MAIS 202 Final Project- Deliverable 1

Tahsin Kabir, Omar Lahlou, Abderrahman Laoufi

Our project will consist of a language detection program that will detect the language of a short text. The scope of the project is limited to the 235 languages found on Wikipedia. We will use the WiLi-2018 data set, as it contains 235000 paragraphs in 235 languages from Wikipedia, and is already cleaned and separated into training and test sets. The size of the data set, as well as its convenience, were the main reasons for choosing it.

The data is already processed, as it contains cleaned paragraphs for each language. It is additionally separated into training and data sets, allowing us to begin working directly on the model.

We will use a confusion matrix and accuracy loss as this is a supervised learning, classification problem. We aim to have above 90% accuracy.

We will employ a Multinomial Naive Bayes model as it is well-suited to multiclass classification problems, as well as large data sets, data with multiple labels, and for natural language processing problems (Kharwal 2021).

The user would input any text into a user-friendly interface consisting of a text box, and the return would simply be a display consisting of the language detected.

Works cited:

Kharwal, Aman. "Multinomial Naive Bayes in Machine Learning." *Data Science | Machine Learning | Python | C++ | Coding | Programming | JavaScript*, 6 Aug. 2021, thecleverprogrammer.com/2021/08/06/multinomial-naive-bayes-in-machine-learning/. Accessed 10 Oct. 2022.