

# PrivaSEE : Statement of Work

Gloria Umutoni ( [gumutoni@g.harvard.edu](mailto:gumutoni@g.harvard.edu) ), Shira Aronson ( [saronson@hbs.edu](mailto:saronson@hbs.edu) ), Aditi Raju ( [araju@college.harvard.edu](mailto:araju@college.harvard.edu) ), Sammi Zhu ( [sammizhu@college.harvard.edu](mailto:sammizhu@college.harvard.edu) ), Yeabsira Mohammed ( [yeabsiramohammed@college.harvard.edu](mailto:yeabsiramohammed@college.harvard.edu) )

## Background and Motivation

In today's digital age, it is hard to find an internet user who has not signed a terms of service document, and harder still to find one who has read it. In 2008, Aleecia M. McDonald and Lorrie Faith Cranor, two Carnegie Mellon professors, calculated that a reasonable reading of all the privacy policies that one encounters in a year would require 76 full workdays at a national opportunity cost of \$781 billion.<sup>1</sup> More recently, a 2017 Deloitte survey found that over 90% of consumers accept terms and conditions without reading them. The number is surely only increasing as consumers navigate an increasingly complex digital landscape.<sup>2</sup> These “contracts of adhesion” or “click-wrap contracts” grant rights without negotiation, establishing practices before users and regulators even notice.<sup>3</sup>

## Problem Statement

When deciding on a messaging app, for example, the average consumer is unlikely to read or understand the terms and conditions of multiple apps and decide which one to use accordingly. This project aims to bridge consumers' knowledge gaps around their data privacy by building an app that reviews terms and conditions agreements and informs users about the aspects of the privacy they cede by using a certain app or website. PrivaSEE would allow users to understand the implications to their data privacy, and compare options in a way that aligns with their personal privacy priorities.

## Data Sources

The project involves collecting and processing the text of legal user agreements for a variety of apps and websites. The non-profit project ToS;DR (“Terms of Service; Didn’t Read”) provides such a dataset, which includes the user agreements for over 10,000 sites and apps.<sup>4</sup> Many agreements are further labeled with a letter grade indicating the level of users' privacy. Some user agreements additionally include annotations which are largely crowdsourced, and reviewed by subject matter experts with legal expertise. For this project, we will utilize the data in two distinct ways to support the app's primary functions:

1. For apps where the full user agreement is available, PrivaSEE will be able to reference the text in its memory without users having to upload the terms themselves. The data for this aspect of the app is the raw full text of user agreements as collected by ToS;DR, with accompanying letter grades.

---

<sup>1</sup> Aleecia M. McDonald and Lorrie Faith Cranor, “The Cost of Reading Privacy Policies,” *Journal of Policy for the Information Society*, 4, no. 3 (2008), <http://hdl.handle.net/1811/72839>.

<sup>2</sup> Deloitte. *The Consumer Data Privacy Paradox: How to Balance Privacy and Personalization*. Deloitte Insights, 2020, <https://www2.deloitte.com/us/en/insights/industry/technology/consumer-data-privacy-paradox.html>.

<sup>3</sup> Kim, Nancy S. *Wrap Contracts: Foundations and Ramifications*. Oxford University Press, 2013.

<sup>4</sup> <https://edit.tosdr.org/services>

2. For apps where the user agreements have been annotated, these annotations will be used to further train the model to break down legal statements into clear takeaways. This data will support the PrivaSEE model's ability to break down legal agreements that were not included in the training set.

The ToS;DR dataset is well-suited for training a robust model capable of tackling agreements across a variety of services. We will identify any broad categories of user agreements missing from the data, and spot-check to verify the accuracy of annotations, supplementing with other data as needed (Eg. Common Crawl, LexPredict).

## **Objectives**

Our goal is to create a web app that achieves the following functionality. An initial design can be referenced [here](#).

1. Upload a Terms and Conditions document for processing into simplified, easily understood language, highlighting key data privacy considerations
2. Provide a privacy grade based on the factors the user values most (e.g., prioritizing protection against data leaks versus monitoring communications)
3. Search by privacy concerns (i.e. consumers search for a specific concern and see which companies are doing well or worse on a certain category)
4. Recommendations: a user puts in their requirements, and we recommend an app based on their priorities.

## **Scope**

The project aims to develop an NLP model that translates complex legal jargon into simple language and creates a user-friendly web application for users to upload and analyze terms and conditions. The application will provide a customized privacy grade and a search/recommendation feature based on user privacy preferences. The learning emphasis includes fine-tuning an LLM for legal text processing, using RAG to support classification and recommendations, and deploying a scalable web app. Research efforts focus on exploring existing literature on NLP in legal analysis, reviewing case studies like ToS;DR for public engagement insights, and examining privacy laws for compliance. This information will be used for retrieval of relevant legal references and standards, which are then used to generate summaries and grades. The project also incorporates an interactive UI as a fun factor to simplify legal texts. However, there are limitations, such as the potential inaccuracy of crowd-sourced annotations, which will be addressed through a verification system. Initially, the data collection will focus on popular consumer apps, with potential expansion as the project grows. Lastly, technical constraints, including high computational demands due to complex NLP models and the amount of text data, will need to be managed.

## **High-Level Project Stages**

The project will begin with data collection from ToS;DR, followed by data cleaning and exploratory analysis. Initial models like BERT and GPT will be trained on a data subset. A prototype with user interfaces and NLP integration will be developed, followed by iterative model refinement, adding features like privacy grading. A beta version will then be deployed for user feedback and testing before the final release.



## Problem Statement

In today's digital age, it is hard to find an internet user who has not signed a user terms of service document, and harder still to find one who has read it. In 2008, Aleecia M. McDonald and Lorrie Faith Cranor, two Carnegie Mellon professors, calculated that a reasonable reading of all the privacy policies that one encounters in a year would require 76 full workdays at a national opportunity cost of \$781 billion.<sup>5</sup> More recently, a 2017 Deloitte survey found that over 90% of consumers accept terms and conditions without reading them. The number is surely only increasing as consumers navigate an increasingly complex digital landscape.<sup>6</sup> These user terms of privacy contracts, known as “contracts of adhesion” or “click-wrap contracts” result in acquisition of rights without bargaining and an establishment and entrenchment of practices before users, and regulators, realize what has happened.<sup>7</sup>

**When deciding on a messaging app, for example, the average consumer, no matter how concerned about their data privacy, is unlikely to read or understand the terms and conditions of multiple apps and decide which one to use accordingly. This project aims to bridge consumers' knowledge gaps around their data privacy by building an app that reviews terms and conditions agreements, and informs users about the aspects of their privacy they cede by using a certain app or website. PrivaSEE would allow users to understand the implications to their data privacy, and compare options in a way that aligns with their personal privacy priorities. We hope that by translating these obtuse contracts users will become more informed and advocate for the protection of their privacy both to the companies that produce these products and to the government.**

---

<sup>5</sup> Aleecia M. McDonald and Lorrie Faith Cranor, “The Cost of Reading Privacy Policies,” *Journal of Policy for the Information Society*, 4, no. 3 (2008), <http://hdl.handle.net/1811/72839>.

<sup>6</sup> Deloitte. *The Consumer Data Privacy Paradox: How to Balance Privacy and Personalization*. Deloitte Insights, 2020, <https://www2.deloitte.com/us/en/insights/industry/technology/consumer-data-privacy-paradox.html>.

<sup>7</sup> Kim, Nancy S. *Wrap Contracts: Foundations and Ramifications*. Oxford University Press, 2013.

**Objectives:** *List the primary goals or outcomes, which should align with your problem statement and the minimum components outlined above.*

## **Data**

The project involves collecting and processing the text of legal user agreements for a variety of apps and websites. A thorough dataset of user agreements across a variety of services is crucial to train the model to recall and interpret a range of user agreements. The non-profit project ToS;DR (“Terms of Service; Didn’t Read”) provides such a dataset, which includes the user agreements for over 10,000 sites and apps.<sup>8</sup> Many agreements are further labeled with a letter grade indicating the level of users’ privacy. Some user agreements additionally include annotations which are largely crowdsourced, and reviewed by subject matter experts with legal expertise.

For this project, we will utilize the data in two distinct ways to support the app’s primary functions:

3. For apps where the full user agreement is available, PrivaSEE users will be able to ask questions without manually providing the user agreement text. The data for this aspect of the app is the raw full text of user agreements as collected by ToS;DR, with accompanying letter grades.
4. For apps where the user agreements have been annotated, these annotations will be used to further train the model to break down legal statements into clear takeaways. This data will support the PrivaSEE model’s ability to break down the implications of legal agreements that were not included in the training set.

The ToS;DR dataset provides user agreements for a variety of websites and apps, indicating that the dataset is well-suited for training a robust model capable of tackling agreements across a variety of services. We will identify any broad categories of user agreements missing from the data, and spot-check to verify the accuracy of annotations, supplementing with other data as needed (Eg. Common Crawl, LexPredict).

Product: Web Application

---

<sup>8</sup> <https://edit.tosdr.org/services>

1. upload a Terms and Conditions document for processing into simplified, easily understood language, highlighting key data privacy considerations
2. provide a privacy grade based on the factors the user values most (e.g., prioritizing protection against data leaks versus monitoring communications)
3. search by privacy concerns (i.e. consumers search for a specific concern and see which companies are doing well or worse on a certain category)
4. Recommendations: a user puts in their requirements, and we recommend an app based on their priorities.

**Learning Emphasis:** *Opt for models and methods that your team understands. The project should reflect your grasp of course concepts.*

**Application Mock Design:** *Include a preliminary design or sketch for the application you intend to develop. This could range from simple wireframes to a more detailed, clickable prototype.*

**Research and Development:** *Reference papers, blog posts, or other scholarly materials that aid your project and align with your objectives.*

**Fun Factor:** *The project should also be a space for you to enjoy both the subject matter and the developmental process.*

- data visualization and aesthetics

**Limitations and Risks:** *Discuss any anticipated challenges or limitations, such as data quality issues or technical constraints.*

- Scope of data collection in terms of categories (focus more on one general type of apps over another or no?)
- Crowd-sourced nature of our annotations. Some could be wrong!
- 

**Milestones:** *List key milestones for both your project and application development. Include tentative deadlines, if possible.*

- 1.
- 2.

3.

The goal of the project is to empower consumers by enhancing their understanding of the privaSEE implications tied to the terms and conditions of various apps and websites they use. The application will interpret and simplify complex legal documents into concise, clear summaries with an emphasis on privacy concerns, making this essential information more accessible and actionable for the average user. The project utilizes a large and heterogeneous dataset, specifically datasets from ToS;DR, which has over 10,000 entries of user agreements. This dataset is crowd-sourced annotations that add depth and diversity to the data, crucial for training robust models capable of understanding and interpreting a wide range of legal documents. To accommodate the scale of data and user base, The application will be capable of handling multiple users simultaneously without performance degradation. This is achieved through the use of scalable cloud services and efficient data handling strategies, ensuring that the infrastructure can grow with user demand.

Moreover, the project employs advanced natural language processing (NLP) models, to tackle the challenge of interpreting and summarizing the dense and complex legal text found in terms and conditions documents. These models are crucial for extracting meaningful information from legal jargon and presenting it in a way that is easily understandable for the average user. The inference processes involved in these models are computationally expensive, leveraging deep learning techniques to provide real-time responses to user queries. This aspect of the project not only aligns with the computational challenges present in many real-world AI applications but also ensures that the project delivers timely and relevant information to its users, enhancing their experience and the application's overall utility.

### **Objectives:**

1. Develop an NLP model capable of translating complex legal jargon into simple language.
2. Create a user-friendly web application that allows users to upload terms and conditions for immediate analysis.
3. Provide a customized privacy grade for each document based on user-specified priorities.
4. Offer a search and comparison feature for users to find apps that align with their privacy preferences.
5. Recommend applications based on user-entered privacy concerns.

### **Learning Emphasis:**

- Application of NLP techniques to process and summarize text.

- Utilization of machine learning models for classification and recommendation systems.
- Deployment of a scalable web application.

### **Application Mock Design:**

- SAMMI

### **Research and Development:**

- Exploration of existing literature on NLP applications in legal document analysis.
- Review of case studies from projects like ToS;DR for insights into public engagement and data annotation quality.
- Examination of privacy laws and regulations to ensure compliance and relevance.

### **Fun Factor:**

- Engaging and interactive UI elements that demystify legal texts.

### **Limitations and Risks:**

- **Data Quality:** Reliance on crowd-sourced annotations may lead to inaccuracies. Plan to implement a verification system for annotations.
- **Scope of Data Collection:** Focused initially on popular consumer apps, potentially expanding to other types as the project scales.
- **Technical Constraints:** High demand for computational resources due to the complexity of the NLP models.

### **References (Cited of consulted as literature review)**

1. Aleecia M. McDonald and Lorrie Faith Cranor, "The Cost of Reading Privacy Policies," Journal of Policy for the Information Society, 4, no. 3 (2008), <http://hdl.handle.net/1811/72839>.
2. Deloitte. *The Consumer Data Privacy Paradox: How to Balance Privacy and Personalization*. Deloitte Insights, 2020, <https://www2.deloitte.com/us/en/insights/industry/technology/consumer-data-privacy-paradox.html>.
3. Kim, Nancy S. *Wrap Contracts: Foundations and Ramifications*. Oxford University Press, 2013.



4. Yannis Bakos, Florencia Marotta-Wurgler, and David R. Trossen, “Does Anyone Read the Fine Print? Consumer Attention to Standard-Form Contracts,” *Journal of Legal Studies* 43, no. 1 (2014): 1–35, <https://doi.org/10.1086/674424>
5. Tess Wilkinson-Ryan, “A Psychological Account of Consent to Fine Print,” *Iowa Law Review* 99 (2014): 1745
6. Thomas J. Maronick, “Do Consumers Read Terms of Service Agreements When Installing Software? A Two-Study Empirical Analysis,” *International Journal of Business and Social Research* 4, no. 6 (2014): 137–45
7. Mark A. Lemley, “Terms of Use,” *Minnesota Law Review* 91 (2006), <https://papers.ssrn.com/abstract=917926>
8. Nili Steinfeld, “‘I Agree to the Terms and Conditions’: (How) Do Users Read Privacy Policies Online? An Eye-Tracking Experiment,” *Computers in Human Behavior* 55 (2016): 992–1000, <https://doi.org/10.1016/j.chb.2015.09.038>
9. Victoria C. Plaut and Robert P. Bartlett, “Blind Consent? A Social Psychological Investigation of Non-readership of Click-Through Agreements,” *Law and Human Behavior*, June 16, 2011, 1–23, <https://doi.org/10.1007/s10979-011-9288-y>.
10. Margaret Jane Radin, *Boilerplate: The Fine Print, Vanishing Rights, and the Rule of Law* (Princeton, NJ: Princeton University Press, 2012)

