# Privacy Policy Report Card

AC215: Advanced practical data science, MLOps.

# Team members

Chad Stoughton, Christopher Lee

# Problem definition

We are used to consenting to a website’s or app’s privacy policy assuming that our personal data will be used responsibly. This is not always the case. Most privacy policies are either too long to read or too difficult to understand. This means that most of us will have no choice but to continue accepting terms that we don’t fully understand.

If we can create a NLP model that is able to comprehend long-written privacy policies and summarize the key points or classify the risk associated with accepting the terms, this would allow for users of the internet to have a better idea of how they are interacting with organizations online, and provide a safer internet for individuals.

Understanding how data on our interactions online is collected and the way they are used is a key step in acquiring control of our own digital footprints.

# Proposed solution

We propose training a model using NLP to recognize data usage and collection clauses within online privacy policies, and to present a user with a “report card” outlining what data the service collects about the user, and how that data is used. This should allow users to feel more confident in their decisions regarding online privacy.

# 

# Project scope

A minimum viable product would allow a user to enter the plain text of a privacy policy, and receive a simple summary of what kinds of data the service collects, whether that data is shared with third parties, and whether the user can opt-out of data collection.

A more sophisticated product might implement web-scraping, allowing the user to enter the URL of a privacy policy, and implement a visual “report card” presenting the model’s classifications to the user in an easily digestible format.

Deployed in the form of an app or website, our product would be able to analyze any privacy policy provided as input in the form of plaintext and deliver the corresponding report card as output.

In particular, questions we would like to see addressed include:

* What type of personal data is collected?
* What is the purpose for collecting such personal data?
* Is the collected personal data shared with third parties?
* ­Will I have access to all of the data that has been collected about me?
* Am I able to request all of my data to be deleted?

# Challenges

While there exist well annotated and researched datasets labeling privacy policy contents and components that are well suited to training models for this task, these datasets are often limited in scale. We have already identified a number of datasets that may provide viable training samples, however their size does not constitute a “big dataset”.

We propose two techniques to increase the complexity and sophistication of this project to meet the needs of this course. The first is to increase the size of the datasets via bootstrapping. By resampling these datasets with replacement, we can create a large dataset with statistical properties mirroring those of the underlying population (i.e. our bootstrapped dataset should show the various data collection and usage clauses at the same rate as they are found in privacy policies generally). This large dataset can then be fed into our data pipeline, demonstrating the scalability of our solution.

Our second proposed technique is to develop an ensemble model consisting of sub-models, each designed to identify one component of the “report card” as a binary response variable. This modularity should both increase the accuracy and ease of evaluating our model, and allow us to further develop our pipeline by training and running our sub-models in parallel.

# A rough timeline and components

* Sep 21-Oct 5:
  + Research previous work done (Github repos, publications)
  + collect data and exploratory data analysis
* Oct. 5 - Oct. 19:
  + Develop pipeline for training data pre-processing and model training
  + Begin training BERT & GPT models for pilot demonstration of one or two categories
  + Begin model evaluation
* Oct. 19 - Nov. 2:
  + Set up cloud environment
  + Designing and setting up APIs
  + Begin developing app frontend
* Nov 2 - Nov 16
  + Migrate application to production environment
  + Implementation of any extra features (web scraping, visualizations etc…)
* Nov 16 - Nov 30:
  + Create cloud deployment architecture and deploy to the web

# Datasets and models being considered

Models

* Pre-trained BERT models through HuggingFace library
* GPT-2 models through OpenAI’s library
* GPT-3 models through OpenAI’s API

Datasets

* OPP-115 Corpus (ACL 2016)
* Opt-out Choice Dataset (WWW 2020)
* APP-350 Corpus (PETS 2019)
* MAPS Policies Dataset (PETS 2019)

# 