Classifying Practices in Privacy Policy Text

Callum Hinchcliffe, BrainStation London Data Science Student, Nov 2022, updated Jan 2023

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 relevant for anyone that has downloaded a smartphone app. The potential for misuse is great, enabling identity theft, hacking and as documented by the Cambridge Analytica scandal, changing political beliefs.

Companies specify how they handle user data in their privacy policies but these legal documents are often long, dense and full of legal jargon, leaving users unable to understand what happens to 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 it could improve the accessibility of legal texts and automate lengthy legal work. Challenges in applying machine learning to legal texts include annotation expense and interpretation difficulty.

A handful of publications have explored approaching the privacy policy domain using machine learning, including some work to identify suitable preprocessing 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, available at UsablePrivacy.org. They collected 350 privacy policies from free apps on the Google Play store. They annotated each ‘segment’

(roughly equivalent to a paragraph), noting whether the segment explained some action using user data (some ‘privacy practice’)

Specifically, the data is structured as a YAML file, with each sub-level containing metadata, the segment text, and a breakdown of the annotations. Each annotation has three parts:

- The data practice, such as the processing of email address as a form of contact details
- The parties involved, 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.

The Goal

The models I am building aim to predict the above annotations in a segment. I build two classifiers to predict each party, two to predict each modality, and 14 to predict different practices.

Cleaning and Preprocessing

Loading and cleaning: 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.

For analysis and modelling, I then added each different target to be classified (data practice, parties and modality annotations) as columns. I cleaned up the text by: removing repeated whitespace and punctuation, removing non-ASCII characters and converting all text to lowercase. For further insights and analysis, I also broke the data down to the sentence level.

Further preprocessing included text vectorization, “Crafted Features” and “Sentence Filtering”.

Crafted Features: as a way to improve performance of the models, I added as columns to the data specific phrases relating to each classifier (crafted by Story et al.). These act as variables that may be predictive of the target. For example, the column “exact device location” could be indicative of whether each segment details how the company uses GPS data.

To improve classifier performance, some segments are filtered out during training by only including those that mention a relevant phrase. This is referred to as sentence filtering.

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 in policies than 1st party practices.

Modelling

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

A goal of the research was to find policies that may break the law by **not** describing certain practices that the company is performing. So a key evaluation metric related was predicting that a segment did **not** contain an annotation. Across all classifiers, the mean average F1 score I achieved for this on the test set was 99%, which is on par with Story et al. Finding when a segment did contain an annotation was more difficult but my classifiers still scored moderately, and I made improvements over some of the results of Story et al., which was state of the art at the time.

Conclusion

Some classifiers for important metrics scored well across the board, showing that we can summarise key information found in a privacy policy using machine learning. I was able to approximately replicate and build upon the work by Story et al.

A number of different preprocessing and modelling methods could still be explored, but the classifiers already show promise in this domain.

A continuation of this project could explore the effects of different preprocessing and the performance of a wider range of models. With so many different classifiers, more specific model evaluation would be interesting, exploring the benefits and limitations of each model.

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]