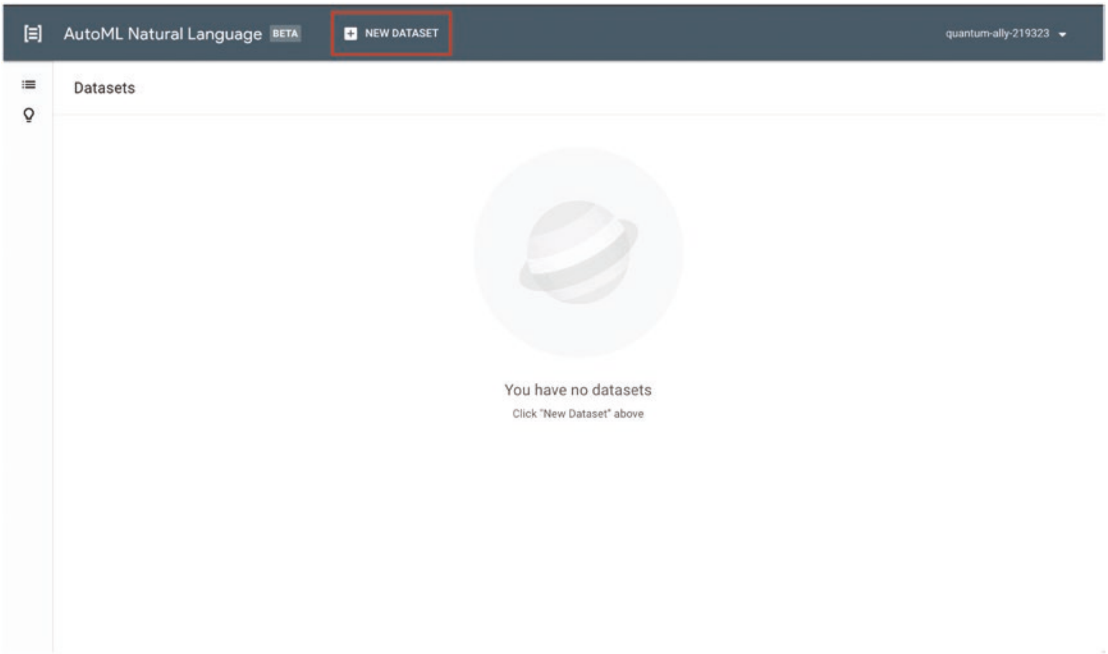
Building a Custom Language Classification Model on Cloud AutoML NLP

This section will walk through creating a document dataset and building a custom language classification model on AutoML Vision.

1. From the Cloud AutoML NLP dashboard, click **NEW DATASET** as shown in Figure 43-5.

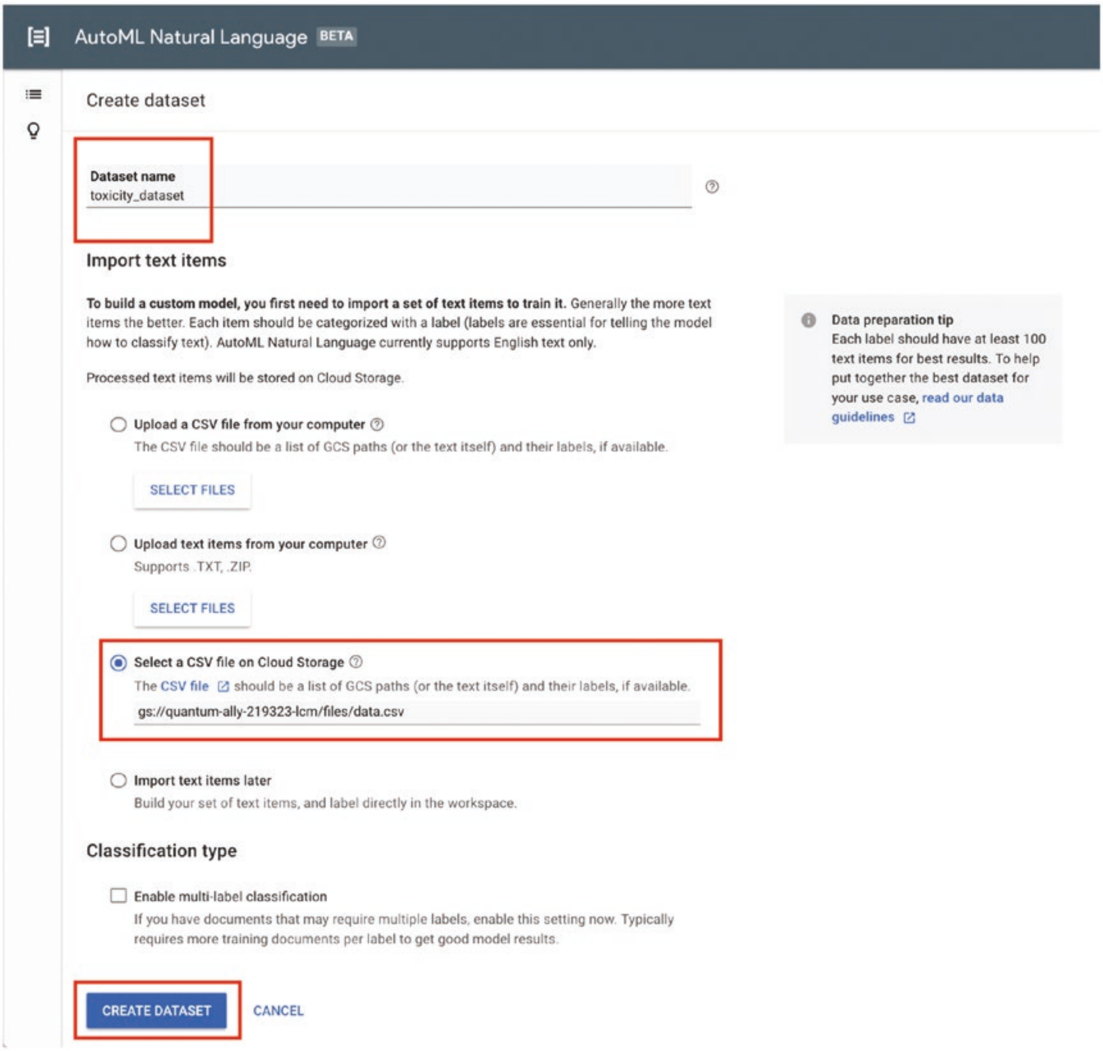


You have no datasets

Click "New Dataset" above

Figure 43-5. *New Dataset on AutoML NLP*

2. To create a Dataset on Cloud AutoML NLP, set the following parameters as shown in Figure 43-6:



AutoML Natural Language BETA

Create dataset

Dataset name
toxicity_dataset

Import text items

To build a custom model, you first need to import a set of text items to train it. Generally the more text items the better. Each item should be categorized with a label (labels are essential for telling the model how to classify text). AutoML Natural Language currently supports English text only.

Processed text items will be stored on Cloud Storage.

○ Upload a CSV file from your computer
The CSV file should be a list of GCS paths (or the text itself) and their labels, if available.

SELECT FILES

○ Upload text items from your computer
Supports .TXT, .ZIP.

SELECT FILES

● Select a CSV file on Cloud Storage
The CSV file should be a list of GCS paths (or the text itself) and their labels, if available.
gs://quantum-ally-219323-lcm/files/data.csv

○ Import text items later
Build your set of text items, and label directly in the workspace.

Data preparation tip
Each label should have at least 100 text items for best results. To help put together the best dataset for your use case, read our data guidelines

Classification type

□ Enable multi-label classification
If you have documents that may require multiple labels, enable this setting now. Typically requires more training documents per label to get good model results.

CREATE DATASET CANCEL

Figure 43-6. Create a Dataset on Cloud AutoML NLP

a. Dataset name: toxicity_dataset.

b. Select a CSV file on Cloud Storage (this is the CSV file placed on the bucket created when Cloud AutoML was configured that contains the path to the text documents): gs://quantum-ally-219323-lcm/file/data.csv.

c. Click **CREATE DATASET** to begin importing images (see Figure 43-7).

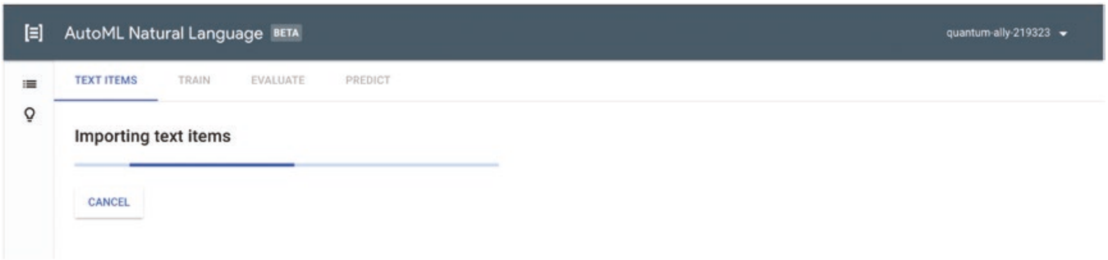
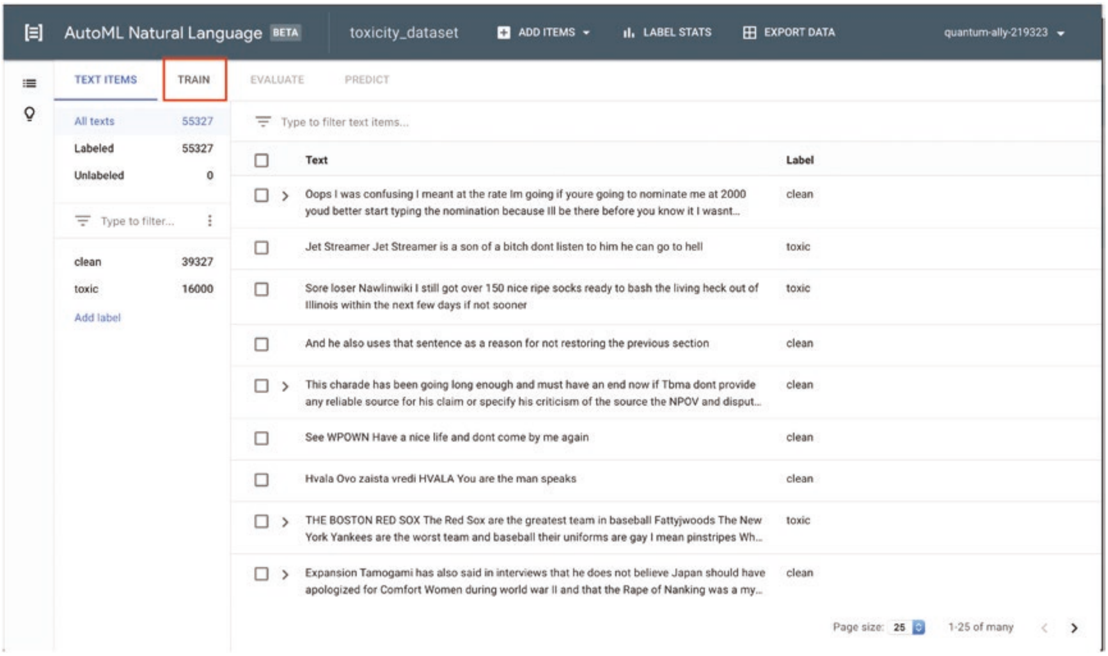


Figure 43-7. *Cloud AutoML NLP: Importing text items*



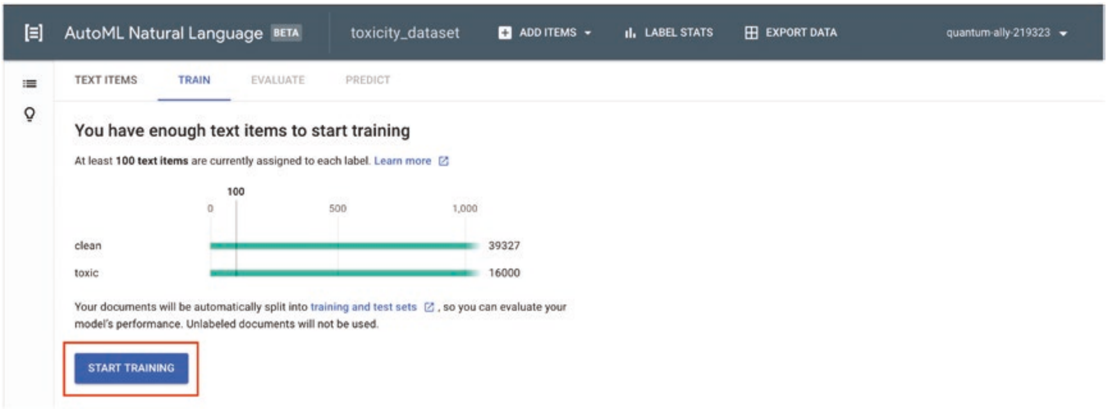
All texts	55327
Labeled	55327
Unlabeled	0
clean	39327
toxic	16000

Text	Label
Oops I was confusing I meant at the rate Im going if youre going to nominate me at 2000 youd better start typing the nomination because Ill be there before you know it I wasnt...	clean
Jet Streamer Jet Streamer is a son of a bitch dont listen to him he can go to hell	toxic
Sore loser Nawlinwiki I still got over 150 nice ripe socks ready to bash the living heck out of Illinois within the next few days if not sooner	toxic
And he also uses that sentence as a reason for not restoring the previous section	clean
This charade has been going long enough and must have an end now if Tbma dont provide any reliable source for his claim or specify his criticism of the source the NPOV and disput...	clean
See WPOWN Have a nice life and dont come by me again	clean
Hvala Ovo zaista vredi HVALA You are the man speaks	clean
THE BOSTON RED SOX The Red Sox are the greatest team in baseball Fattyjwoods The New York Yankees are the worst team and baseball their uniforms are gay I mean pinstripes Wh...	toxic
Expansion Tamogami has also said in interviews that he does not believe Japan should have apologized for Comfort Women during world war II and that the Rape of Nanking was a my...	clean

Figure 43-8. *Cloud AutoML NLP: Imported text documents and their labels*

3. After importing the Dataset, click **TRAIN** (see Figure 43-8) to initiate the process of building a custom language classification model.

4. In this example, we have a good enough number of training examples as seen in Figure 43-9, so hopefully, it makes sense to expect a good language classification model. Click **START TRAINING** to begin the training job.



AutoML Natural Language BETA toxicity_dataset ADD ITEMS LABEL STATS EXPORT DATA quantum-ally-219323

TEXT ITEMS TRAIN EVALUATE PREDICT

You have enough text items to start training

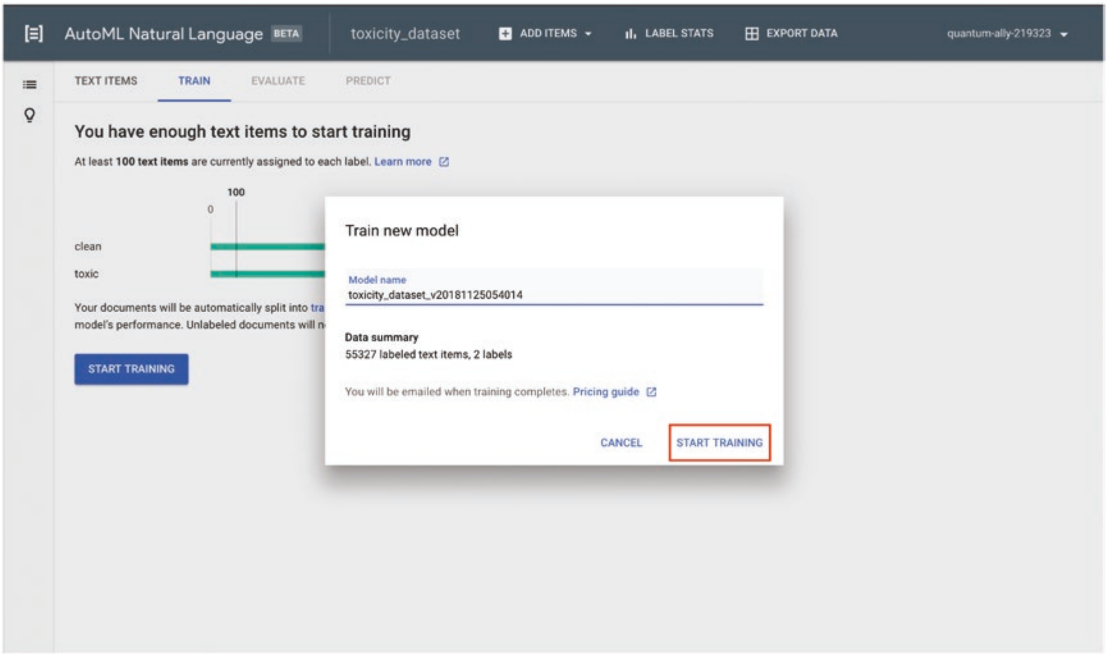
At least **100 text items** are currently assigned to each label. Learn more

Your documents will be automatically split into training and test sets, so you can evaluate your model's performance. Unlabeled documents will not be used.

START TRAINING

Figure 43-9. Cloud AutoML NLP checking the adequacy of training examples

5. Accept the default model name, and click **START TRAINING** (see Figure 43-10) to begin building the model as seen in Figure 43-11. Note that this training might take about an hour to complete. When done, the user will get an email of completion.



AutoML Natural Language BETA toxicity_dataset ADD ITEMS LABEL STATS EXPORT DATA quantum-ally-219323

TEXT ITEMS TRAIN EVALUATE PREDICT

You have enough text items to start training

At least **100 text items** are currently assigned to each label. Learn more

Your documents will be automatically split into tra
model's performance. Unlabeled documents will n

START TRAINING

Train new model

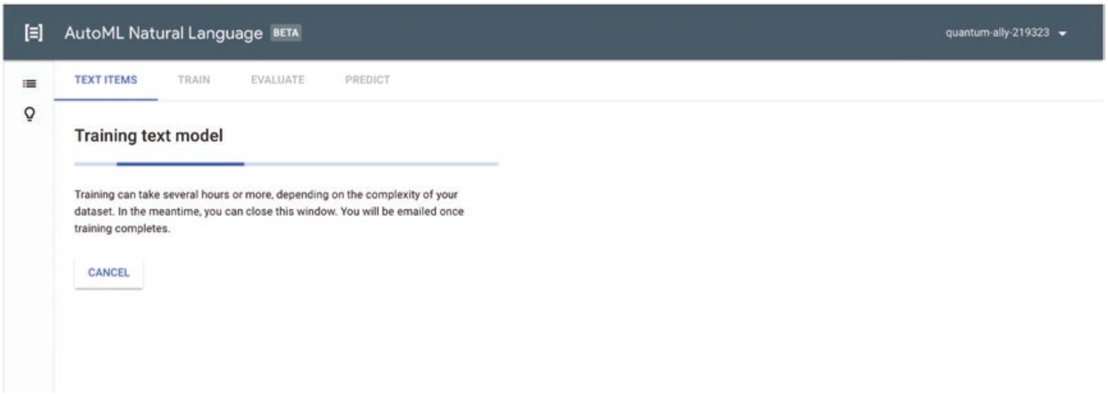
Model name
toxicity_dataset_v20181125054014

Data summary
55327 labeled text items, 2 labels

You will be emailed when training completes. Pricing guide

CANCEL START TRAINING

Figure 43-10. Accept the Model name and click on "Start Training"

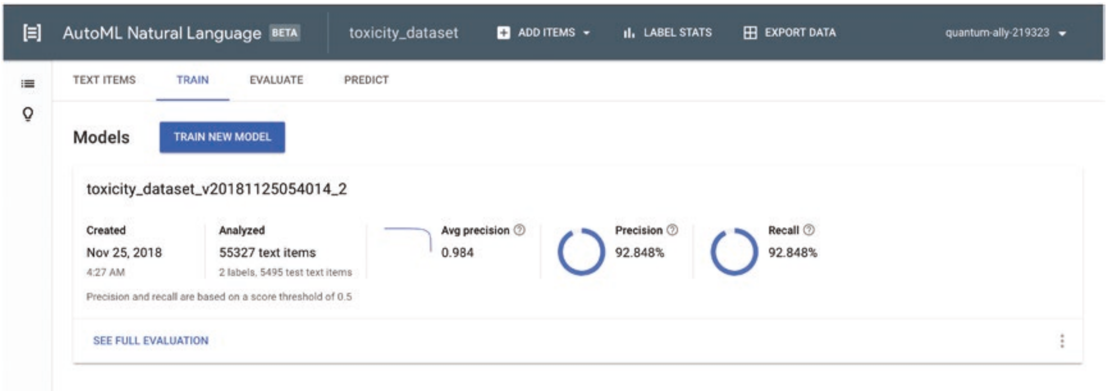


Training text model

Training can take several hours or more, depending on the complexity of your dataset. In the meantime, you can close this window. You will be emailed once training completes.

Figure 43-11. *Training the text classification model on Cloud AutoML NLP*

6. The training summary is shown in Figure 43-12. The training phase lasted for approximately 4 hours and 45 minutes.



Models

toxicity_dataset_v20181125054014_2

Created: Nov 25, 2018, 4:27 AM

Analyzed: 55327 text items, 2 labels, 5495 test text items

Avg precision: 0.984

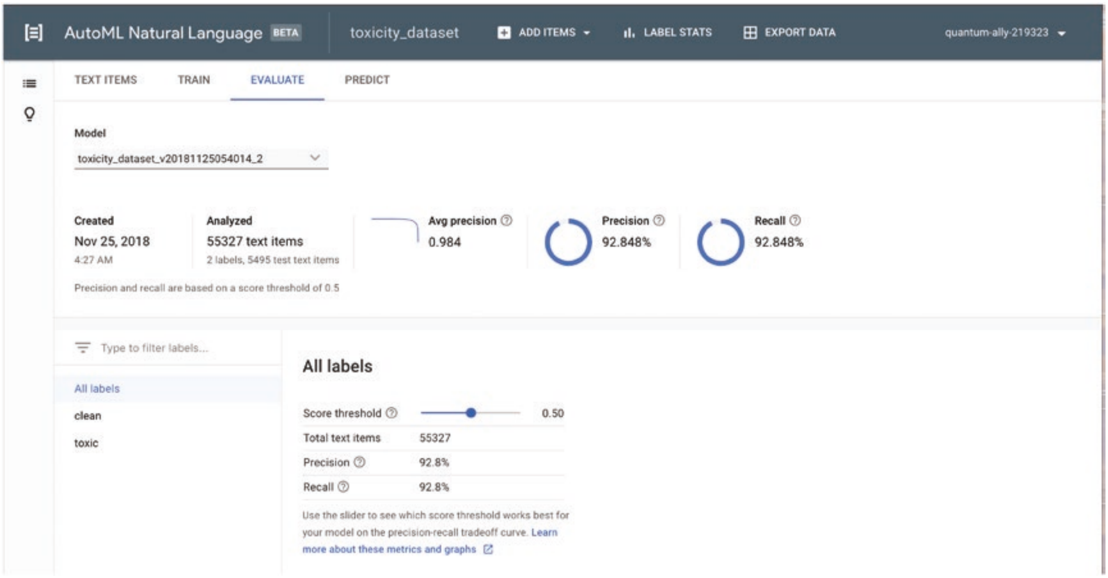
Precision: 92.848%

Recall: 92.848%

Precision and recall are based on a score threshold of 0.5

Figure 43-12. *Cloud AutoML NLP: Training summary*

7. AutoML NLP sets aside a portion of the documents as a test set in order to evaluate the quality of the model after training (see Figure 43-13). The F1 plot shows the trade-off between precision and recall. Also, a confusion matrix provides further insight into the model quality (see Figure 43-14).



Model
toxicity_dataset_v20181125054014_2

Created
Nov 25, 2018
4:27 AM

Analyzed
55327 text items
2 labels, 5495 test text items

Avg precision 0.984

Precision 92.848%

Recall 92.848%

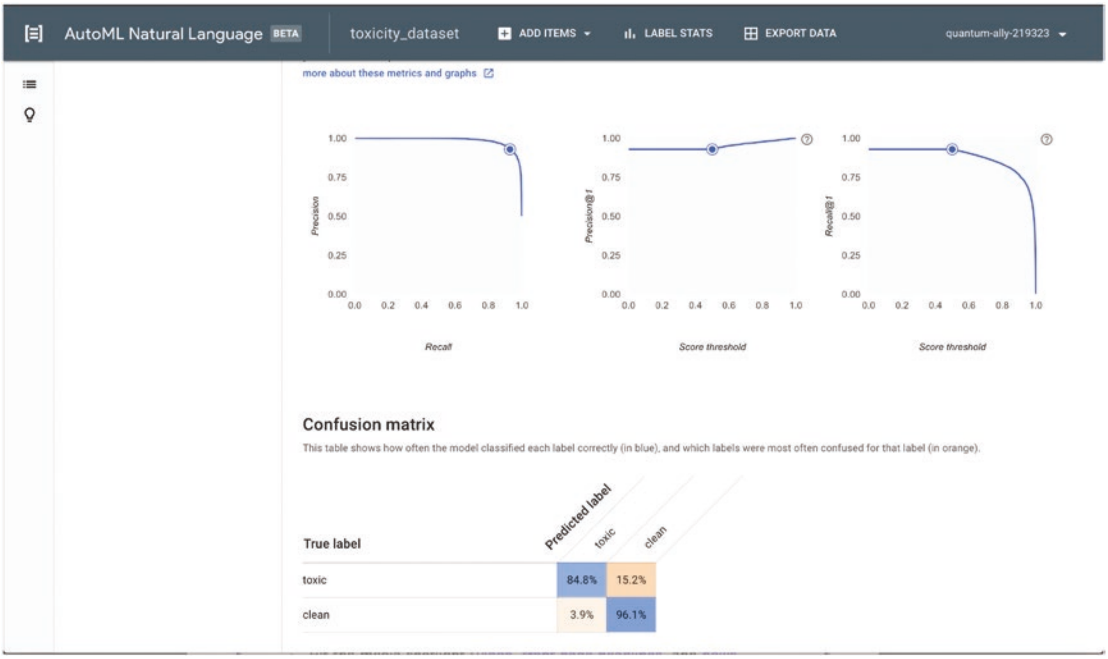
Precision and recall are based on a score threshold of 0.5

All labels

Score threshold	0.50
Total text items	55327
Precision	92.8%
Recall	92.8%

Use the slider to see which score threshold works best for your model on the precision-recall tradeoff curve. Learn more about these metrics and graphs

Figure 43-13. Cloud AutoML NLP: Model evaluation



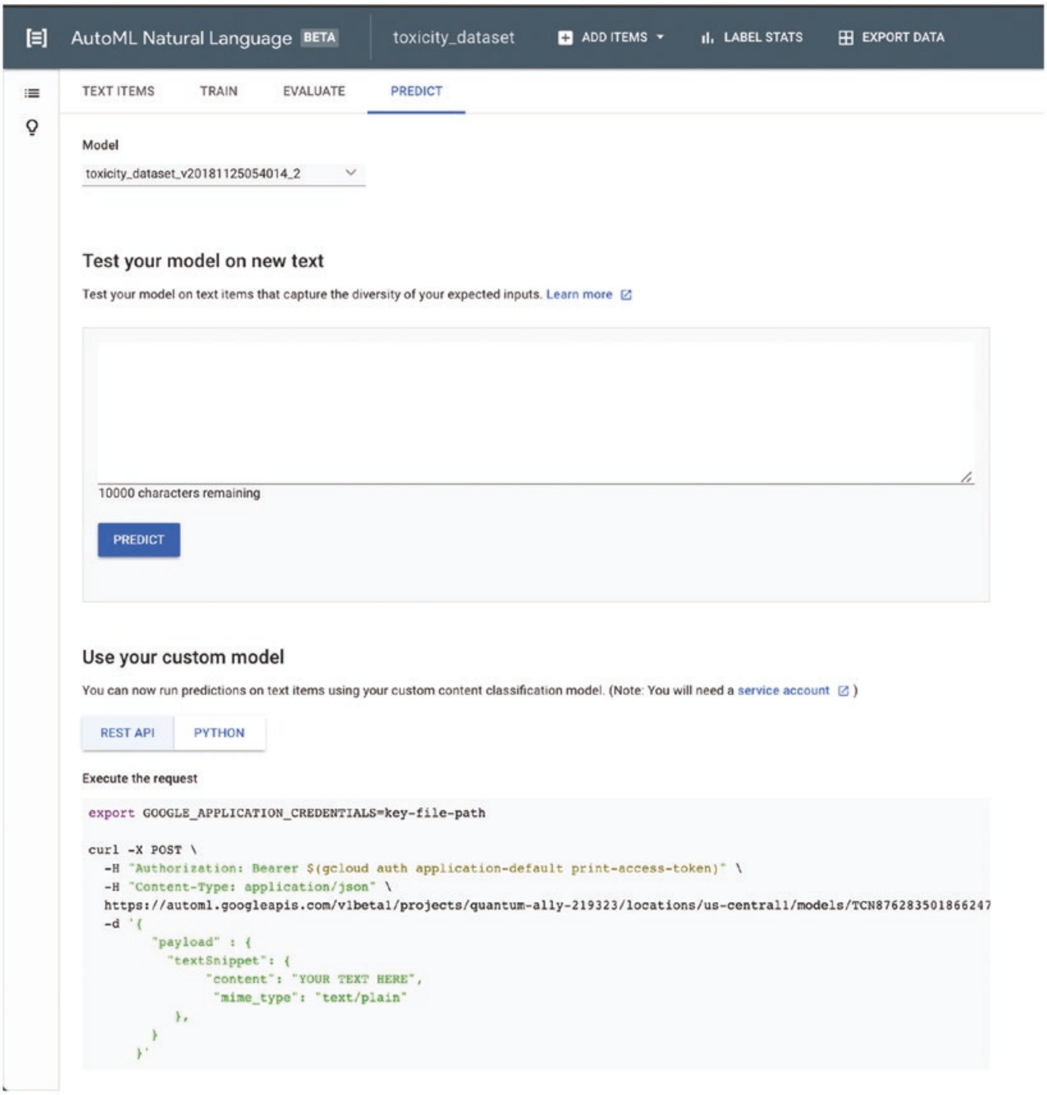
Confusion matrix

This table shows how often the model classified each label correctly (in blue), and which labels were most often confused for that label (in orange).

True label	toxic	clean
toxic	84.8%	15.2%
clean	3.9%	96.1%

Figure 43-14. F1 evaluation plot and confusion matrix on Cloud AutoML NLP

8. The custom text classification model is exposed as a REST or Python API for integration into software applications as a prediction service (see Figure 43-15). We can test our model by uploading a sample image for classification. Figure 43-16 passes a clean text example to the model and it predicts correctly with a probability of 98%, while Figure 43-17 passes a toxic text example to the model. This example is also correctly classified with a probability score of 99%.



Model

toxicity_dataset_v20181125054014_2

Test your model on new text

Test your model on text items that capture the diversity of your expected inputs. Learn more

10000 characters remaining

Use your custom model

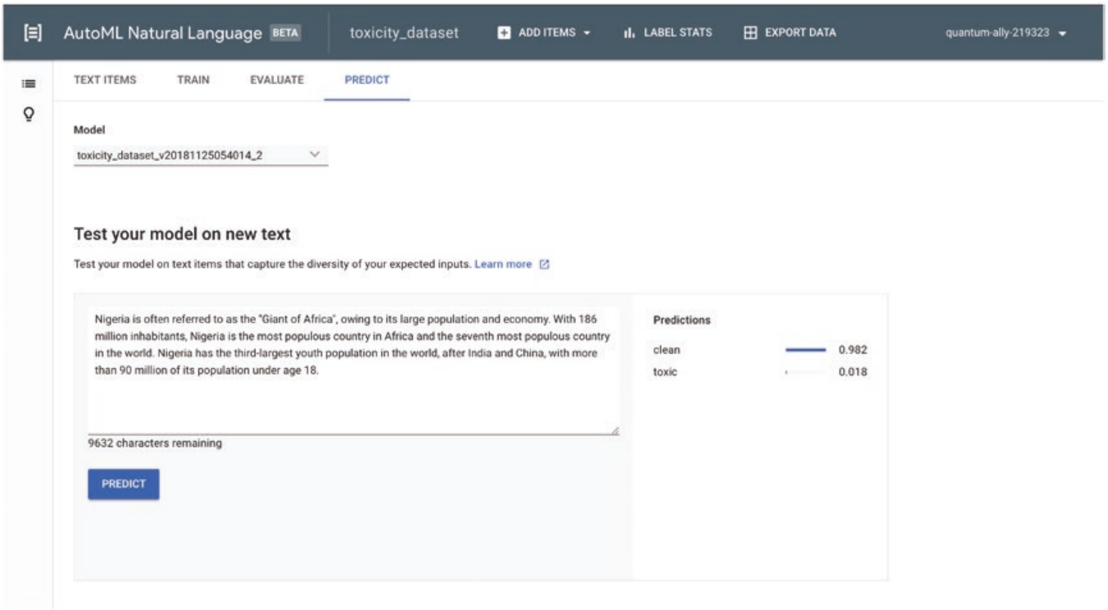
You can now run predictions on text items using your custom content classification model. (Note: You will need a service account)

Execute the request

```
export GOOGLE_APPLICATION_CREDENTIALS=key-file-path

curl -X POST \
  -H "Authorization: Bearer $(gcloud auth application-default print-access-token)" \
  -H "Content-Type: application/json" \
  https://automl.googleapis.com/v1beta1/projects/quantum-ally-219323/locations/us-central1/models/TCN876283501866247
  -d '{
        "payload" : {
          "textSnippet": {
               "content": "YOUR TEXT HERE",
                "mime_type": "text/plain"
            },
         }
      }'
```

Figure 43-15. *Cloud AutoML NLP model as a prediction service*



Model

toxicity_dataset_v20181125054014_2

Test your model on new text

Test your model on text items that capture the diversity of your expected inputs. Learn more

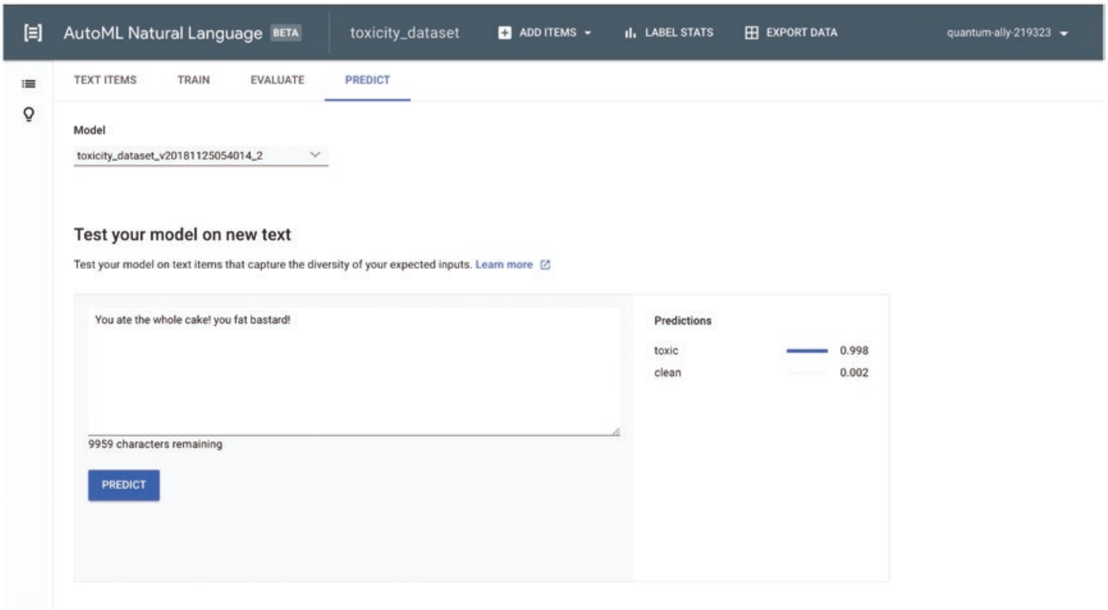
Nigeria is often referred to as the "Giant of Africa", owing to its large population and economy. With 186 million inhabitants, Nigeria is the most populous country in Africa and the seventh most populous country in the world. Nigeria has the third-largest youth population in the world, after India and China, with more than 90 million of its population under age 18.

9632 characters remaining

Predictions

clean	0.982
toxic	0.018

Figure 43-16. Clean words example: AutoML NLP



Model

toxicity_dataset_v20181125054014_2

Test your model on new text

Test your model on text items that capture the diversity of your expected inputs. Learn more

You ate the whole cake! you fat bastard!

9959 characters remaining

Predictions

toxic	0.998
clean	0.002

Figure 43-17. Toxic words example: AutoML NLP

This chapter covered building and deploying custom text classification models using Google AutoML Cloud Vision. In the next chapter, we will build an end-to-end data science product on GCP.

CHAPTER 44

Model to Predict the Critical Temperature of Superconductors

This chapter builds a regression machine learning model to predict the critical temperature of superconductors. The features for this dataset were derived based on the following superconductor properties:

- Atomic mass
- First ionization energy
- Atomic radius
- Density
- Electron affinity
- Fusion heat
- Thermal conductivity
- Valence

And for each property, the mean, weighted mean, geometric mean, weighted geometric mean, entropy, weighted entropy, range, weighted range, standard deviation, and weighted standard deviation are extracted. Thus, this results in a total number of 8 x 10 = 80 features. In addition to this, a feature that contains the number of elements in the superconductor is added to the design matrix. The predictor variable is the critical temperature of the superconductor. Hence, the dataset has a total of 81 features and 21,263 rows.

E. Bisong, *Building Machine Learning and Deep Learning Models on Google Cloud Platform*,
https://doi.org/10.1007/978-1-4842-4470-8_44