**\*\*\* This document to be submitted via email to** [**challenge2020@gsa.gov**](mailto:challenge2020@gsa.gov) **\*\*\***

# **Contact information for Official Representative:**

**Name:** Sample Submission

**Email:** sample.submission@gmail.com

**Team Name:** Sample Submission

# **Names of additional team members:**

**Name:**

**Name:**

**Name:**

# **Introduction to Team:**

Sample is an undergraduate student studying computer science at University. She is fascinated by natural language processing and its applications, especially in text sentiment analysis and topic modeling.

# **Executive Summary of Solution:**

This solution consists of a custom text classifier in Amazon Comprehend that can be used to determine whether a certain contract clause is acceptable under federal contractor regulations, or if it appears to be problematic. The classifier accepts text input from users, and labels the input from the user as either Acceptable or Toxic. The classifier was trained on 3036 clauses from End User License Agreements (EULAs) provided by the GSA.

# **Amazon Comprehend Architecture:**

## **Technology Scope:**

* This solution uses services in AWS, including AWS Identity and Access Management (IAM) to manage access to AWS services and resources, Amazon Comprehend for text classification services, Amazon S3 to store training and validation data, and Amazon Cloud9 to allow for code implementation within a browser.
  + Amazon Comprehend is a NLP service, hosted by Amazon Web Services, that utilizes machine learning (ML) to discover relationships in texts. It can be used to perform custom classification on various documents.

## **Functionality and User Interface:**

* This solution consists of an endpoint that enables users to input text in the AWS Command Line Interface to conduct real-time analysis using an Amazon Comprehend custom classifier, as well as an AWS Cloud9 integrated development environment (IDE) that allows users to run classification jobs for batches of documents in their browser.
* To analyze a singular clause, the user can simply input the clause text into their Amazon Comprehend real-time analysis console, and the classifier will output (1) its prediction for the input as either Acceptable or Toxic, and (2) classification confidence metrics like Accuracy, Precision, and Recall.
* To analyze multiple clauses in a document or batch of documents, the user can store data in a CSV file (one per document) with one column, such that every row consists of one contract clause. After uploading the CSV file(s) to the Amazon S3 Bucket associated with the classifier, the user can implement an inference job using the AWS Cloud9 IDE.

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

* The Amazon Comprehend technology used in this solution leverages supervised machine learning, as it maps input text to output labels (Acceptable vs. Toxic). The classifier uses Amazon Comprehend’s capabilities for training, which is the process of creating a machine learning algorithm based on the GSA-provided dataset of acceptable and toxic EULA clauses. The Comprehend classifier then provides inference, which is the process of using the trained machine learning algorithm of the classifier to make a prediction. Namely, users can input a clause, a document, or a batch of documents, and the classifier will predict whether the input is Acceptable or Toxic.

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