, and informative clusters of items. An example of the quality of the clusters is found in Cluster 4 from the Question 2 parsed data. This cluster features 281 items, 267 of which includes the term 'Geolocation' and 14 of which contain the term "Location' referring to the type of data a DPS indicates is collected from a website user. The term 'Geolocation" only appears in one other cluster[15] with six items, Cluster 63. Reviewing other clusters showed similarly narrowly focused results.

4.3.3 Clustering Results. The items in the clusters are labeled to allow a for a heuristic review of how many items in a cluster are from a non-California DPS and how many are from a California DPS.[16] Let us take a look at the contents of the clusters in and aggregate sense. In the following table:

- the column labeled 'Clus.' presents the total number of clusters created by DBSCAN for the respective question
- the column labeled 'nonCA.' presents the number of clusters for the respective question that contain more non-California items than California items
- the column labeled 'Only' presents the number of clusters for the respective question that only contain items from non-California DPS
- the column labeled 'CA.' presents the number of clusters for the respective question that contain more California items than non-California items
- the column labeled 'CA Only' presents the number of clusters for the respective question that only contain items from California DPS
- the column labeled 'Dom" indicates which kind of DPS is shown to have more items clustered overall, *i.e.*, is dominant: non-California or California, with '—' meaning the results are inconclusive.

Questions	Clus.	nonCA	Only	CA	CA Only	Dom
Question 1	23	3	3	20	13	CA
Question 2	71	32	14	39	9	—
Question 3	38	11	3	27	17	CA
Question 4	42	9	8	33	22	CA
Question 5	20	7	5	13	6	CA

Reviewing the data presented in the table it can be seen that for questions 1 and 3-5 the number of clusters dominated by items from California DPS is clear for both those clusters that have items from both types of DPS and for clusters that contain items from a single type of DPS. For Question 2 on the other hand, there are nearly as many clusters dominated by non-California items and more clusters that only include items from non-California DPS

5 RESULTS

While the more sophisticated clustering analysis produced meaningful results, the heuristic analysis has a fundamental advantage. Recall that the goal here is to identify if the prescriptive rules of the CCPA and its accompanying regulations lead to website visitors from California being presented with more information that visitors from elsewhere. That necessarily begs for the comparison of non-California and California DPS *on the same website* for a true comparison.

While the clustering was performed on all items parsed by GPT in the aggregate, the heuristic was a direct comparison of the separate pair of DPS on a single website: one for California residents, and one for everyone else.

Overall, even if the heuristic is deemed to be more focused, the clustering confirms the conclusion that it appears that the prescriptive rules in California result in DPS of websites aimed at residents

[14] It should be noted that unlike kmeans, DBSCAN makes the decision to *not* cluster a significant number of items.

[15] The term *does* appear in 30 of the unclustered items.

[16] In the .txt files available in GitHub, each item in the output of a DBSCAN cluster is labeled with 'GPT' denoting non-California DPS origin or CAGPT denoting California DPS origin. In the text of this paper we will refer to these as non-California and California for clarity.

of California receiving more information regarding how their personal information is collected, who their personal information is being provide to, how long their personal information of retained, and what their rights are regarding their personal information.

6 DISCUSSION

The heuristic analysis and NLP techniques employed in this study have yielded critical insights into the comparative richness of CCPA-compliant versus non-CCPA Data Privacy Statements. Utilizing OpenAI's GPT-4 has underlined the transformative capabilities of transformers in legal text interpretation, providing a quantitative measure of the CCPA's influence on DPS informativeness.

Our approach combined the precision of manual review with the breadth of automated analysis, allowing for a comprehensive parsing of DPS content. We have reflected on the strengths and limitations of these methodologies, highlighting their congruence with the overarching objectives of enhancing transparency and consumer awareness as mandated by the CCPA.

Future research avenues are broad and crucial, particularly regarding how diverse state legislation can impact DPS informativeness. This can reveal multifaceted data privacy standards across jurisdictions and the consequent variation in consumer rights and corporate disclosure practices. Moreover, exploring the synergistic use of NLP techniques beyond BERT and GPT-4 promises to refine our content analysis approach, aiming for nuanced interpretations of privacy policies in adherence to an increasingly complex legal framework.

Through such future explorations, we aim to broaden the corpus of knowledge in the domain of computational legal analysis, thereby enabling stakeholders to navigate the complex matrix of data privacy with greater efficacy and legal foresight.

7 LIMITATIONS

Over the course of this work, we encountered several limitations in the both the collection of data and the analysis of that data.

7.1 Scraping Limitations

While the BeautifulSoup library is a well-loved and useful tool, it is a bit of a blunt instrument when it comes to scraping text from websites. We overcame this shortcoming with a manual look at the deep URLs and text of the DPS identified using BeautifulSoup. Future searchers may ease their burden using Selenium or some other, sharper tool. Overall, out of approximately 10,000 websites our location of only 230 pristine pairs of DPS on the same site where one is for residents of California and the other if for everyone else may be limiting, but we maintain our sample size was sufficient.

7.2 Analysis Limitations

The nuances and possible combination of natural language processing and transformer tools delivers a daunting task for data scientists. A not insignificant amount of work was put into our BERT embeddings/kmeans analysis, only to produce results that were quickly dismissed. And while we are confident in our GPT embeddings/DBSCAN analysis there is surely a better pathway - perhaps one not yet developed but arriving soon - to take when investigating our hypothesis.

7.3 Direct Comparison Limitation

As mentioned in the our Results section above, while the heuristic analysis did so, our clustering analysis did not directly pit a non-California DPS against a California DPS from the exact same website. We are confident our clustering analysis is meaningful, but perhaps an algorithm other than clustering could be used to more directly compare DPS from the same website using sophisticated natural language processing.

8 CONCLUSION

This project has demonstrated the utility of integrating advanced computational techniques, specifically NLP and machine learning algorithms, in the analysis of Data Privacy Statements (DPS) under the scope of the California Consumer Privacy Act (CCPA). Through the innovative use of web scraping methodologies, manual data curation, and the application of OpenAI's GPT-4, this study has provided a foundational analysis of the comparative informativeness of CCPA-compliant versus non-compliant DPS.

Key findings indicate that CCPA-compliant DPS are significantly more detailed, offering greater transparency and potentially enhancing consumer awareness of their privacy rights. The application of transformers and clustering algorithms, such as DBSCAN, has allowed for a nuanced interpretation of legal texts, presenting new opportunities for legal analysts to understand and predict compliance trends.

The challenges encountered, particularly in the accurate extraction of DPS content due to the limitations of current web scraping technologies, underscore the need for ongoing advancements in computational tools to better handle the complexities of legal document analysis. The success of manual review highlights the indispensable role of human oversight in the validation of automated processes, ensuring data accuracy and reliability.

9 FUTURE WORK

Future work in this area of research could aim to expand these methodologies and findings to include data privacy statements from other jurisdictions besides California to compare the impacts of different legislative frameworks on data privacy norms. For example, one could compare the additional protections granted by the General Data Protection Regulation (GDPR) in the European Union to those additional protections granted by the CCPA. Even within the United States, there are fourteen states besides California which have passed data privacy legislation of varying strengths at the state level.[17] All of these pieces of legislation could be compared along the axes of the additional protections they claim to grant and how effectively they have been enforced.

Another facet of this work upon which future work could expand is the analysis of the clusters which were generated using DBSCAN. Further analysis could reveal which clusters are the most useful in distinguishing between different types of privacy statements, and what kinds of information are the most likely to be included in one type and not the other. Another type of analysis that could be performed would be analysis in search of data about the relative quality of the information and protections in each of the privacy

[17] Colorado, Connecticut, Delaware, Indiana, Iowa, Kentucky, Montana, New Hampshire, New Jersey, Oregon, Tennessee; Texas, Utah, and Virginia.

statements. While our findings focus on the quantity of information provided in each, it is possible that even if a Californian DPS is shorter, it might provide stronger protections in some cases.

An additional direction in which this work could expand would be towards the exploration of a wider array of NLP techniques beyond BERT and GPT-4. Some other techniques which could be applied in this field include other types of word embeddings, such as GloVe or FastText, and sequence models such as Long Short-Term Memory (LSTM) and Gated Recurrent Unit (GRU). Sequence models in particular have shown to be effective for text classification due to their ability to learn long-term dependencies in data. An expansion of techniques could be utilized to further refine data analysis methodologies, facilitating more sophisticated insights into privacy policy structures and compliance.

10 APPENDIX

10.1 Data Privacy Statement (DPS)

A legal document provided by a website that explains how the company collects, uses, stores, and protects users' personal data. It is also known as a privacy policy or privacy notice.

10.2 California Consumer Privacy Act (CCPA)

A state statute intended to enhance privacy rights and consumer protection for residents of California, USA. It requires businesses to inform consumers about the types of personal data they collect and to provide them with certain rights regarding their personal information. [10]

10.3 Uniform Resource Locator(URL)

The address of a web page on the internet, which typically consists of a protocol (http or https), a domain name, and optionally, a path and parameters specifying a resource within the domain.

10.3.1 Base URLs. Base URLs contain only the domain name and extension, allowing for navigation to a website's home page.

10.3.2 Deep URLs. Deep URLs contain the path portion of a URL relative to the current page's base URL, allowing for the direct navigation to a DPS presented by a website.

10.4 DBSCAN

DBSCAN (Density-Based Spatial Clustering of Applications with Noise)[11] is an intuitive and widely used clustering algorithm that excels in identifying clusters of varying shapes and sizes in a data set, which is particularly useful in the analysis of complex datasets like Data Privacy Statements (DPS). The algorithm operates based on two primary parameters:

10.4.1 eps (epsilon). This parameter sets the maximum distance between two points that can be considered neighbors. It effectively defines the radius of the neighborhood around a point.

10.4.2 min samples. This parameter specifies the minimum number of points required to form a dense region, which DBSCAN treats as a single cluster. Points in low-density areas that do not meet these criteria are typically marked as noise or outliers.

REFERENCES

[1] Mike Hintze, "In Defense of Long Privacy Statement," 2017: 76 Md. L. Rev. 1044 (2016-2017)

[2] Kambiz Ghazinour, Maryam Majedi, and Ken Barker. 2009. A model for privacy policy visualization. 2009 33rd Annual IEEE International Computer Software and Applications Conference (2009), 335-340. DOI:http://dx.doi.org/10.1109/compsac.2009.156

[3] Norman Sadeh, Alessandro Acquisti, Travis Breaux, and Lorrie Faith Cranor. 2013. The Usable Privacy Policy Project: Combining crowdsourcing ... (December 2013). Retrieved February 26, 2024 from http://reports-archive.adm.cs.cmu.edu/anon/isr2013/CMU-ISR-13-119.pdf

[4] Hamza Harkous, Kassem Fawaz, Rémi Lebret, Florian Schaub, Kang G. Shin, and Karl Aberer. 2018. Polisis: Automated Analysis and presentation of privacy policies using Deep Learning. (August 2018). Retrieved February 26, 2024 from https://www.usenix.org/conference/usenixsecurity18/presentation/harkous

[5] Shomir Wilson *et al.* 2018. Analyzing Privacy Policies at Scale: From Crowdsourcing to Automated Annotations. ACM Transactions on the Web 13, 1 (December 2018), 1–29. DOI:http://dx.doi.org/10.1145/3230665

[6] Abhilasha Ravichander, Alan W. Black, Shomir Wilson, Thomas Norton, and Norman Sadeh. 2019. Question answering for privacy policies: Combining Computational and Legal Perspectives. Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP) (November 2019), 4947-4958. DOI:http://dx.doi.org/10.18653/v1/d19-1500

[7] Yan Shvartzshanider, Ananth Balashankar, Thomas Wies, and Lakshminarayanan Subramanian. 2020. Beyond the text: Analysis of privacy statements through syntactic and semantic role labeling. Proceedings of the Natural Legal Language Processing Workshop 2023 (October 2020). DOI:http://dx.doi.org/10.18653/v1/2023.nllp-1.10

[8] Muhammad Sajidur Rahman *et al.* 2022. PermPress: Machine learning-based pipeline to evaluate permissions in App Privacy Policies. IEEE Access 10 (August 2022), 89248–89269. DOI:http://dx.doi.org/10.1109/access.2022.3199882

[9] Isabel Wagner. 2023. Privacy Policies across the Ages: Content of Privacy Policies 1996–2021. ACM Transactions on Privacy and Security 26, 3 (May 2023), 1–32. DOI:http://dx.doi.org/10.1145/3590152

[10] Cal. Civ. Code § 1798.100 *et seq.* and Cal. Code Regs. tit. 11, § 7000 *et seq.* [11] Ester, M., Kriegel, H.-P., Sander, J., & Xu, X. (1996). A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise. In Proceedings of the 2nd International Conference on Knowledge Discovery and Data Mining (KDD-96). Portland, OR: AAAI Press. pp. 226–231.