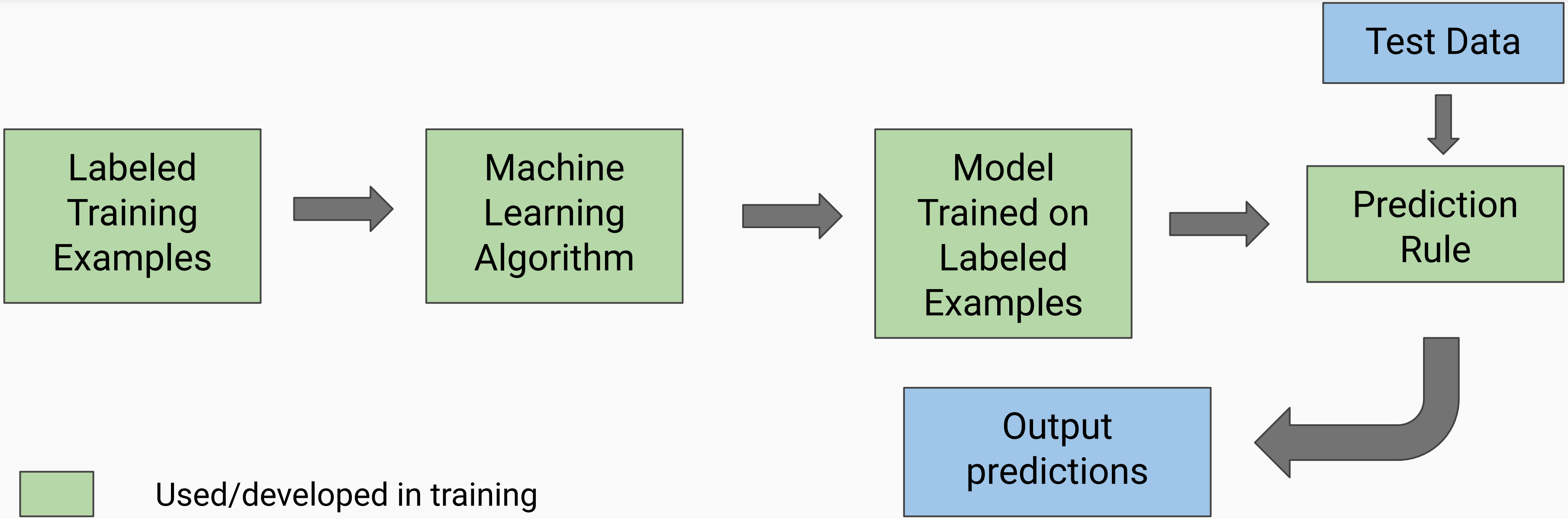
GSA EULA Machine Learning

Rohan Narain + Andrew Kim

The Problem

- End User License Agreements must adhere to specific regulations when involving software and products sold to the government.
- Some End User License Agreements contain clauses that may violate regulations, and it can be very difficult to find these clauses.
- By leveraging Artificial Intelligence/Machine Learning, one can detect the presence of clauses that violate regulations and figure out which clauses violate them.

Supervised Machine Learning Process



Labeled Training Examples

Machine Learning Algorithm

Model Trained on Labeled Examples

Test Data

Prediction Rule

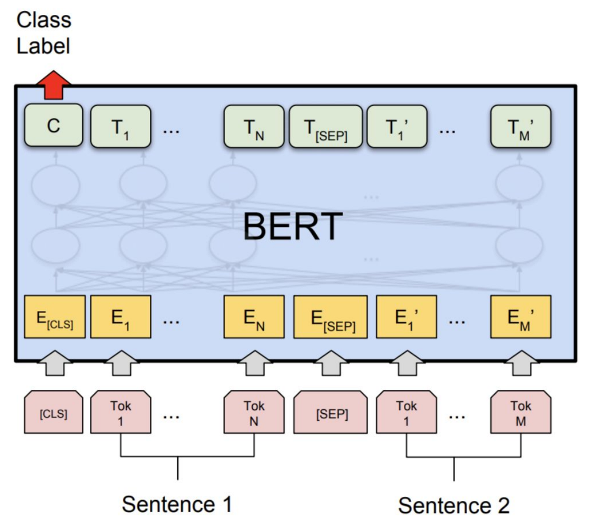
Output predictions

🟩 Used/developed in training

🟦 Provided/accessed by user

The Model: BERT

- BERT is a transformer model--a deep neural net that efficiently extracts **contextual information** from every part of the input sentence
- Transformer models are the **state-of-the-art** in many natural language processing tasks, including text classification
- Models released by research company HuggingFace are pre-trained on massive text corpus--using a pre-trained model and fine-tuning it is called **transfer learning**



Scaling BERT: DistilBERT

- In 2019, HuggingFace released DistilBERT, a version of BERT with only 60% of the parameters, but **maintains 95% of the performance** of the original BERT
- DistilBERT makes using BERT in a production system very easy

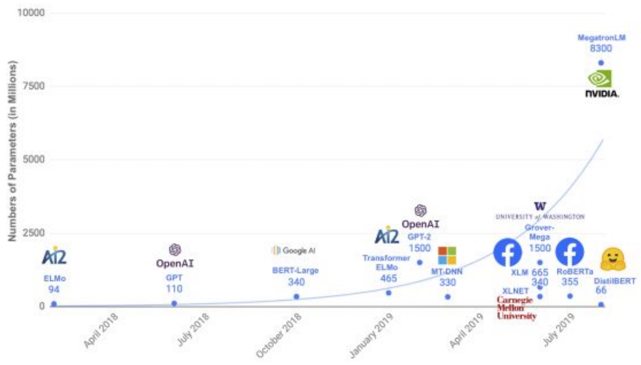
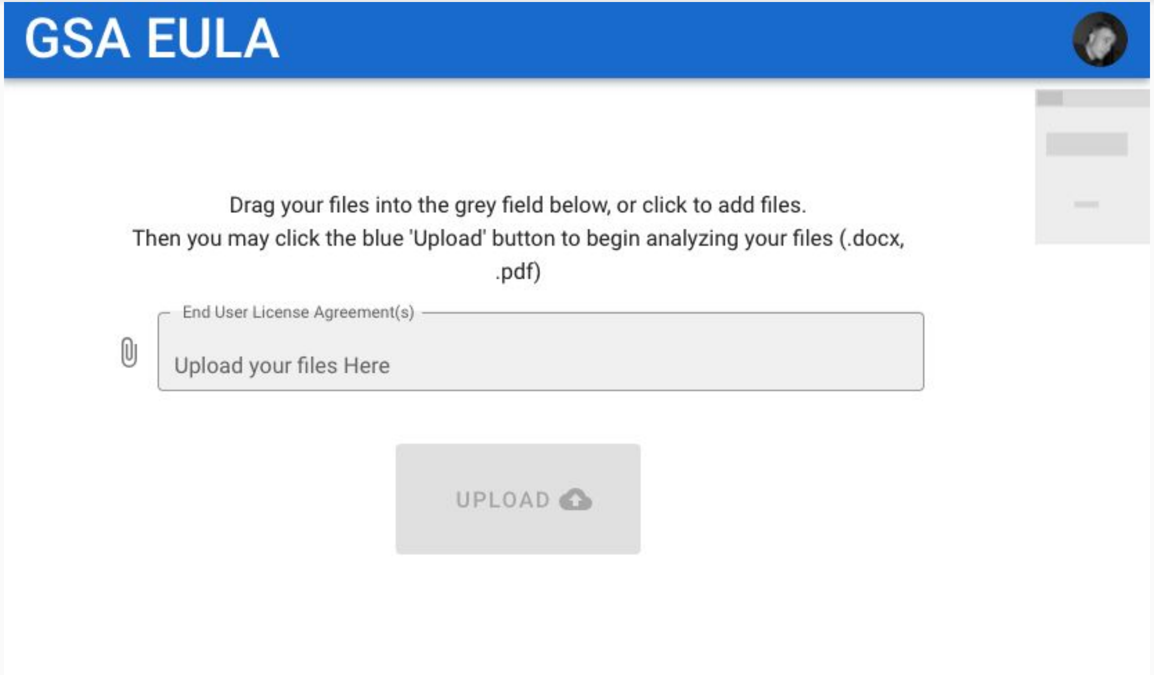


Figure 1: **Parameter counts of several recently released pretrained language models.**

Data Format

- Sample data was given in the following format:
  - Three columns: Clause ID, Clause Text, Classification
  - Classification: 0 (acceptable) or 1 (unacceptable)
- Data can be provided by the user as either a Microsoft Word Document in .docx format, or in Portable Document Format (PDF).

The Solution: EULA-ML



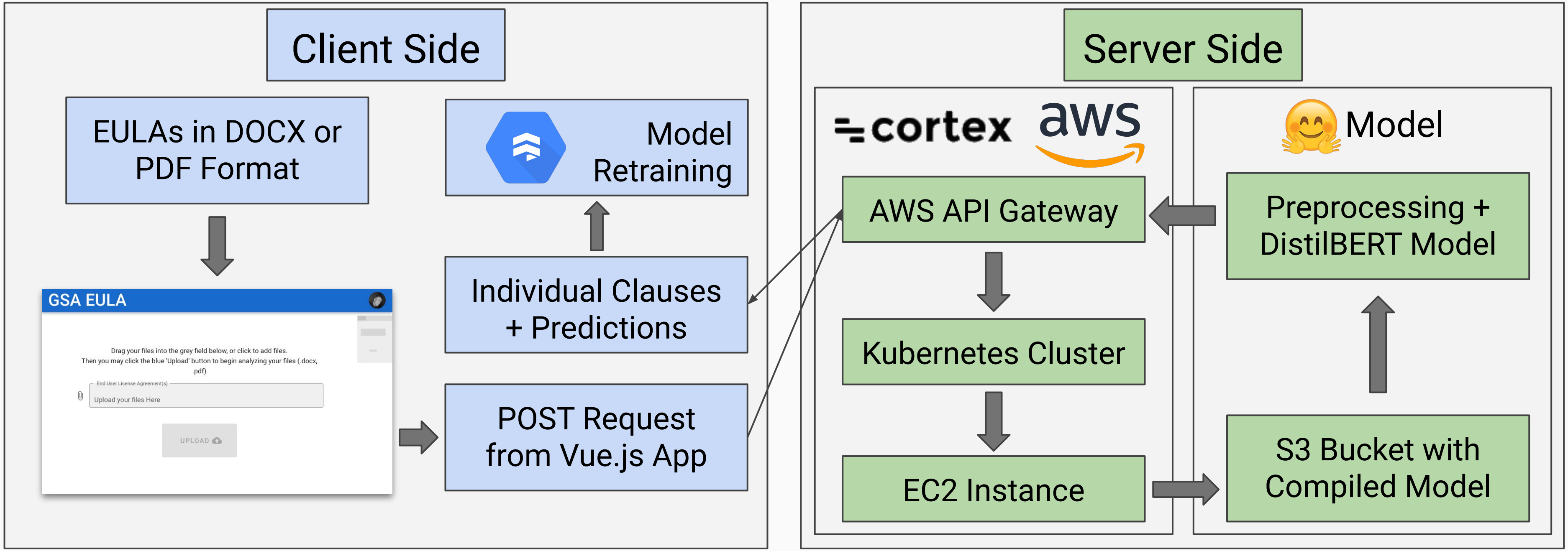
GSA EULA

Drag your files into the grey field below, or click to add files.
Then you may click the blue 'Upload' button to begin analyzing your files (.docx, .pdf)

End User License Agreement(s)

Upload your files Here

How Does It Work?



Client Side

EULAs in DOCX or PDF Format

GSA EULA

Drag your files into the grey field below, or click to add files.
Then you may click the blue 'Upload' button to begin analyzing your files (.docx, .pdf)

POST Request from Vue.js App

Individual Clauses + Predictions

Model Retraining

Server Side

AWS API Gateway

Kubernetes Cluster

EC2 Instance

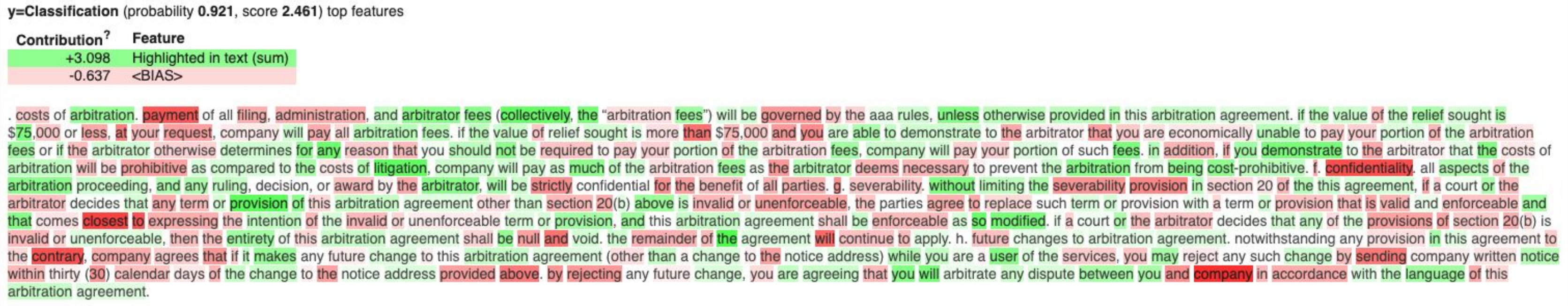
Model

Preprocessing + DistilBERT Model

S3 Bucket with Compiled Model

Model Performance and Explanation

- Weighted F-1 score of 0.87 / 1
- Brier score of 0.09 / 1
- Baseline model scores -- F-1: 0.73, Brier: 0.54
- Non-neural models: XGBoost, SVM, Random Forests…
  - Best F-1: 0.78, Best Brier: 0.18



y=Classification (probability **0.921**, score **2.461**) top features

Contribution?	Feature
+3.098	Highlighted in text (sum)
-0.637	<BIAS>

. costs of arbitration. payment of all filing, administration, and arbitrator fees (collectively, the “arbitration fees”) will be governed by the aaa rules, unless otherwise provided in this arbitration agreement. if the value of the relief sought is $75,000 or less, at your request, company will pay all arbitration fees. if the value of relief sought is more than $75,000 and you are able to demonstrate to the arbitrator that you are economically unable to pay your portion of the arbitration fees or if the arbitrator otherwise determines for any reason that you should not be required to pay your portion of the arbitration fees, company will pay your portion of such fees. in addition, if you demonstrate to the arbitrator that the costs of arbitration will be prohibitive as compared to the costs of litigation, company will pay as much of the arbitration fees as the arbitrator deems necessary to prevent the arbitration from being cost-prohibitive. f. confidentiality. all aspects of the arbitration proceeding, and any ruling, decision, or award by the arbitrator, will be strictly confidential for the benefit of all parties. g. severability. without limiting the severability provision in section 20 of the this agreement, if a court or the arbitrator decides that any term or provision of this arbitration agreement other than section 20(b) above is invalid or unenforceable, the parties agree to replace such term or provision with a term or provision that is valid and enforceable and that comes closest to expressing the intention of the invalid or unenforceable term or provision, and this arbitration agreement shall be enforceable as so modified. if a court or the arbitrator decides that any of the provisions of section 20(b) is invalid or unenforceable, then the entirety of this arbitration agreement shall be null and void. the remainder of the agreement will continue to apply. h. future changes to arbitration agreement. notwithstanding any provision in this agreement to the contrary, company agrees that if it makes any future change to this arbitration agreement (other than a change to the notice address) while you are a user of the services, you may reject any such change by sending company written notice within thirty (30) calendar days of the change to the notice address provided above. by rejecting any future change, you are agreeing that you will arbitrate any dispute between you and company in accordance with the language of this arbitration agreement.

Areas of Improvement

- Parsing clauses is imperfect--currently using a simple heuristic, but the formatting of word documents and PDFs can be tough to work with.
- Model retraining
- Adding in explainability
- UI enhacement for sorting clauses by character length, probabilities, predictions (perhaps a table)
- Allow multiple users to provide their feedback on re-labeling our predictions (several stakeholders may have differing opinions)
- Better client side validation of file input (where upload of files happens)

Thank you!