Predicting your English level

Rickard Ramhöj

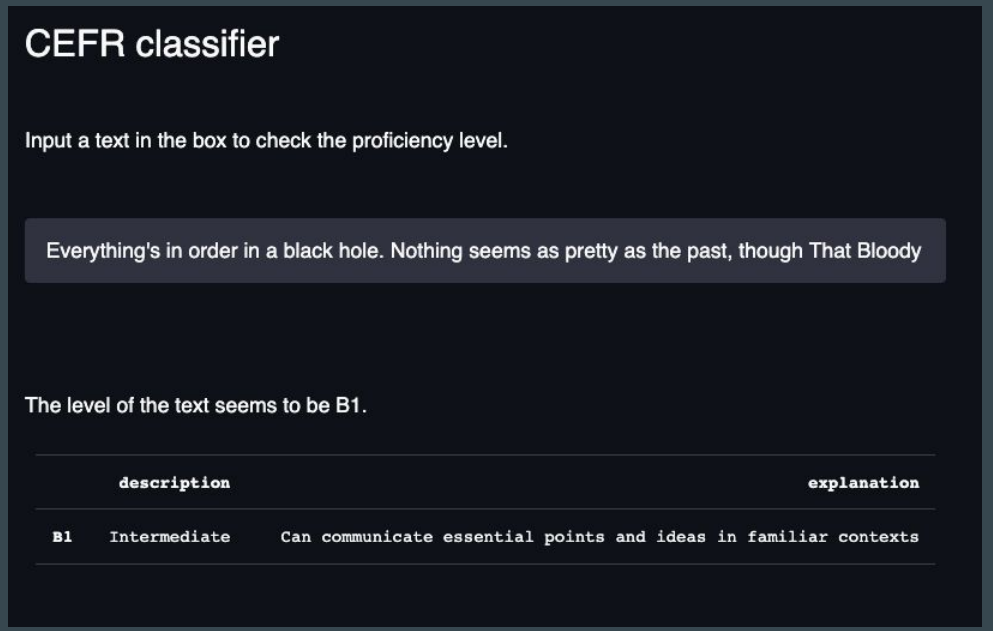
Overview

1. Product idea (CEFR Classifier)
2. Introducing dataset
3. Feature engineering
4. Descriptive and inferential stats
5. Classification modelling (multiclass logistic regression)
6. Create product and next steps

1. Product idea

- Enter a text
- See the proficiency level of the writer

Web interface



2. Dataset: CEFR levelled English texts

Source

https://www.kaggle.com/amontgomerie/cefr-levelled-english-texts

Structure

text	label
bla bla bla bla bla bla [...]	A1
blaaaa blaa blaaa bla [...]	B2
bla bla bla [...]	A2
bla bla blassss bla blaaaa bla [...]	C1

3. Feature engineering

- Grammatical features (parts of speech)
  - 'ADV', 'NOUN', 'ADP', 'VERB', 'PRON', 'ADJ', 'PUNCT','PROPN', 'DET', 'PART', 'CCONJ', 'SPACE', 'AUX', 'NUM', 'SCONJ', 'INTJ', 'SYM', 'X' (using spaCy)
- Lexical features (TF-IDF)
  - TF = term frequency = how frequent is the term in a document
  - IDF = inverse document frequency = how common or rare is the term across all documents?

4. Descriptive and inferential stats

• Balanced dataset?	YES
• Outliers?	YES (227 outliers with z score > 3)
• Normality?	NO (H0 rejected for 17/18 columns)
• Homogeneity of variances?	NO (H0 rejected for 15/18 columns)

4. Descriptive and inferential stats

PARAMETRIC

• ANOVA	H0 rejected for all variables
• T-test (mean)	H0 rejected for 48/85 pairs
• Confidence intervals	E.g. interval for mean % of PRON for A1

NON-PARAMETRIC

• Mann-Whitney U rank test (median)	H0 rejected for 62/85 pairs
• Spearman correlation	Used to remove variables for classifier

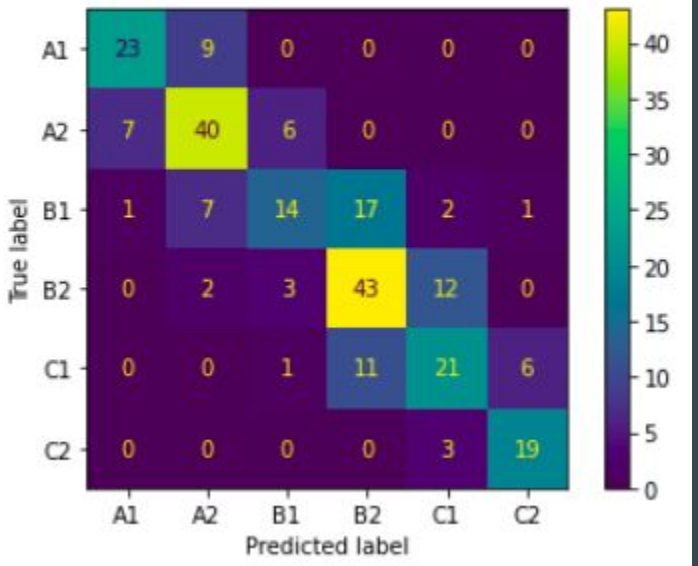
5. Classification modelling

Algorithm = multinomial logistic regression

- Independent variables (X):
  - 'NOUN', 'ADP', 'VERB', 'PRON', 'ADJ', 'PUNCT', 'PROPN', 'DET', 'CCONJ', 'AUX', 'SCONJ', 'INTJ' (scaled)
  - TF-IDF
- Dependent variable (y):
  - proficiency levels (A1, A2, B1, B2, C1, C2)

5. Evaluate model

```
Precision: 0.640166300188252
Recall: 0.6451612903225806
F1 score: 0.63492002539497
```



6. Create product and next steps

Steps to make into a product

- saving pipeline using pickle
- function to apply pipeline to new texts
- create web application to apply classifier to user-input text

Next steps

- increase accuracy
- enable folder upload
- enhance user interface
- add more features to interface (such as displaying all nouns in the text etc.)

Questions?