Kaggle Competition:

Detecting the difficulty level of french texts

Presented by Katelyn Bonner and Ahmed Farhat

Project Overview

When learning a new language it is important to develop reading skills… However, it can be difficult locating texts that match one's reading level.

The aim of this project is to create a model that can predict the difficulty level of a given french text in order to recommend an appropriate text. The levels are based on the Common European Framework of Reference for Languages (CEFR) namely A1, A2, B1, B2, C1, and C2.

Our problem can therefore be labeled as a classification problem with 6 labels.

Our Approach - Basic Models

The project provided the following list of basic models (meaning no data cleaning, stop word removal, etc.) with Tf-idf vectorization and hyperparameter tuning to build and evaluate: logistic regression, K-nearest neighbors, Decision Tree, and Random Forest.

We obtained the following results:

	Logistic Regression	KNN	Decision Tree	Random Forest
Precision	0.4865	0.4037	0.3464	0.3886
Recall	0.4854	0.3990	0.3396	0.3969
F1-Score	0.4823	0.3767	0.3305	0.3723
Accuracy	0.4854	0.3990	0.3396	0.3969

Our Approach - Data Augmentation

Next, we decided to include data augmentation in order to add features that may indicate the difficulty of the text and improve accuracy. We included:

- entity recognition and count
- part of speech recognition and count
- number of words per sentence
- average word length
- number of stop words per sentence
- share of stop words per sentence

Additionally, we removed stop words and punctuation and convert all words to lowercase. We also tried lemmetization but found it lowered accuracy. We obtained the following results:

Precision: 0.4215 **Recall:** 0.3000 **F1-Score:** 0.2770 **Accuracy:** 0.3000

Our Approach - Pre-trained Model

Finally, we tried using a pre-trained model found on hugging face: CamemBERT model

This model is the french version of the BERT model. We use the sequence classification head and camemBERT tokenizer.

We obtained the following results:

Precision: 0.5985 **Recall:** 0.5948 **F1-Score:** 0.5952 **Accuracy:** 0.5948

Future Recommendations

- Add data augmentation and word embeddings to the CamemBERT model
- Train the model on additional resources of french texts to improve accuracy