Outline of presentation (based on the current solution)

**I. Preprocessing**

- Introduce the dataset & our task

- Read the files & remove any non-alphabet characters

- Create data frame with the authors’ names as labels

**II. Classification**

- Vectorize the documents with “stop words”

- Use LinearSVM as the best solution

- Our achievement

**III. Challenges**

- No access to HPC

- Surprising poor performance of KNN, Logistic Regression, etc.

- Tuning parameters of models

**VI. Conclusion**

- Professional news reporters have similar professional wording style

- The sea of text classification approaches and combinations of methods are worth our deeper diving

- CNN, RNN could be next tries