Classifying Practices in Privacy Policy Text

Callum Hinchcliffe, BrainStation London Data Science Student, November 2022

This project seeks to identify data practices described by app developers in their data privacy policy. Using natural language processing and machine learning, can we summarise key information found in a privacy policy?

The challenge

How apps use our data is of relevance for anyone that has downloaded a smartphone app. The potential for misuse is great – if a company misuses or loses a user's data, that company or another party can commit identity theft, hacking and as documented by the Cambridge Analytica scandal, has the power to change political beliefs.

Companies specify how they handle user data in their privacy policies but these legal documents are often long, dense with legal jargon and generally unreadable, leaving users unable to understand what the company will do with their data and thus what the implications are for using the app.

But with modern natural language processing techniques it may be possible to extract the key information. If users understand the key ways that their data will be used, they can make an informed choice about using the app, and apps can be judged by their data privacy practices.

Background

The potential gains from applying machine learning techniques to the legal domain is great, as machine learning has, in other domains, made it faster to navigate the domain and reduced the expertise required. Challenges to applying ML to the legal domain include the expense required to annotate legal texts and the difficulty level of interpreting texts. Progress is being made as work is done to find better statistical and machine learning techniques to apply to the domain, as well as the increased interest in annotation.

A handful of publications have explored approaching the privacy policy domain using machine learning, including some work to identify suitable pre-processing and machine learning models for categorising privacy policies. This project seeks to replicate work by Story et al. 2019 who found some success with their methods.

Data Available

Data has been provided by Story et al. 2019 (link) and is available on UsablePrivacy.org. They collected 350 privacy policies from free apps on the Google Play store and annotated each segment (roughly equivalent to a paragraph), stating whether it listed that the app did or did not conduct an action with user data (some 'privacy practice'), and whether this practice related to a third party or not.

The data is structured as a YAML file, with each sub-level containing: metadata about the policy; policy text; each segment; annotations applied to that segment; and specific sentences from the segment with the annotation applied. Each annotation has three parts:

- The data practice, such as the processing of email address as a form of contact details;
- The parties involved, consisting of two options, either "1st party" (the app company) or "3rd party" (information is processed by a third party, such has as an analytics service provider);
- The "modality", of whether the practice is being described as being "performed" or "not performed", to account for cases in which an app developer makes it clear that they do not take some action with regards to the data.

Each segment can have multiple annotations.

Cleaning and preprocessing

Loading and cleaning data: I converted the data from YAML format into a table (Pandas DataFrame). At the highest level, each row contains a policy.

This was broken down so that each row contains a segment. I created separate classifiers to predict each part of the annotation. So for modelling, I then added each different target to be classified (data practice, parties or modality) as columns.

I also broke the data down further to the sentence level for further analysis and insights.

For modelling, I cleaned up the text by: removing repeated whitespace and punctuation, removing non-ASCII characters and converting all text to lowercase.

Further pre-processing involves "Crafted Features", "Sentence Filtering" and text vectorization. Specific classifiers were made to predict each of the three parts of an annotation, so both Crafted Features and Sentence Filtering relate to a specific classifier.

Crafted Features: as a way to improve performance of the ML model, specific phrases (produced by Story et al.) relating to each classifier were added as columns to the data to act as variables that may be predictive of the target. For example, the column relating to the phrase "exact device location", indicating whether a segment mentions this phrase, could help predict whether the segment details how the company uses GPS data.

Sentence Filtering: when training a classifier, in some cases performance is improved by only training it on segments that don't mention a related phrase. So different segments are filtered out when training each classifier.

Finally the segments are 'vectorized', a natural language processing technique that identifies every word and two-word-combination.

Initial Insights

- Despite 350 privacy policies consisting of a lot of text, some annotations are quite rare and so there is not so much data with which to train a model.
- The most common phrases were legal jargon
- Although sharing information with 3rd parties is an area of concern for users, this is detailed significantly less often than 1st party practices.

Modelling

For modelling, all policies were split into train, validation and test data. After initial baseline with logistic regression, following Story et al. I used SVM to predict each part of each annotation. I created a pipeline to create 18 different classifiers, tuning the processing and SVM setup for each classifier.

Evaluation metrics related to looking for policies that may not comply with the law by **not** describing certain practices that the company is performing. Many classifiers struggled to accurately predict when a class was not present but as a whole there is the ability to successfully achieve the goal of finding descriptions of some common and relevant data practices. There is also good room for improvement over the results of Story et al.

Conclusion

I was able to follow similar steps and approximately replicate the work by Story et al. It is not exactly clear how they trained each classifier and I would have to try more varieties of processing to see which come close to their performance. But even without trying many variations of preprocessing, the classifiers show promise in this domain. Most classifiers performed much better predicting the absence of the target than the presence of it.

In further work I will review my preprocessing pipeline to minimise data leakage, build an OOO framework, add further unit tests, explore the effects of different preprocessing and explore the performance of a wider range of models.

Paper mentioned:

Peter Story, Sebastian Zimmeck, Abhilasha Ravichander, Daniel Smullen, Ziqi Wang, Joel Reidenberg, N. Cameron Russell, and Norman Sadeh, "Natural Language Processing for Mobile App Privacy Compliance", AAAI Spring Symposium on Privacy Enhancing AI and Language Technologies (PAL 2019), Mar 2019 [pdf]