# **Description of Methods:**

First, we ran the provided training and test data through the Microsoft AI key phrase extractor. This gave us a list of relevant words and phrases which best describe the subject of each clause provided. This information allowed our models to find patterns in key words and phrases which were more likely appear within invalid clauses.

We split the training data to have a separate validation set to calculate metrics from our trained models. We ran several standard machine learning classification models and found that they all performed similarly in terms of F1 scores and Brier loss scores. However, they had different rates of false negatives and false positives. Our team decided that minimizing false negatives was more important in the context of identifying invalid clauses. We were able to tolerate false positives more since it would be better if a contracting officer to review more documents that are flagged as invalid than to miss problematic documents that are flagged as valid. Thus, we chose the random forest classifier.

A user is able to use this by opening the document on Microsoft Word Online using an Ictect account. Under the “EULA Review Tool” tab, users can click the “Review EULA” button to scan the document for potentially invalid clauses; clauses which our model predicts are invalid will be highlighted in yellow. If the user wants more insight as to why a particular clause was predicted to be invalid, they can click on the “Why Was This Flagged?” button for the clause in question. If the clause contains any words or phrases which are included in the list of the 50 most important features to our random forest model (see: Reasons for predictions made in the Validation Data File submission), this button will highlight them red.

# **Description of the data, methods and software used to complete the solution:**

In order to extract the key phrases, we utilized our Microsoft Azure account’s access to the Microsoft key phrase extractor.

We drew from the Python sci-kit learn package for the machine learning models.

We utilized mainly Node.js to implement our User Interface.

# **Process used to complete the model used in the solution, including data inputs and visualizations:**

* + - 1. Get key phrases from Microsoft AI
      2. Remove punctuation and stop-words
      3. Vectorize text using bag-of-words approach
      4. Split training data into a training set and validation set
      5. Run various machine learning classification models
      6. Generate metrics on the validation set from part (4) for all the models (see figure below)
      7. Choose model based on criteria explained in part (1)
      8. Use the trained model to predict on the test set provided by GSA

A screenshot of a cell phone

Description automatically generated

# **Reasons for predictions made in the Validation Data File submission:**

The 50 most important features to the random forest model are listed below. Words, such as 'expenses', 'fees', 'damages', 'payment', and 'costs', could refer to unspecified future fees or penalties, which is in violation with the FAR Clause. 'Rights', 'license', 'software' may indicate indemnities. If a clause contains the words 'termination' or 'breach', this would raise a concern for possible unilateral termination by the contractor for a breach.

['agreement', 'company', 'terms', 'notice', 'customer', 'fees', 'software', 'party', 'written', 'services', 'state', 'conditions', 'law', 'license', 'days', 'gsa', 'rights', 'right', 'parties', 'breach', 'expenses', 'damages', 'applicable', 'time', 'use', 'event', 'liability', 'accordance', 'information', 'licensee', 'subscription', 'date', 'payment', 'entity', 'termination', 'term', 'order', 'ordering', 'taxes', 'contract', 'activity', 'courts', 'legal', 'companys', 'obligations', 'costs', 'section', 'writing', 'consumer', 'service']

# **Self-reported metrics:**

## **Brier score:** 0.19322

## **F1 score:** 0.81932

## **Accuracy:** 0.8068

## **Optimal Threshold:** 0.23

## **Recall:** 0.78114

## **Precision:** 0.71585