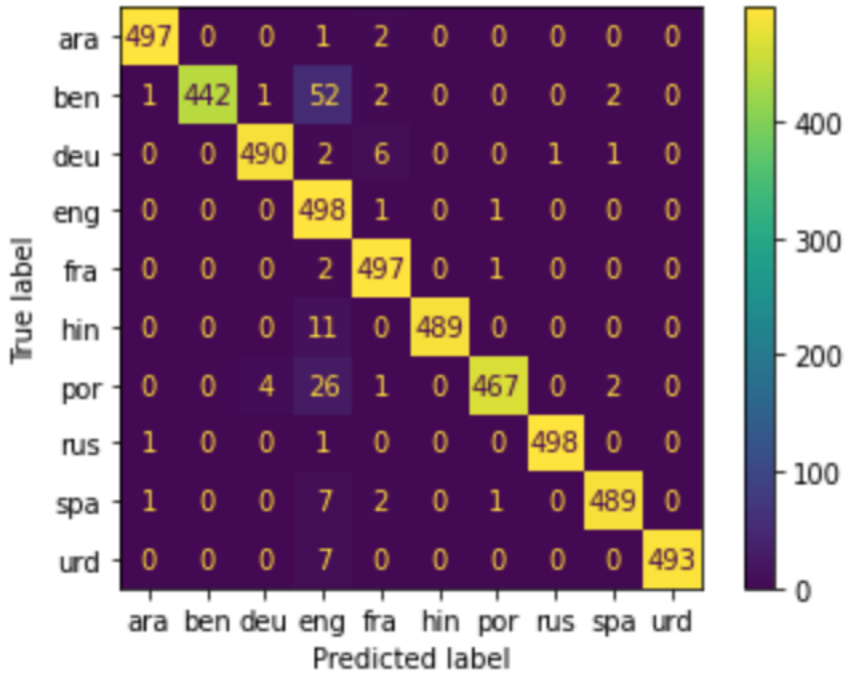
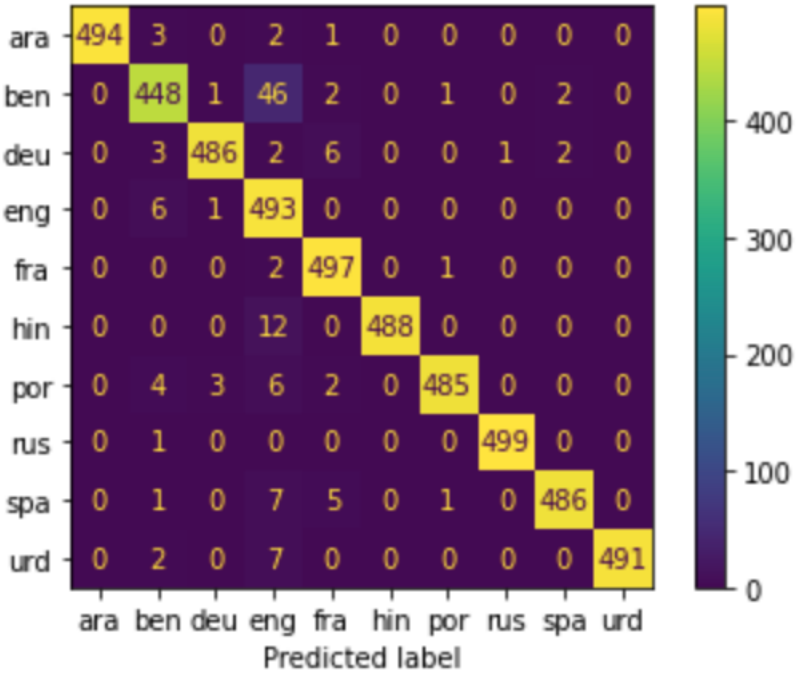
MAIS 202 Final Project deliverable 3

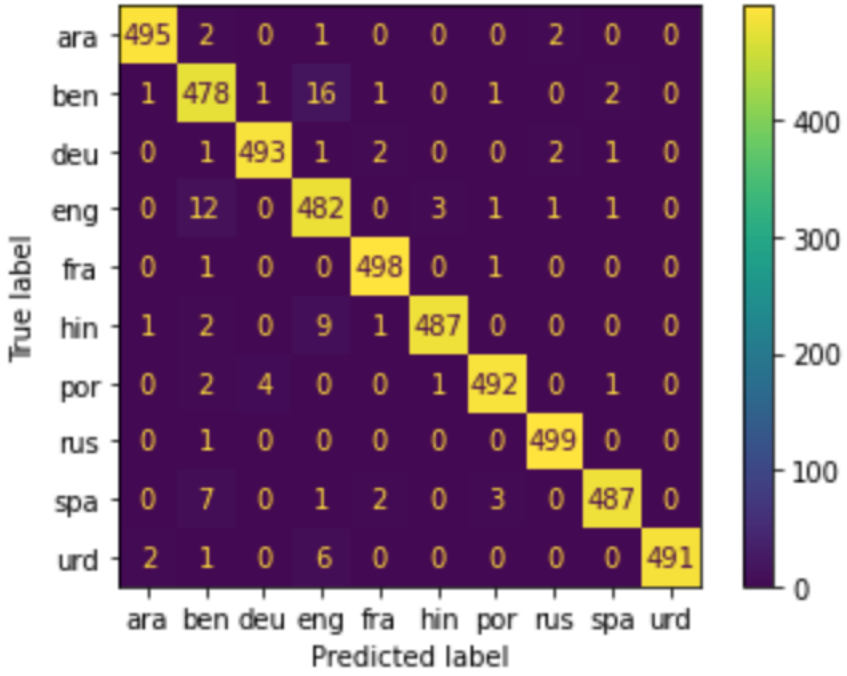
Tahsin Kabir, Omar Lahlou, Abderrahman Laoufi

We have successfully implemented our model, however we had to greatly reduce the data for it to run. Our model is too computationally expensive to vectorize the entirety of the 500 sentences per 250 languages. While we tried to circumvent this by using Scikit's vectorizer function, it resulted in our model being woefully inaccurate, with an accuracy of around 4%. We thus stuck to our original vectorizer, but reduced the number of languages, which allowed the model to successfully train, and have a testing accuracy of around 97% for the three models we used, which are Naïve Bayes, random forests, and support vector machines.

The accuracy of our models can be seen in the following confusion matrices, representing Naïve Bayes, random forests, and support vector machines, respectively.






We will develop a web application in which the user will input a short text and it will return the language that it is most likely to be. If it is not one of the languages that the data is trained on due to limited computational power, the model will say that it is not available. It will dynamically display the results, remaining on the same page. We currently do not have any

experience developing web apps, and because of such will reach out to older students as well as online tutorials for help.