Briefly describe what method you used (minimum one paragraph) (This might appear in the overview paper)

The method involves character n-gram TF-IDF vectorization as a representation scheme, then applying several machine learning algorithms such as SVM, SGD, KNN, Naïve Bayes and multi-Layer Perceptron classifiers. To enhance the representation scheme, word length has been added to the features. A wide range of character n-gram has been tested and the best performed range, uni-gram to 4-gram, has been used. In addition, SVM outperforms all other classifiers.

What is the most interesting aspect of your system (i.e., if we have to briefly describe your system in the overview paper, what would you want us to mention)?

By augmenting the dataset with the "word length" feature, we capture additional information about the words beyond their textual content. This feature provides valuable insights into the relationship between word length and language classification. Moreover, the utilization of character n-gram TF-IDF vectorization allows us to effectively handle the impact of frequently occurring characters, enhancing the performance and interpretability of the model. The combination of these techniques enables a comprehensive analysis of the dataset, leading to improved accuracy and a better understanding of the underlying patterns in language identification.