Final Project:
Increasing the Readability of Privacy Policies

Special Topics: Text Analysis
(36-468/668)

Danny Nguyen

Fall, 2023

1 Introduction

In our modern age of technology, data has become the hot new commodity exchanged between users and platforms. Ubiquitously, all websites today have a privacy policy that they notify users with. The idea of effective notice has been a crucial part of many different Fair Information Practices and Principles, which are doctrines about informational privacy. However, privacy policies are notorious for being hard to understand due to both their length and legal jargon. In a 2004 study, it was found that only 0.24% of 5,158 users navigate towards the privacy policy of a website, and only 6% of policies are readable by the most vulnerable 28.3% of the population [1]. Our goal is to determine what language choices are used to make a privacy policy easier or harder to understand. Specifically, what are the linguistic features of difficult privacy policies as opposed to easier privacy policies? Using these findings, we hope to craft better methods of writing privacy policies for all users.

2 Data

The data used in this project comes from the Princeton-Leuven Longitudinal Corpus of Privacy Policies. This corpus is found here.

The Princeton-Leuven Longitudinal Corpus of Privacy Policies was created by Princeton professors Amos, Lucherini, Kshirsagar, Mayer, Narayanan, and Radboud professor Acar. This corpus was created for their research paper on the various changes to privacy policies over time [2]. In total, it consists of over 1 million privacy policy snapshots from more than 100,000 websites, spanning over two decades. For this project, we limit our focus to the policy texts and the Flesch Ease Score. The Flesch Ease Score is a categorical labeling using the Flesch-Kincaid score, a numerical score of readability. The specific main corpus and subcorpra that we use are shown in Table 1.

Table 1: Number of Privacy Policies and Words in Corpus and Sub-corpora.

Readability	Number	Number %	Words	Word %
easy	1	0.00	2441	0.00
fairly_easy	24	0.01	217096	0.01
standard	985	0.44	8562270	0.31
fairly_difficult	16265	7.24	137117875	4.99
difficult	186266	82.93	2297743589	83.69
very_confusing	21069	9.38	302056621	11.00
Total	224610	100.00	2745699892	100.00

We see that our full corpus has 224,610 privacy policies after filtering out policies that did not have a Flesch Ease Score. From both the percentages of policies and words, we observed that our corpus is extremely unbalanced with an overwhelming number of difficult privacy policies compared to the other categories. This can be an issue as we are analyzing the privacy policies concerning their readability category, so we decided to downsample for a balanced corpus.

3 Methods

First, we needed to create a balanced sample to analyze. We observed that privacy policies overwhelmingly fell into the harder reading comprehension categories, so our data is unbalanced. Specifically, 1,010 privacy policies fell under the categories of **easy**, **fairly easy**, and **standard** whereas there were 223,600 policies under **fairly difficult**, **difficult**, and **very confusing**. To create a balanced sample, we sampled 1,009 privacy policies that had a Flesch Ease score of **fairly easy** and **standard** to represent the easier policies. We did not include **easy** because there was only 1 policy as an outlier. These categories were chosen to represent easier as there are 6 total categories, and these were the first 3 (though, we did not use **easy**). Then, we randomly sampled 1,009 privacy policies that had a Flesch Ease score of **fairly difficult**, **difficult**, and **very confusing** to present the harder policies. This sample size was chosen to match the total number of easier texts, and it should be representative of the harder texts as a random sample. Similarly, these categories were chosen for harder texts as the last 3 categories on the reading difficulty scale. Finally, we combined these to create the final balanced corpus of 2,018 privacy policies. We acknowledge the concern that this final corpus is only a fraction of our original corpus, less than 1% specifically. However, this corpus is balanced and contains fair representation for both the easier and harder privacy policies.

Next, we performed factor analysis using Biber's tagging method. The specifics of how this tagging system includes aspects of language such as personal pronouns and adverbs [3]. We determined the number of factors to use based on the scree plot's acceleration factor, which was 2 for our corpus. For this paper, our analysis was determining what specific linguistic elements are used in easier policies vs harder ones rather than determining a broad dimension to categorize the texts. Aside from practicality, these texts come from the same genre and serve the same purpose, so interpreting the dimension is not the most useful part of our analysis.

Next, we developed a model for predicting whether a privacy policy was **EASIER** or **HARDER** to read. Once again, the privacy policies categorized as **fairly easy** and **standard** were labeled **EASIER**, and **fairly difficult**, **difficult**, and **very confusing** were labeled as **HARDER**. We used lasso regression based on Mosteller and Wallace's methods for determining the authorship of the Federalist Papers [4]. While Mosteller and Wallace needed to identify potentially productive tokens to use in their model, we utilized a keyness table with **EASIER** as the reference corpus and **HARDER** as the target corpus. This allowed us to effectively determine the tokens that appear more frequently in one category over the other. We separated our subcorpa into a training set of 1,600 randomly selected policies and the other 400 was our test set. After creating our model, we focused on analyzing the tokens used in the model rather than the classification accuracy of the model itself. We still tested our model's accuracy to see if this model reasonably identifies whether a policy is **EASIER** or **HARDER**, but this was not the focus.

4 Results

First, we analyzed the linguistic features of easier vs harder privacy policies using factor analysis using Biber's linguistic tags.

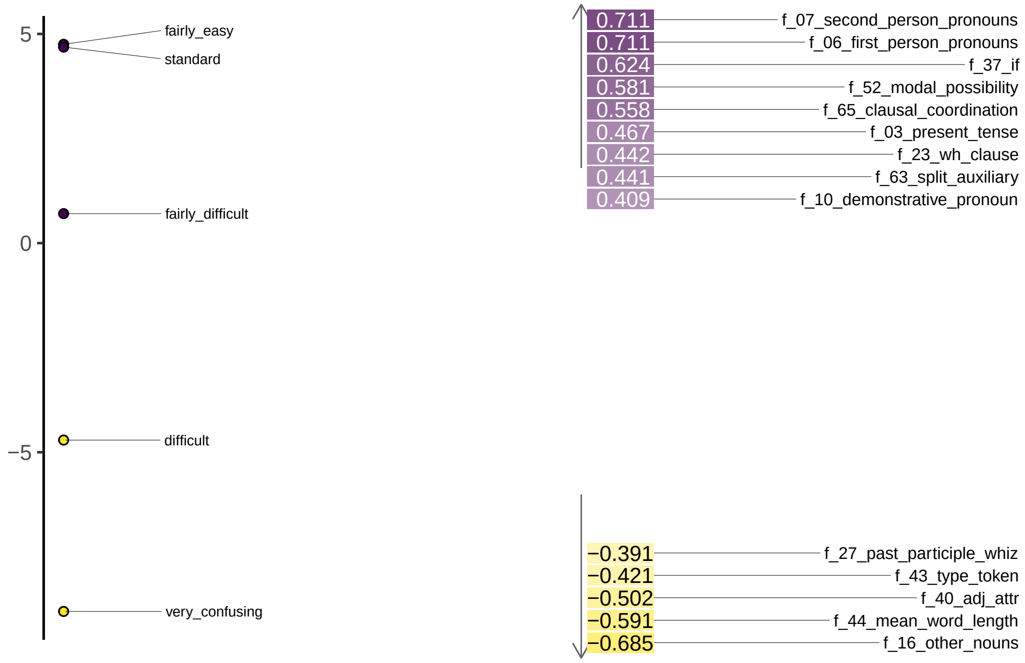

Figure 1: Dimension score by Readability plotted along Factor 2 using Biber.

In Figure 1, we observe the most notable linguistic features of the easier privacy policies were the presence of **second-person** and **first-person** pronouns. The most notable linguistic features of the harder privacy policies were **mean word length** and **other nouns**. This suggests the idea that easier privacy policies use a more interactive communication style that more directly defines the roles of the website and user.

Next, we created a lasso regression model to categorize whether a privacy policy is considered **EASIER** or **HARDER**.

Table 2: Tokens with the highest keyness values in the easier policies when compared to the harder policies.

Token	LL	LR	AF_Tar	AF_Ref
we	2737.71	0.56	39282	34612
i	2320.74	2.52	3092	696
you	2270.12	0.44	49276	46932
might	1969.74	2.64	2494	519
get	1693.70	3.38	1657	206

We first determined a list of candidate tokens to use in our lasso model by creating a keyness table shown in Table 2. We used all of our **EASIER** policies in our sample as the reference corpus and all of our **HARDER** policies in the sample as the target corpus. We made sure that the tokens chosen had appeared within BOTH corpora. We noted that first-person pronouns **we** and **i** alongside second-person pronoun **you** appear in our top 5 tokens with the highest keyness.

Table 3: Tokens Used in Model for Easier Policies vs Harder Policies

Token	Coeff	Token	Coeff
we	-30.20	lawyers	354.92
you	-27.91	analytical	107.40
my	-33.73	corporations	172.94
click	-51.23	lawfulness	539.94
me	-57.48	european	223.51
tools	-41.31	regulation	305.48
write	-45.57	liability	33.41
inquires	-3.06	functionality	327.57

Afterward, we created our lasso regression model and extracted the most interesting tokens shown in Table 3. A negative coefficient means that this token is associated with the **EASIER** policies, and a positive coefficient means it is associated with the **HARDER** policies. We observe once again that first-person pronouns like **we**, **my**, **me**, and second-person pronoun **you** appear to be associated with **EASIER** texts. Additionally, some of the other tokens associated with **EASIER** texts include simple but direct actions that users can take such as **click**, **tools**, **write**, and **inquires**. For the **HARDER** texts, it seems that many of the tokens are associated with many legal or business aspects of the website. Tokens such as **lawful**, **liability**, and **regulation** suggest that websites prioritize the function of their privacy policies as a legal defense. Many websites may contain the token **european** due to the EU's General Data Protection Regulation, a landmark privacy law. If the goal of websites is simply to legally cover their bases, then it detracts from helping users understand what happens. Overall, users seem to understand privacy policies better whenever the policies are more direct with less specialized details.

Table 4: First 10 Lasso Classifications of Policy Readability.

Predicted	Prob	True Label
HARDER	0.89	HARDER
HARDER	0.82	HARDER
HARDER	0.78	HARDER
EASIER	0.08	EASIER
HARDER	0.97	HARDER
HARDER	0.93	EASIER
HARDER	0.95	HARDER
HARDER	0.96	HARDER
EASIER	0.05	EASIER
HARDER	0.83	HARDER

In Table 4, we observe that our lasso model performed reasonably well on the first 10 policies with only 1 error. We tested it on the rest of our test set and found it correctly classified 368 policies out of 400. In other words, our model had an accuracy of 92%. So, it appears that our model performs relatively well for classification, which indicates that the tokens we identified contribute to the ease or difficulty of privacy policies.

5 Discussion

From our analysis, we have seen that harder-to-read privacy policies make up an overwhelming majority of privacy policies. This makes sense in the context of privacy policies as these texts are very dense and contain many legal or specialized terms. This notion is further supported by the notable linguistic features of harder policies being **mean word length** and **other nouns** under the Biber factor analysis. On the other hand, the Biber factor analysis identified the notable linguistic features of easier policies to be **second-person** and **first-person pronouns**. In the context of privacy policies, this suggests that these policies try to directly define the relationship and boundaries between the website and the users. Our lasso regression model further supports this idea with some of the tokens it chose.

From these insights, we can recommend that privacy policies be written in more active and direct ways. One example is having two sections of the privacy policy dedicated to “We” statements about how the website handles data and “You” statements about actions and controls users have. It is important to keep the language relatively simple and direct so that users understand how to do something. Another idea is to implement the privacy policy Q and A style with common questions.

The major limitations of this study are related to sample size. Specifically, we did not have as many easier-to-read privacy policies, which limits what possible linguistic features showed up in our analysis. Another crucial limitation was the sample sizes themselves as we had to limit the main corpus for computation purposes. It is possible that another analysis that utilizes the full corpus would produce different results. Additionally, our study is limited by the time frame that the data was captured. Some of the latest privacy policies in the corpus arrived shortly after the introduction of privacy laws such as the EU’s GDPR; however, future privacy policies may look very different than in the analysis.

Overall, this analysis showed promise that there may be distinct differences in privacy policies based on reading comprehension. Some of the potential directions that this research could take involve more active testing if the recommended strategies make it easier for people to comprehend privacy policies. Another potential direction is analyzing the writing goals of the authors of privacy policies. Privacy policies should be written for the users to understand, first and foremost. This notion of effective notice is what many past privacy scholars believe is a crucial component of preserving privacy.

References

[1] C. Jensen and C. Potts, "Privacy policies as decision-making tools: An evaluation of online privacy notices," in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, ser. CHI '04, Vienna, Austria: Association for Computing Machinery, 2004, pp. 471–478, ISBN: 1581137028. DOI: 10.1145/985692.985752. [Online]. Available: https://doi.org/10.1145/985692.985752.

[2] R. Amos, G. Acar, E. Lucherini, M. Kshirsagar, A. Narayanan, and J. Mayer, "Privacy Policies over Time: Curation and Analysis of a Million-Document Dataset," in *Proceedings of The Web Conference 2021*, ser. WWW '21, Ljubljana, Slovenia: Association for Computing Machinery, Apr. 19, 2021, p. 22. DOI: 10.1145/3442381.3450048. [Online]. Available: https://doi.org/10.1145/3442381.3450048.

[3] D. Biber, S. Conrad, R. Reppen, P. Byrd, and M. Helt, "Speaking and writing in the university: A multidimensional comparison," *TESOL Quarterly*, vol. 36, no. 1, pp. 9–48, 2002, ISSN: 00398322. [Online]. Available: http://www.jstor.org/stable/3588359 (visited on 10/13/2023).

[4] F. Mosteller and D. L. Wallace, "Inference in an authorship problem," *Journal of the American Statistical Association*, vol. 58, no. 302, pp. 275–309, 1963, ISSN: 01621459. [Online]. Available: http://www.jstor.org/stable/2283270 (visited on 12/04/2023).