Personal Data Analysis with Large Language Models

Matias Badino

CMPINF 1205

Professor Song Shi

April 25, 2024

Code Repository: https://github.com/N0taR0b0t/DataAnalysisLLM

Introduction:

This research project was designed to undertake a comprehensive examination and comparison of personal data and privacy policies implemented by major digital platforms, including Instagram, Twitter, Google, and Amazon. The objective was to determine the types of data collected by these platforms and to assess the privacy frameworks in the United States, Canada, and Germany. The study leveraged personal data downloaded from these platforms, supported by advanced language models, to conduct a detailed categorization of the data types involved.

Objectives

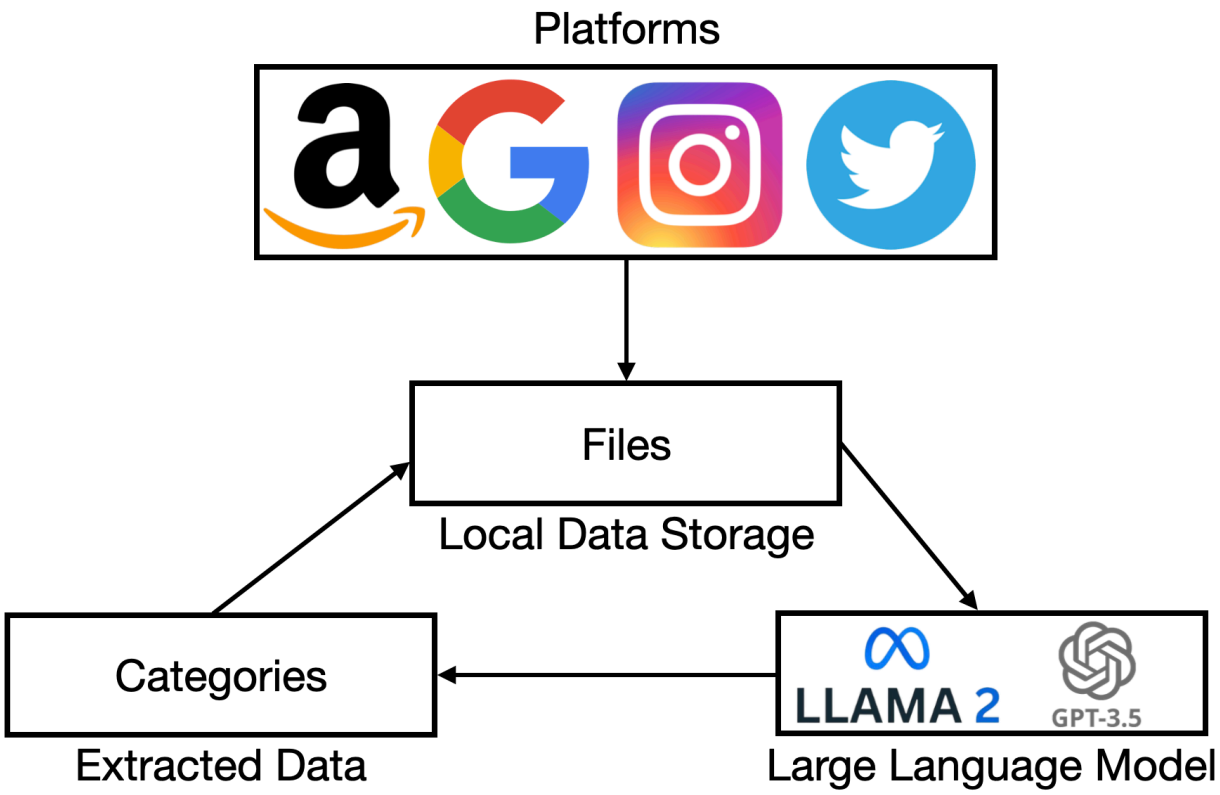
The primary objectives were to delineate and compare the various types of personal data collected by the aforementioned platforms. To achieve a comprehensive understanding of how the data was handled, the project sought to answer the following critical questions:
How do privacy rights of users differ across the United States, Canada, and Germany? What policy obstacles hinder users' abilities to manage their digital privacy on these platforms within these countries? What types of personal data are platforms permitted to collect, and for what specific purposes can this data be utilized? What mechanisms are available to users for accessing

or deleting their personal information held by these digital entities, and how do these mechanisms vary by country?

To support these objectives, a Python program was developed to generate a comprehensive list of categories, offering a detailed representation of the various types of data collected by these platforms from users.

Methodology

The analysis was facilitated through Python scripting and large language models. Initially, the local model, identified as Llama-2-13b-chat.Q5, was planned for primary use; however, OpenAI's GPT-3.5 was predominantly utilized as it processes text approximately 100 times faster (2000 characters per second vs 20 characters per second) and is far less likely to deviate from the instructions. The first version of the code provided the large language model with file



(Figure 1) A simple visualization of the flow of information in the python code

contents along with categories such as Personal Files and Information, Platform and Content Engagement, Advertisements and Products, Settings and Internal References, and Miscellaneous. A dynamic approach was then adopted, where the large language model was provided with a list of categories that it could modify and reuse in future iterations (visualized in Figure 1). This approach allowed the large language model the autonomy to curate categories and assign those labels to files. Due to this newfound autonomy, multiple revisions of the large language model's instructions were necessary to fine-tune the output. The process involved feeding the large language model one file at a time, extracting the categories, and then including the revised category list along with the next file in the following iteration. Occasionally, an excessive number of categories was produced, requiring adjustments in the instructions to mitigate the issue. Implementing dynamic text extraction—which checks for multiple patterns expected in the output—proved crucial for maintaining data integrity.

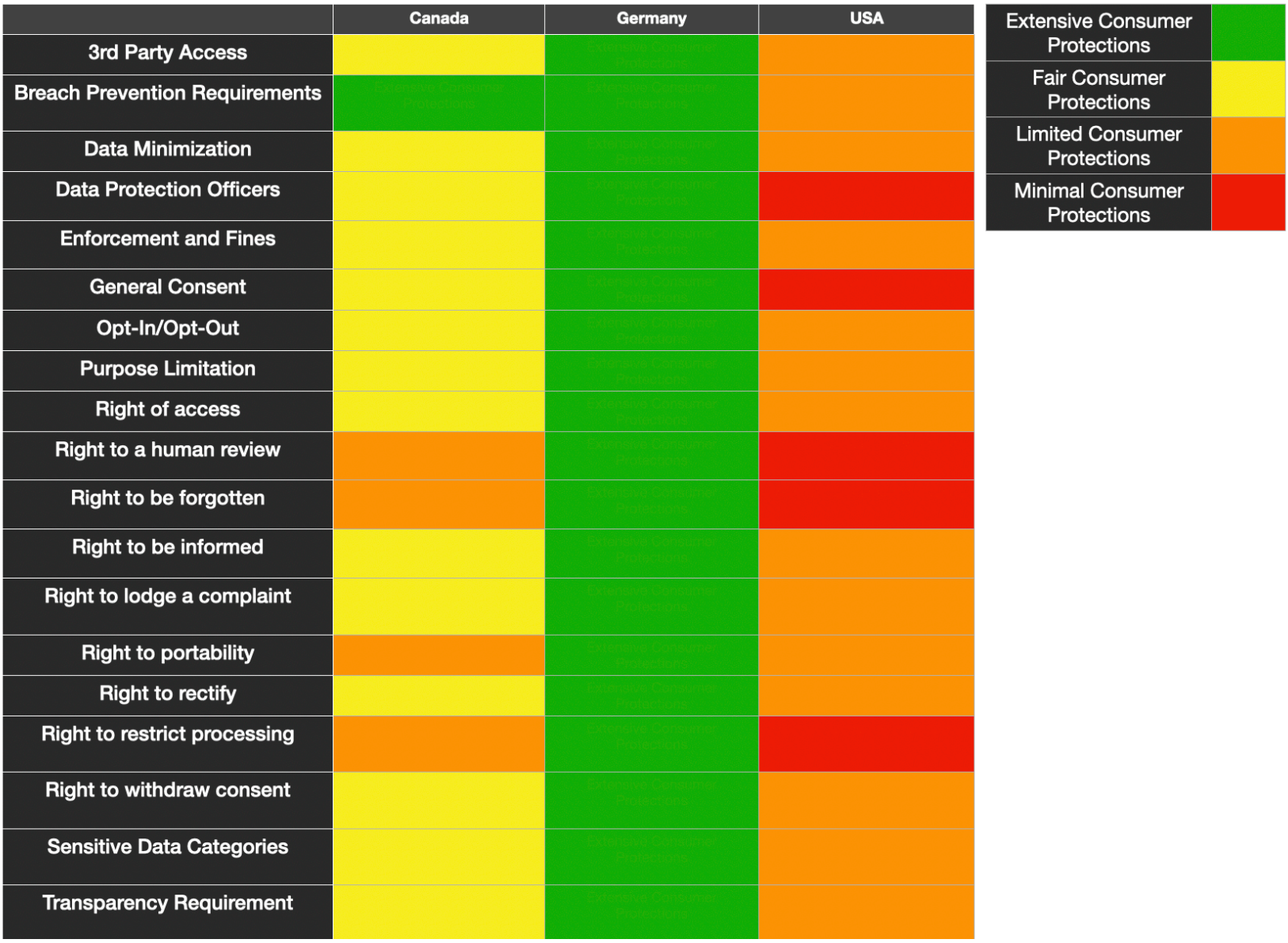
Results

The final result from the code is a list of 93 categories generated by OpenAI's GPT-3.5. Figure 2 on the following page contains the unedited category list, including the redundant categories. Although there are a handful of redundant categories, such as "Participant" and "Time" (we already have "Participants" and "Timestamp"), this is a significant reduction from the approximately 2000 categories generated by the original version of the code which did not reuse the dynamic category list. In addition to automated coding, a manual review of the privacy laws for Canada, Germany, and the United States was conducted. Figure 3 shows a comparison of the privacy protection laws provided by these countries on the national level. Canada's PIPEDA law

1. Academic Information
2. Account Holder Name
3. Action
4. Activity Status
5. Address
6. Ads viewed
7. Advertising Advertiser Audiences
8. App Last Updated Time
9. Audiences in which you are included
10. Author
11. Beneficiary
12. Billing Information
13. Bio
14. Camera Information
15. Camera Metadata
16. Channel Auto Moderation in Live Chat
17. Channel Id
18. Click Data
19. Clicked Items
20. Comment
21. Communication Preferences
22. Company Name
23. Contact Information
24. Country Code
25. Date
26. Deletion Data Type
27. Deletion Status
28. Demographics
29. Device ID
30. Device Information
31. Device Name
32. Device Operating System
33. Device Platform
34. Display Text
35. Email
36. Emoji Interaction Data
37. Enablement Status
38. Encoding of the Audio
39. Energy Usage Data
40. Entity App Names
41. Event
42. Experience Type
43. Feedback Type
44. First Name
45. Followers
46. Following
47. Geolocation
48. IP Address
49. Interactions with Advertiser
50. Item Name
51. Legacy Payment Information
52. Liked Items
53. Liked Threads
54. Likes
55. Live Chat ID
56. Location Data
57. Mac Address
58. Media Owner
59. Merchant Name
60. Message Content
61. Message ID
62. Notification Preferences
63. Order Status
64. Participant
65. Participants
66. Payment Information
67. Phone Number
68. Playlist Video Creation Timestamp
69. Playlist Visibility
70. Product Name
71. Profile Photos
72. Question and Answer
73. Recent Searches
74. Recently Deleted Content
75. Search Contributions
76. Sensor Data
77. Sentiment Score
78. Skill Name
79. Smart Home Devices
80. Song Title
81. Song URL
82. Subscription Data
83. Threads and Replies
84. Time
85. Timestamp
86. Transaction Details
87. Transcription
88. URL
89. User ID
90. Username
91. Video ID
92. Watch Event Data
93. Word or Phrase Searches

(Figure 2) The 93 categories extracted by the most recent version of the code

is relatively comprehensive compared to the USA's scarce consumer protections. Germany's privacy laws are comprehensive across the board and seem to be keeping up with the changes in the sale, collection, processing, transfer, and storage of consumer data.



(Figure 3) A visualization of the privacy protections afforded by each country on the national level

Conclusion and Future Work

Further improvements to this project will involve utilizing multiple sets of instructions for the large language model. The instructions are simply a series of sentences in english that explain the context and output requirements to the large language model. In the next version of the code, the large language model will generate a list of categories independently of the existing list. Then, it will review these new categories and compare them against the existing dynamic list, where it will be instructed to merge or create new categories. By separating the process and having the category identification and category merging processes occur individually, the large language model can more effectively concentrate its attention on each of these tasks. This change is expected to reduce the number of redundant categories further and, along with more precise prompt tuning, will enhance the quality of the category names. Additionally, updates to the code will facilitate easier adoption and utilization by the open-source community, turning this program into a more accessible tool.

Bibliography

Banks, T. M. (2020, May 6). GDPR matchup: Canada's Personal Information Protection and Electronic Documents act. GDPR matchup: Canada's Personal Information Protection and Electronic Documents Act. https://iapp.org/news/a/matchup-canadas-pipeda-and-the-gdpr/

German Federal Ministry of Justice and the Federal Office of Justice. (n.d.). BDSG - Englisch. https://www.gesetze-im-internet.de/englisch_bdsg/englisch_bdsg.pdf

The European Parliament and of the Council of 27 April 2016. (2016, May 4). *Regulation - 2016/679 - en - GDPR - EUR-lex*. EUR-Lex. https://eur-lex.europa.eu/eli/reg/2016/679/oj

Meta. (n.d.). *Llama 2: Open Foundation and fine-tuned chat models*. Paper page. https://huggingface.co/papers/2307.09288

Meta. (n.d.). *Meta-llama/LLAMA-2-13B-CHAT-HF · hugging face*. meta-llama/Llama-2-13b-chat-hf · Hugging Face. https://huggingface.co/meta-llama/Llama-2-13b-chat-hf

(OCR), O. for C. R. (2022, January 19). Your rights under HIPAA. HHS.gov. https://www.hhs.gov/hipaa/for-individuals/guidance-materials-for-consumers/index.html

OneTrust DataGuidance. (n.d.). Comparing privacy laws: GDPR v. pipeda. https://www.dataguidance.com/sites/default/files/gdpr_v_pipeda.pdf