Applying Explainable AI techniques to the APP-350 Corpus - MCS MidTerm Capstone Report

Tristan Koh

October 21, 2022

Abstract

This is the for the capstone project YSS4107 under the double degree law and liberal arts program. It is also intended to fulfil the MCS major capstone requirements. The code can be found here.

1 Definitions

- Explainable AI is a branch of AI that seeks to explain the predictions of machine learning models through statistical explanations and visualisations.
- Natural Language Processing (NLP) is a subset of machine learning that aims to quantifiably analyse natural language.
- APP-350 Corpus is a corpus of 350 Android app privacy policies annotated with privacy practices (i.e. behaviour that can have privacy implications).

2 Objective of capstone

Using Interpret Text, I aim to train a machine learning model using the APP-350 corpus to predict the type of privacy practices within app privacy policies. I then use Interpret Text tools to visualise why the model made such predictions from a machine learning perspective. I then compare and contrast this with how a lawyer would reason about the privacy policy.

3 Rationale of capstone

Natural language forms the bread and butter of the legal industry, as expressed in contracts, judgements and legislation. There has been an increasing trend within the legal industry to adopt more machine learning techniques to automate and assist low level legal analysis. Within the specific context of data privacy, it would be useful

to have a tool that assesses the possible data privacy risks of user policies. Such a tool would naturally use NLP techniques. Nevertheless, these tools should still be accessible to the layperson lawyer that might not be trained in data science.

While NLP techniques have substantially increased in performance in recent years, it has come at the cost of explainability of predictions because of the usage of neural networks which are architecturally more complex than traditional machine learning models. This lack of explainability of neural networks could potentially be a significant hindrance towards their adoption within the legal industry because the lawyer / law firm which uses these models still ultimately bear the responsibility of ensuring that the analysis is legally sound.

However, the intersection in skillset between data science and legal analysis is still nascent and it is unrealistic to expect all legally trained personnel to be trained in data science to the extent required to interpret the predictions of machine learning models without aid.

Recent research within the legal NLP space have focused on building higher performing models and word representations, but there have been comparatively few papers that assess the explainability of such models. Therefore, my capstone aims to bridge the gap between the lawyer and the data scientist by using Explainable AI techniques to explain the predictions of machine learning models in the context of predicting data privacy practices of app policies.

4 Explanation of Dataset ("The APP-350 corpus")

The APP-350 Corpus consists of 350 annotated Android app privacy policies. Each annotation consists of a practice and a modality.

A "privacy practice" (or "practice") describes a certain behaviour of an app that can have privacy implications (e.g., collection of a phone's device identifier or sharing of its location with ad networks). Altogether, 30 different practices were annotated.

There are two modalities: PERFORMED (i.e. a practice is explicitly described as being performed) and NOT_PERFORMED (i.e. a practice is explicitly described as not being performed).

The following is a table of the 30 practices and their descriptions.

Data Type	Description
Contact	The policy describes collection of unspecified contact data.
Contact_Address_Book	The policy describes collection of contact data from a user's address book on the phone.
Contact_City	The policy describes collection of the user's city.
Contact_E_Mail_Address	The policy describes collection of the user's e-mail.
Contact_Password	The policy describes collection of the user's password.
Contact_Phone_Number	The policy describes collection of the user's phone number.
Contact_Postal_Address	The policy describes collection of the user's postal address.
Contact_ZIP	The policy describes collection of the user's ZIP code.
Demographic	The policy describes collection of the user's unspecified demographic data.
Demographic_Age	The policy describes collection of the user's age (including birth date and age range).
Demographic_Gender	The policy describes collection of the user's gender.
Identifier	The policy describes collection of the user's unspecified identifiers.
Identifier_Ad_ID	The policy describes collection of the user's ad ID (such as the Google Ad ID).
Identifier_Cookie_or_similar_Tech	The policy describes collection of the user's HTTP cookies, flash cookies, pixel tags, or similar identifiers.
Identifier_Device_ID	The policy describes collection of the user's device ID (such as the Android ID).
Identifier_IMEI	The policy describes collection of the user's IMEI (International Mobile Equipment Identity).
Identifier_IMSI	The policy describes collection of the user's IMSI (International Mobile Subscriber Identity).
Identifier_IP_Address	The policy describes collection of the user's IP address.
Identifier_MAC	The policy describes collection of the user's MAC address.
Identifier_Mobile_Carrier	The policy describes collection of the user's mobile carrier name or other mobile carrier identifier.
Identifier_SIM_Serial	The policy describes collection of the user's SIM serial number.
Identifier_SSID_BSSID	The policy describes collection of the user's SSID or BSSID.
Location	The policy describes collection of the user's unspecified location data.
Location_Bluetooth	The policy describes collection of the user's Bluetooth location data.
Location_Cell_Tower	The policy describes collection of the user's cell tower location data.
Location_GPS	The policy describes collection of the user's GPS location data.
Location_IP_Address	The policy describes collection of the user's IP location data.
Location_WiFi	The policy describes collection of the user's WiFi location data.
SSO	The policy describes receiving data from an unspecified single sign on service.
Facebook_SSO	The policy describes receiving data from the Facebook single sign on service.

Table 1: List of annotated data privacy practices and their descriptions.

The APP-350 Corpus was used in a broader project to train machine learning models to conduct a privacy census of 1,035,853 Android apps. In that project, the researchers downloaded the data privacy practices of all apps from the Play Store with more than 350 million installs (which totalled 247 apps) and 103 randomly selected apps with 5 million installs. In total, the researchers collected the data privacy policies of 350 apps.

All 350 policies were annotated by one of the authors, a lawyer with experience in data privacy law. To ensure reliability of annotations, 2 other law students were hired to double annotate 10% of the corpus. With a mean of Krippendorff's $\alpha = 0.78$[1], the agreement between the annotations exceeded previous similar research.

For more information about how the Corpus was annotated, see the paper "MAPS: Scaling Privacy Compliance Analysis to a Million Apps", Section 3, Pg 69 to 70.

5 Rationale for utilising the APP-350 corpus

Since the focus of this capstone is to assess the interpretability of XAI models specifically within a legal context, this dataset was chosen for the following reasons:

[1]Krippendorff's α is a measure of agreement, with $\alpha > 0.8$ indicating good agreement, $0.67 <= \alpha <= 0.8$ indicating fair agreement, and $\alpha < 0.67$ indicating doubtful agreement.

1. APP-350 contains real-world data privacy practices as they were scraped from Google PlayStore apps. Thus training XAI models on such a dataset would provide a realistic insight into the extent of which AI models are explainable in the legal context.

2. Legal tech companies are also using such datasets to train models as part of their contract / document review products. By using APP-350 to train XAI models, the results can be used as a (simple)[2] proxy for the explainability of models that are currently used in the industry.

3. APP-350 is a labelled dataset, allowing easy validation of results. If an unlabelled dataset was used, unsupervised training would have to be conducted. The performance of the models would likely be much lower because NLP models for specific vocabulary like law are still not as sophisticated as models trained on general vocabulary. Further, there are few law specific labelled datasets to begin with.

4. APP-350 is labelled on both the sentence and segment (i.e. paragraph) level. This provides more granular data for training the AI models.

6 Methodology

There are three major steps to the capstone: First is data pre-processing, second is model training and applying XAI techniques to visualise their predictions ("Model Training and XAI visualisation"), and the last is to survey law and non-law students about whether they find these explanations interpretable ("Survey to assess interpretability"). I describe the specific methodology to the parts below.

6.1 Data pre-processing

The annotated privacy policies were originally in .yml format, with one .yml file containing one app data privacy policy. As explained above, each data privacy policy is labelled at both the sentence and segment level. Therefore I restructured the data from .yml to .csv, with one .csv file containing annotated sentences and the other containing annotated segments. By having two levels of text data for model training, this would provide another feature to compare model performance on.

6.2 Model training and XAI visualisation

As the original researchers used the same dataset to train classifiers to predict on unseen data privacy policies, I use their training methodology and model choice as a guide for this capstone.

[2]The datasets used in industry are usually much larger and the models used are more complicated. However, APP-350 would be sufficiently complicated to serve as a toy example at an undergraduate level.

The original researchers feature engineered the policy text as follows (Page 71 to 72):

1. Tokenisation: Lowercase all characters, remove non-ASCII characters, no stemming, normalisation of whitespace and punctuation, unigrams and bigrams.
2. Word representation: Union of TF-IDF and manually crafted features.
   (a) The manually crafted features consist of Boolean values indicating the presence or absence of indicative strings the researchers observed in the data.

Individual classifiers were then trained for every policy classification. For all the classifications (except for four policy classifications), they trained a model using scikitlearn SVC implementation with a linear kernel, with five-fold cross validation. For the four policy classifications, word-based rule classifiers were used instead because of the limited number of training data.
(Add performance of researchers' models here.)

6.2.1 Proposed model training methodology

There are three main factors that I vary in order to produce the different explanations of AI models: Model, word representation and XAI method used to explain the trained model.

6.2.2 Model choice

As the researchers found that the scikit-learn SVC classifier produces the best performance, I use SVC as well. I use the following models:

1. Logistic regression
2. SGDClassifier
3. SVC
4. Ensemble classifiers

I include logistic regression as a baseline classifier for its simplicity. SGDClassifier functions as a possible alternative to SVC since SGDClassifier can use a linear SVM loss function. I include ensemble classifiers as they have generally worked well with NLP scenarios (?)

The two broad types of AI models that are usually used are traditional AI models and neural networks. I will only be experimenting with the first type of AI model as these models are considered "glass-box" models: A glass-box model implies a model that is innately explainable, where the user can fully observe and dissect

the process adopted by the model in making a prediction. The hyperparameters of neural networks are much more complicated. Generally, neural networks require (add number here) datapoints to perform well. However, we only have (add number here) data points spread out across 30 different privacy practices. Therefore, there is insufficient data for robust training of neural networks. Hence, neural networks are out of the scope for the purposes of this capstone.

6.2.3 Word representation

1. Word representations: Bag of words and word embeddings.
2. Different combinations of n-grams.

6.2.4 XAI packages

1. LIME
2. SHAP
3. OmniXAI
4. Interpret Text

   Interpret Text allows visualisation of the predictions of NLP models across a range of complexities.

   For linear and tree-based models, Interpret Text uses a Classical Text Explainer. The Explainer leverages this inherent explainability by exposing weights and importances over encoded tokens as explanations over each word in a document. In practice, these can be accessed through the visualization dashboard or the explanation object.

   The explanations provided by the aforementioned glass-box methods serve as direct proxies for weights and parameters in the model, which make the final prediction.

6.3 Survey to assess interpretability

To be added.

7 Findings so far

7.1 Exploratory Data Analysis

7.1.1 Sentence level

7.1.2 Segment level

7.2 Performance of classifiers for top N practices

(Add statistics on existing categories here). Given that there are 20 categories for the entire dataset, but there is not a uniform distribution of occurrences.
Thus, to find an optimal balance between model performance and still maintain a realistic sample of practices that could appear in a real world dataset, I chose to assess model performance by taking the top N frequently occurring practices.

7.3 Performance of individual classifiers for top 5 practices

7.4 XAI visualisation of models

8 Further steps to take in Semester 2