# **Contact information for Official Representative:**

**Name: Alan deLevie**

**Email:** [**adelevie@gmail.com**](mailto:adelevie@gmail.com)

**Team Name: TeamAlan**

# **Names of additional team members:**

**Name:**

**Name:**

**Name:**

# **Introduction to Team:**

Alan deLevie is a software engineer working on data science applications in the legal domain.

# **Executive Summary of Solution:**

The solution is an attempt at balancing simplicity and performance. The model architecture is based on BERT, which revolutionized NLP in 2018. Although BERT is very advanced, it is decently well-understood. Moreover, the code powering much of the modeling is from the HuggingFace/transformers library. Transformers is incredibly easy to use and represents a giant leap forward in terms of developer ergonomics around AI/ML.

The model is pre-tuned on privacy policies and EULAs from around the web, and fine-tuned an GSA’s EULA data.

On top of the model, the user interface is very spartan. It’s just a text box where the user can paste text copied from any source, including PDFs and Word docs. The predictions visualization is also very simple. It’s just a Python Flask server and HTML (no javascript or stylesheets). While usable on its own, it can quicky and simply be integrated into other systems.

There is an included notebook that shows how the models were trained, and also includes a section on interpreting any given prediction. The notebook was run in Google Colab and can run from start to finish in just about an hour.

The model achieved a Brier loss of 0.102729 and an F1 score of 0.543611. The included notebook details several iterations of the model and the performance metrics for each iteration (including precision and recall).

# **BERT Architecture:**

## **Technology Scope:**

* The solution uses a DistilBERT model (a version of BERT with fewer weights), with additional pre-training on EULAs from around the web and privacy policies from the OPP-115 data set.

## **Functionality and User Interface:**

* What type of user interface does the solution provide (e.g. web interface, command line interface).
* The user interface is a web interface, served by a Python Flask server
* Out of the box, the web interface does not handle batches of documents, but the server and notebook provide example code to enable batch document processing.

## **Application of Artificial Intelligence/Machine Learning (AI/ML):**

* The solution uses a DistilBERT model (a version of BERT with fewer weights), with additional pre-training on EULAs from around the web and privacy policies from the OPP-115 data set.
* This model utilizes semi-supervised learning with supervised learning and transfer learning.
  + The model pre-training creates ablations in unlabeled text and trains by guessing the missing word.
  + The model fine-tuning is supervised training on the EULA data provided by GSA.
  + The model utilizes transfer learning, first by building a general language model on large corpora of text (e.g. Wikipedia) and second by further pre-training on text in the same domain as the downstream classification task.
* More info on DistilBERT here: <https://arxiv.org/pdf/1910.01108.pdf>

NOTE: Please do not submit any sensitive or classified information.