Submission for GSA Artificial Intelligence (AI) Machine Learning (ML) EULA Challenge 2020

1. Contact information for Official Representative:

Name: Mohamed Hassan Kane
Email: hassanmohamed0407@gmail.com
Team Name: mhkane

2. Names of additional team members:

Name: Pascal Notsawo

3. Introduction to Team:

Hassan is a machine learning enthusiast with over 5 years of experience developing AI systems in academic and industry settings. In my current industry role,I am Lead Data Scientist at Emtropy Labs, where I build NLP models to analyze customer service conversations. Before diving into the startup world, I studied Computer Science at MIT. As an undergraduate student, I co-founded the undergraduate machine intelligence community, interned at Facebook twice and developed radar perception algorithms for Uber ATG.

Pascal Notsawo is a rising senior at Ecole Nationale Superieure Polytechnique de Yaounde (ENSPY) where he studies computer science. He has taken numerous classes in machine learning and NLP online to supplement his coursework.

4. Executive Summary of Solution:

The provided solution has 3 components: a notebook showing how to train and finetune RoBERTa, a version of BERT by Facebook which is better pre-trained. BERT is a language representation model which stands for Bidirectional Encoder Representations from Transformers. BERT is designed to pre-train deep bidirectional representations from unlabeled text by jointly conditioning on both left and right context in all layers. As a result, the pre-trained BERT model can be fine-tuned with just one additional output layer to create state-of-the-art models for a wide range of tasks, such as question answering and language inference, without substantial task-specific architecture modifications.

Besides the notebook, we have a web application enabling a business user to upload a file in PDF or Doc format and have it be parsed clause by clause with an output per clause.

Finally, we also have a command line interface enabling a user to specify a model, input file and returns an output file containing the parsed clause with acceptability or unacceptability.

5. Solution Architecture:

a. Technology Scope:

- For the transformer model, we use ktrain which is a lightweight keras wrapper to help build, train, debug, and deploy neural networks in the deep learning software framework, Keras. (As of v0.7, ktrain uses tf.keras in TensorFlow instead of standalone Keras.)
- For the web application, we use a Django server to process file and use lighter bag of word and TF-IDF + logistic regression to process files

b. Functionality and User Interface:

- The most accurate interface is the notebook which enable users to use transformers to analyze documents at a clause by clause level.
    - We obtain a 0.7412 F1 score on the training set
    - We obtain a 0.069 brier score on the training set
- We also have a web application enabling users to input PDF or MS Word files which are individually parsed into clauses and then ran through our models
- We also have a command line interface supporting google docs, PDFs

c. Application of Artificial Intelligence/Machine Learning (AI/ML):

- The models are trained using supervised learning on the training set