ESL TEXT CLASSIFIER

HOW TO FIND THE READING THAT'S RIGHT FOR YOU

“TO LEARN A LANGUAGE, YOU NEED A DREAM”

BY ME

- DATA
- REPETITION
- EXPERIENCE
- ACTION
- MOTIVATION

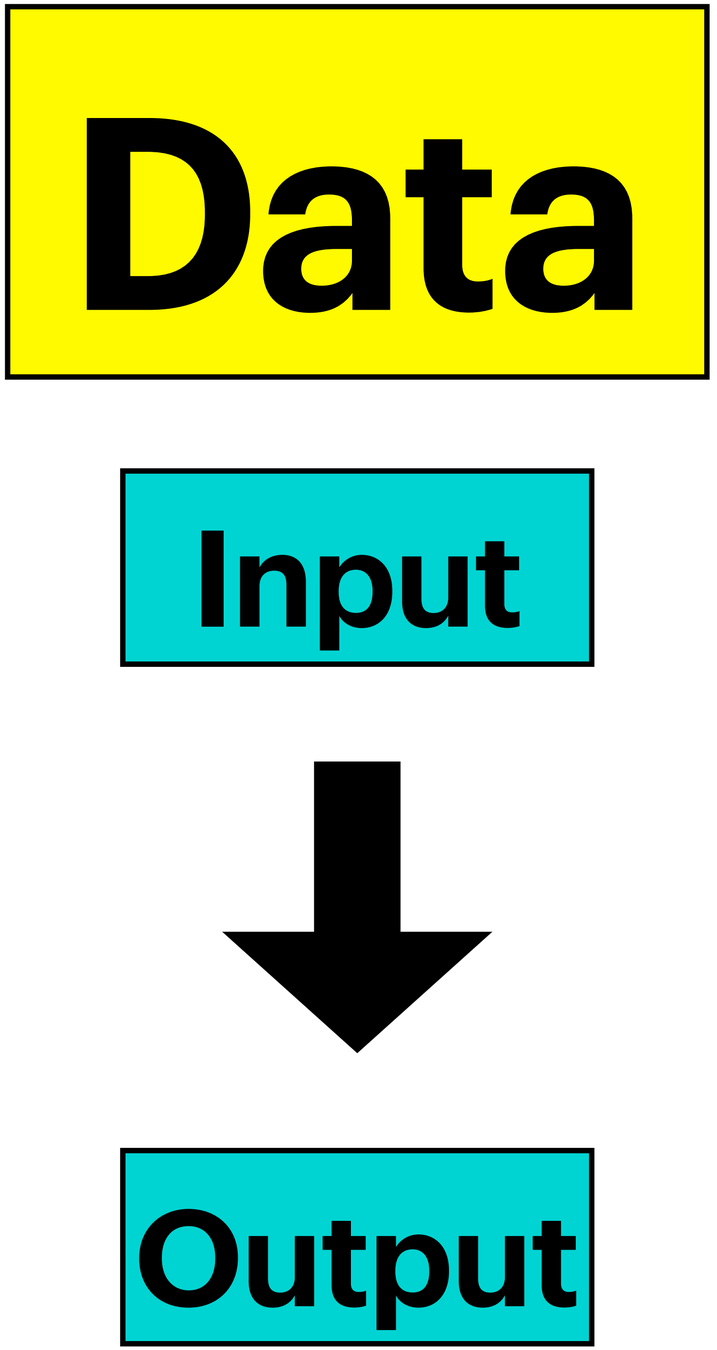
DATA

REPETITION

EXPERIENCE

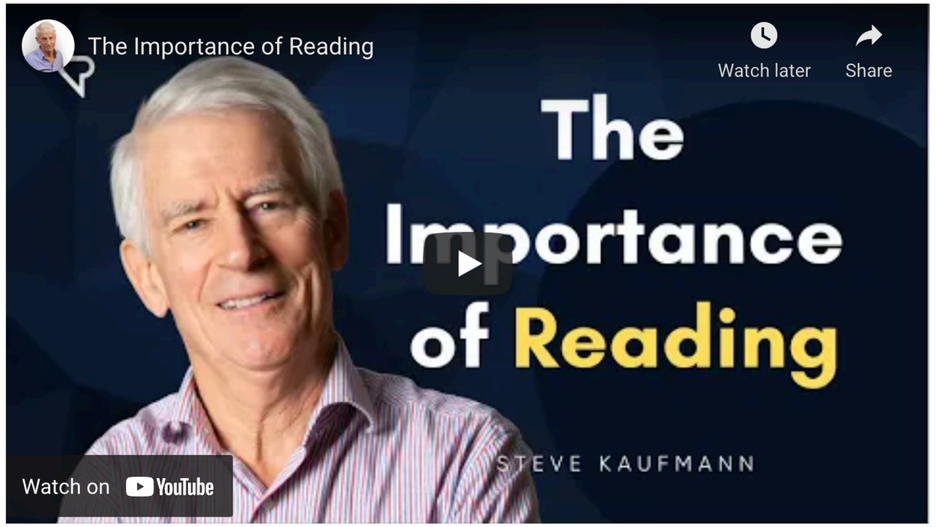
ACTION

MOTIVATION



The Linguist | Steve Kaufmann

The Importance of Reading



Data

Reading

Listening

LEVEL APPROPRIATE READING

LEVEL APPROPRIATE READING

Storytelling

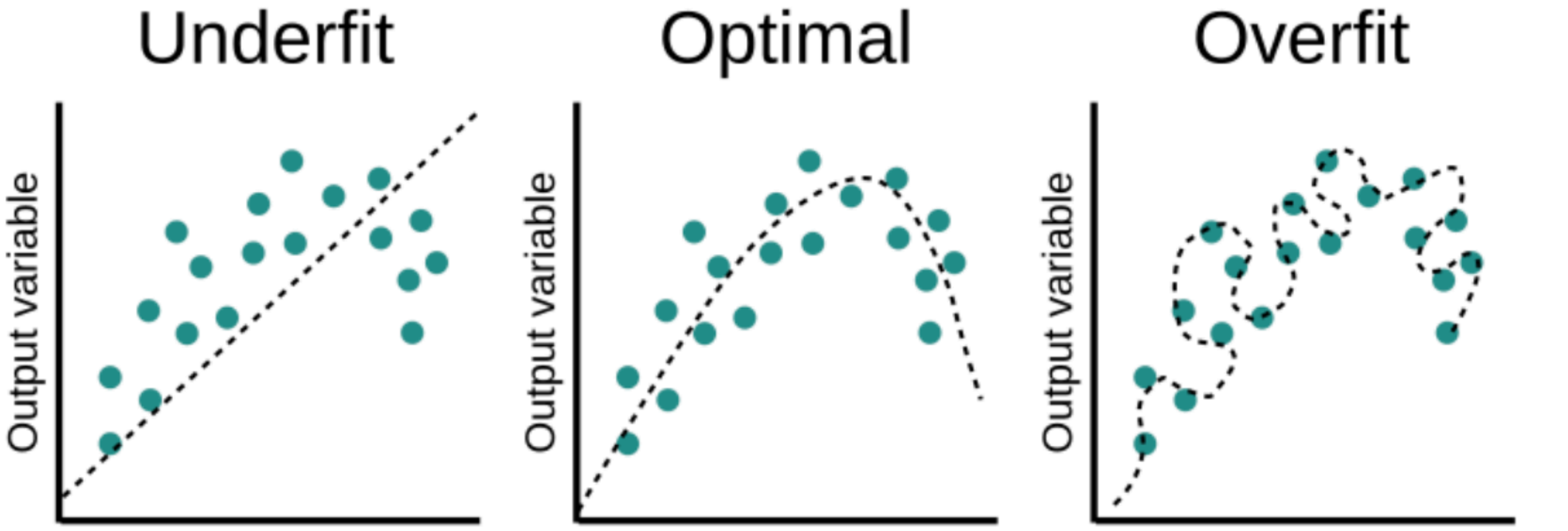
Context

Data

Grammar

Vocabulary

ROBUST READING TEXT CLASSIFIER



I hardly understand anything!

It’s just right

I understand everything!

Pandas

Data Acquisition

Natural Language
Analyses with NLTK

Data Cleaning

K Keras

Simple. Flexible. Powerful.

Modeling

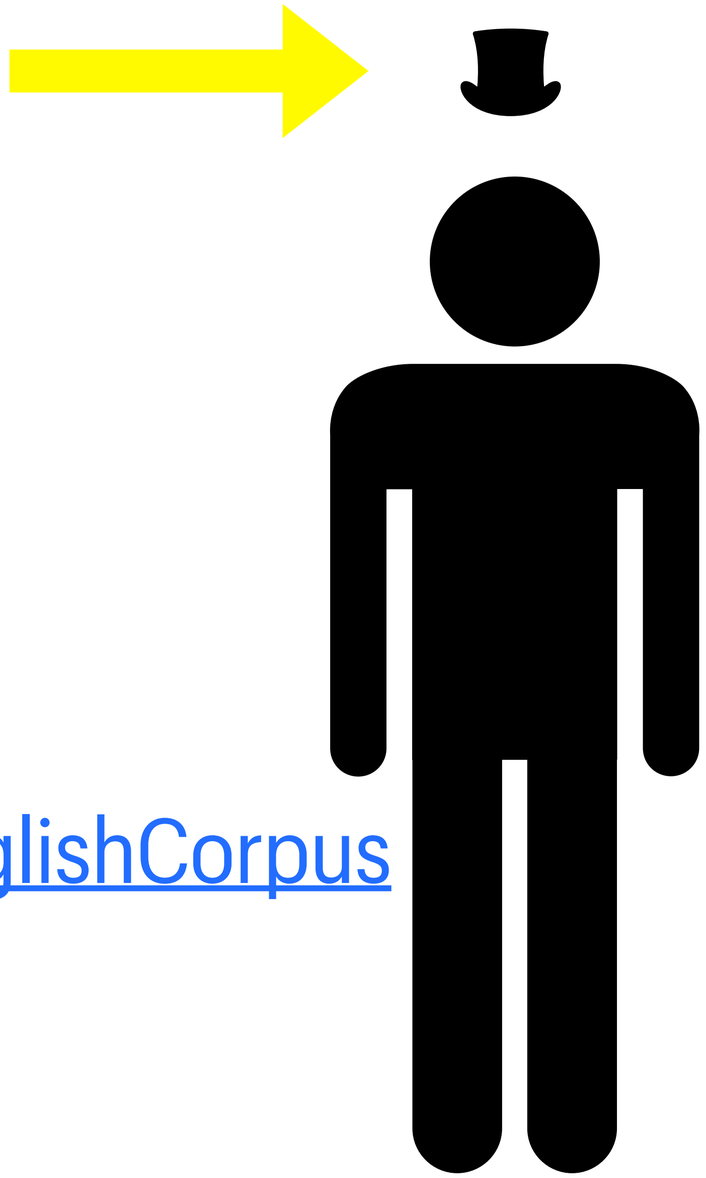
Data Sets:

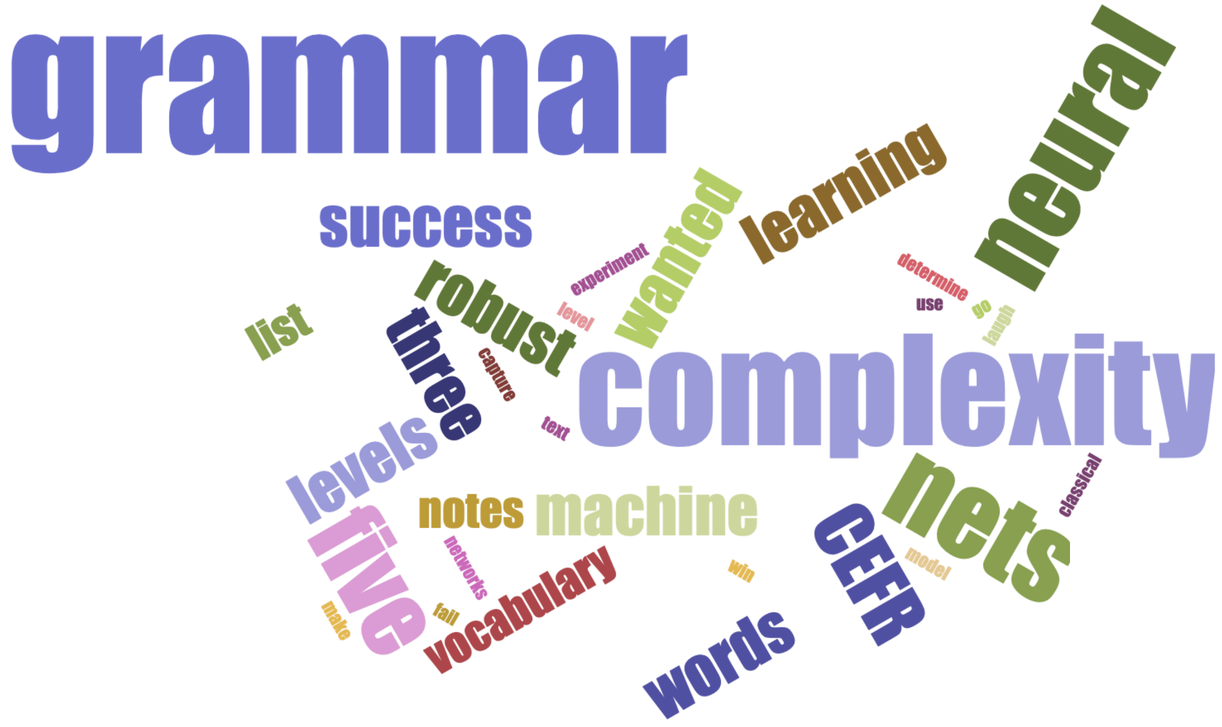
- Cambridge Readability Data Set

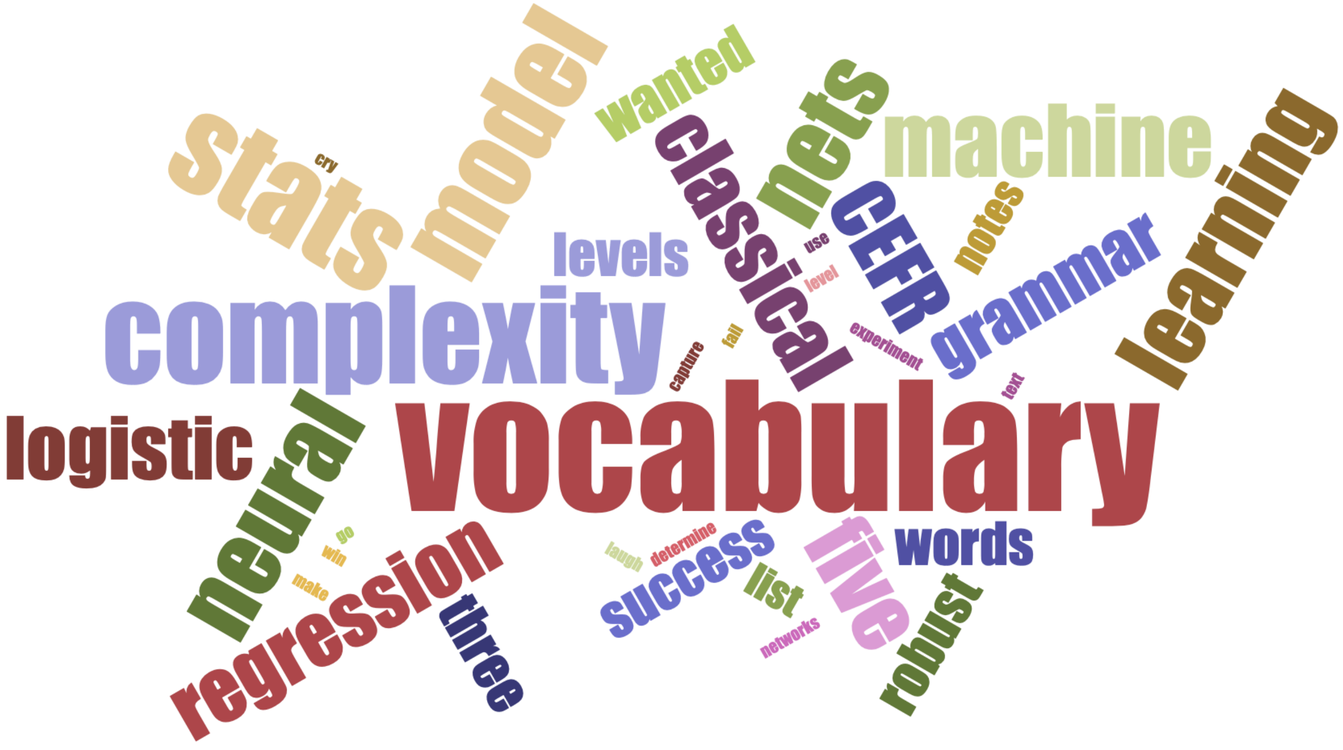
https://ilexir.co.uk/datasets/index.html

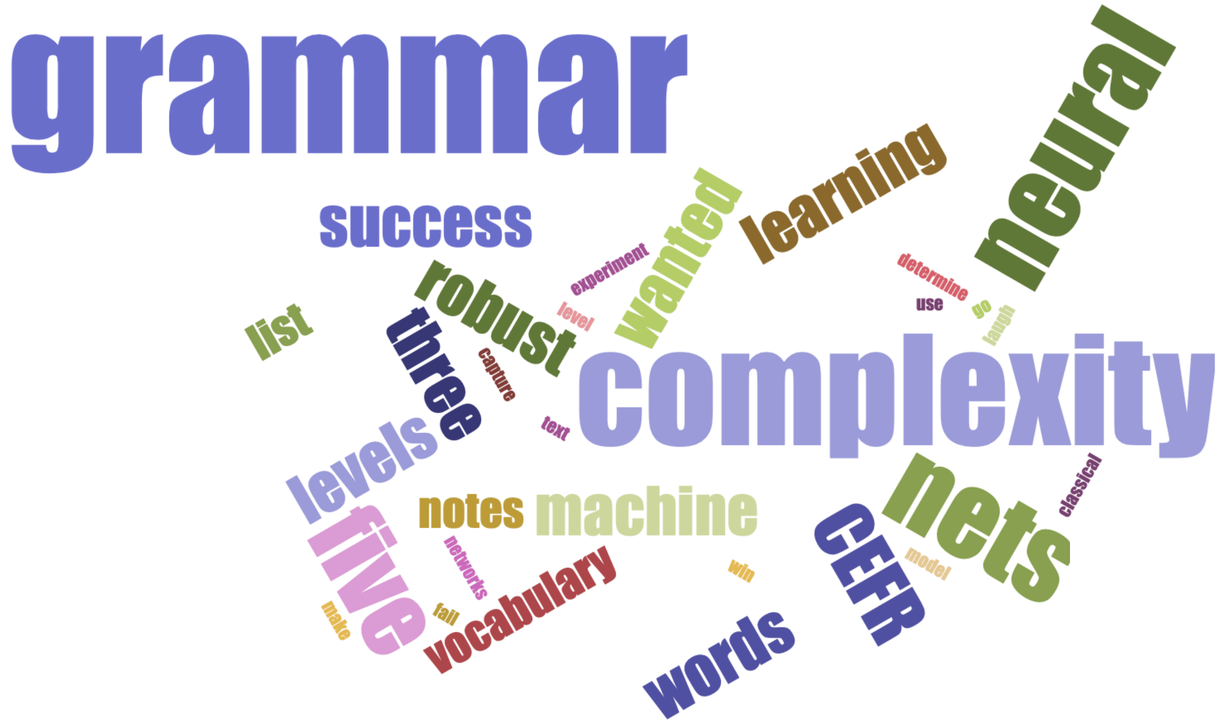
- One Stop English Corpus

https://github.com/nishkalavallabhi/OneStopEnglishCorpus

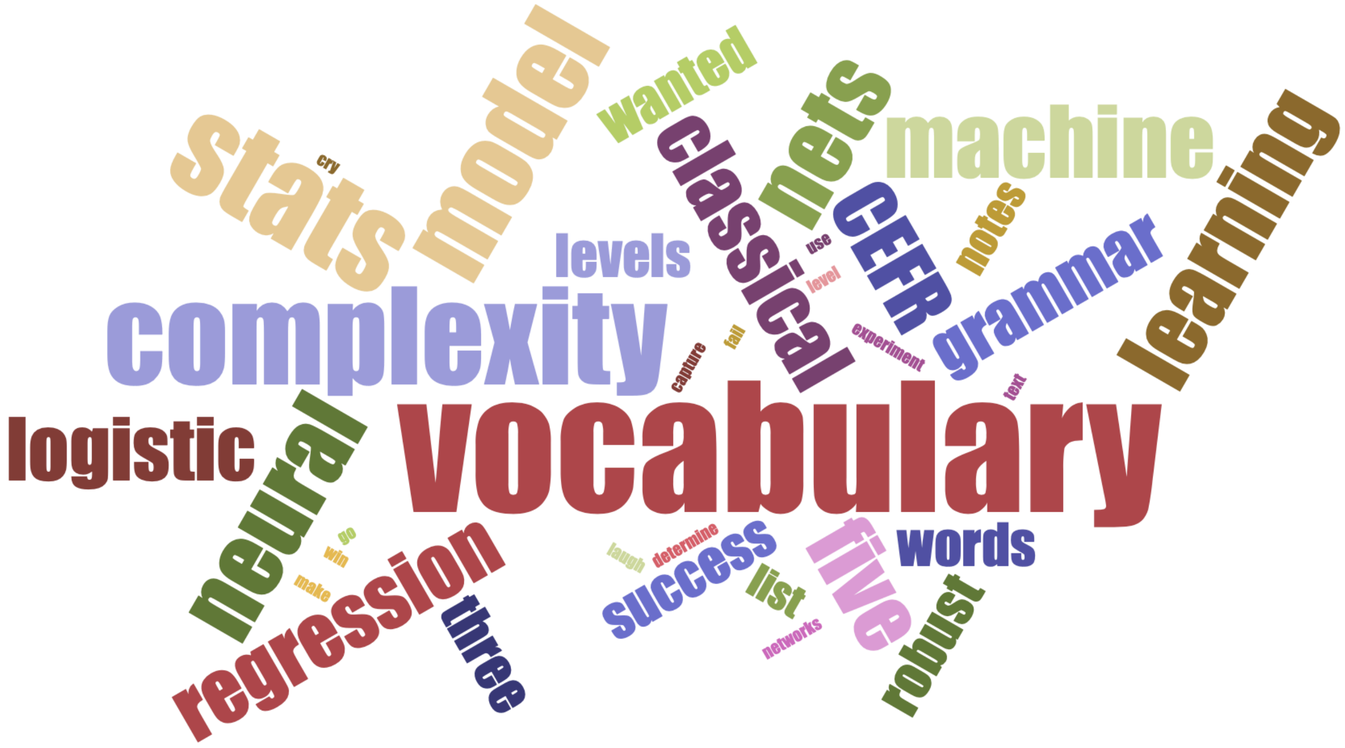



grammar
success
learning
neural
wanted
robust
experiment
determine
use
go
laugh
list
level
three
capture
complexity
text
levels
nets
classical
notes
machine
CEFR
networks
win
model
five
make
fail
vocabulary
words


stats
model
wanted
cry
classical
nets
machine
learning
levels
use
CEFR
notes
level
grammar
complexity
experiment
capture
fail
text
logistic
vocabulary
neural
go
win
laugh
determine
words
success
list
five
regression
make
three
networks
robust

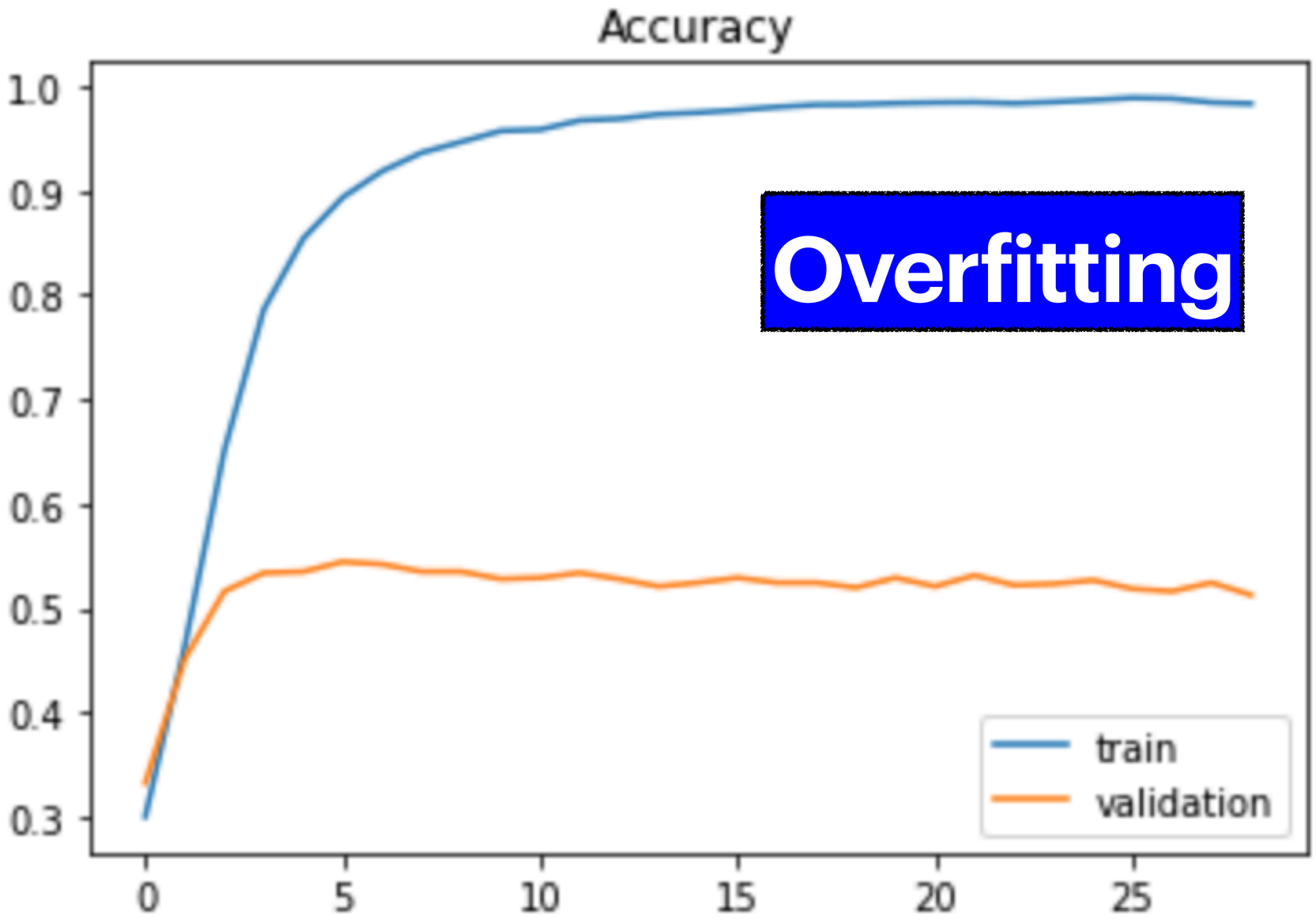


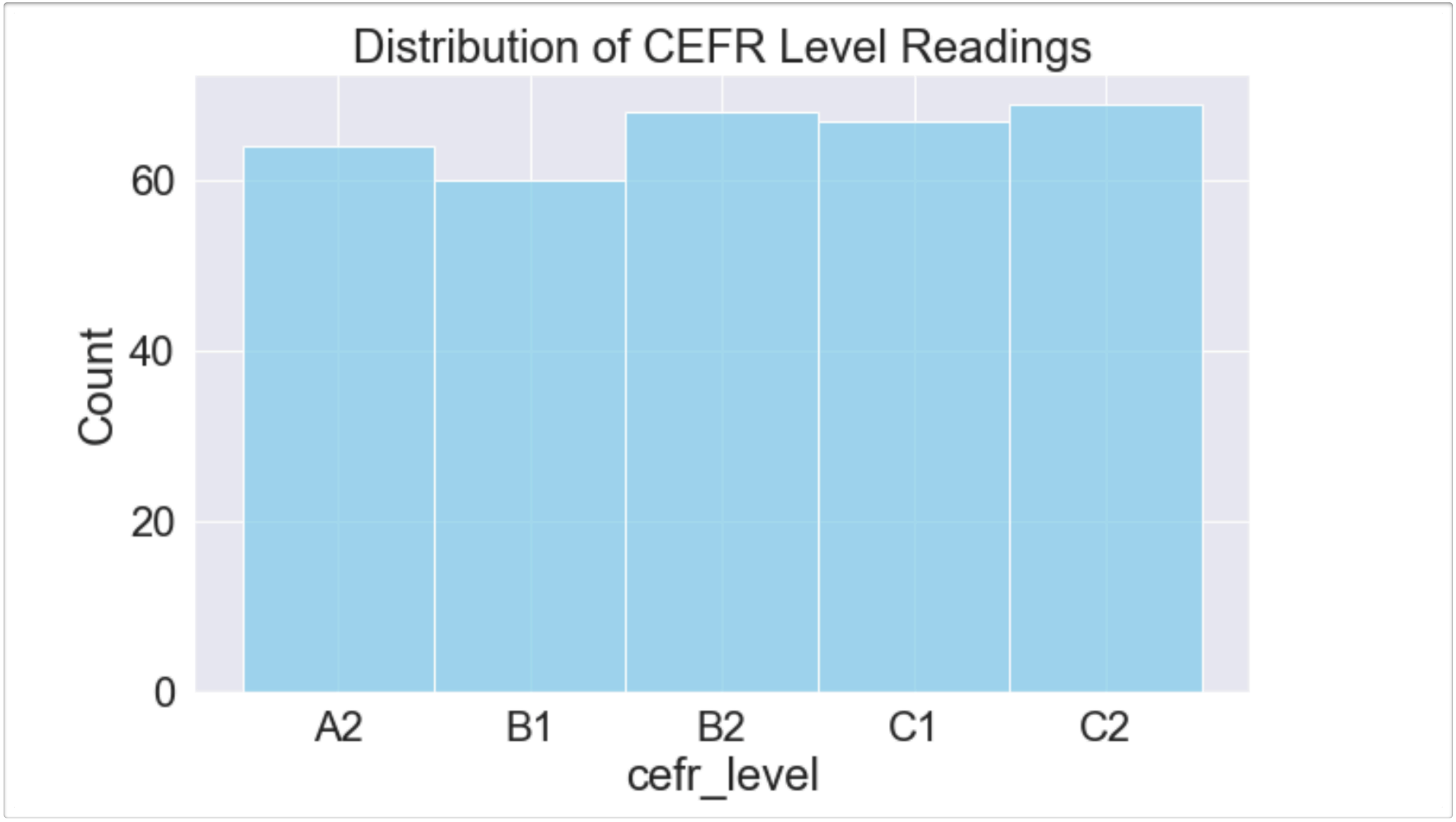
Classical Machine Learning with statistical vectorization

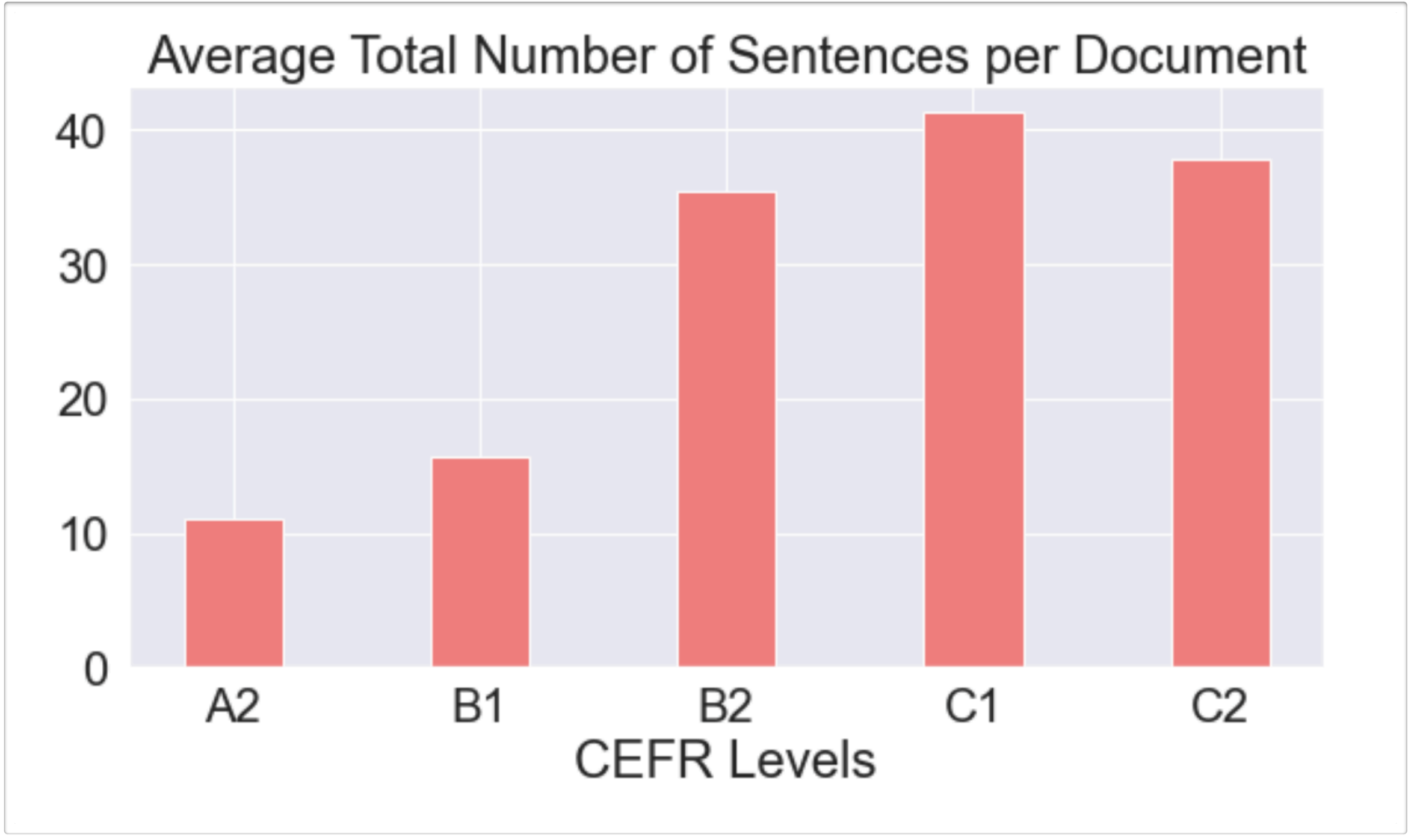


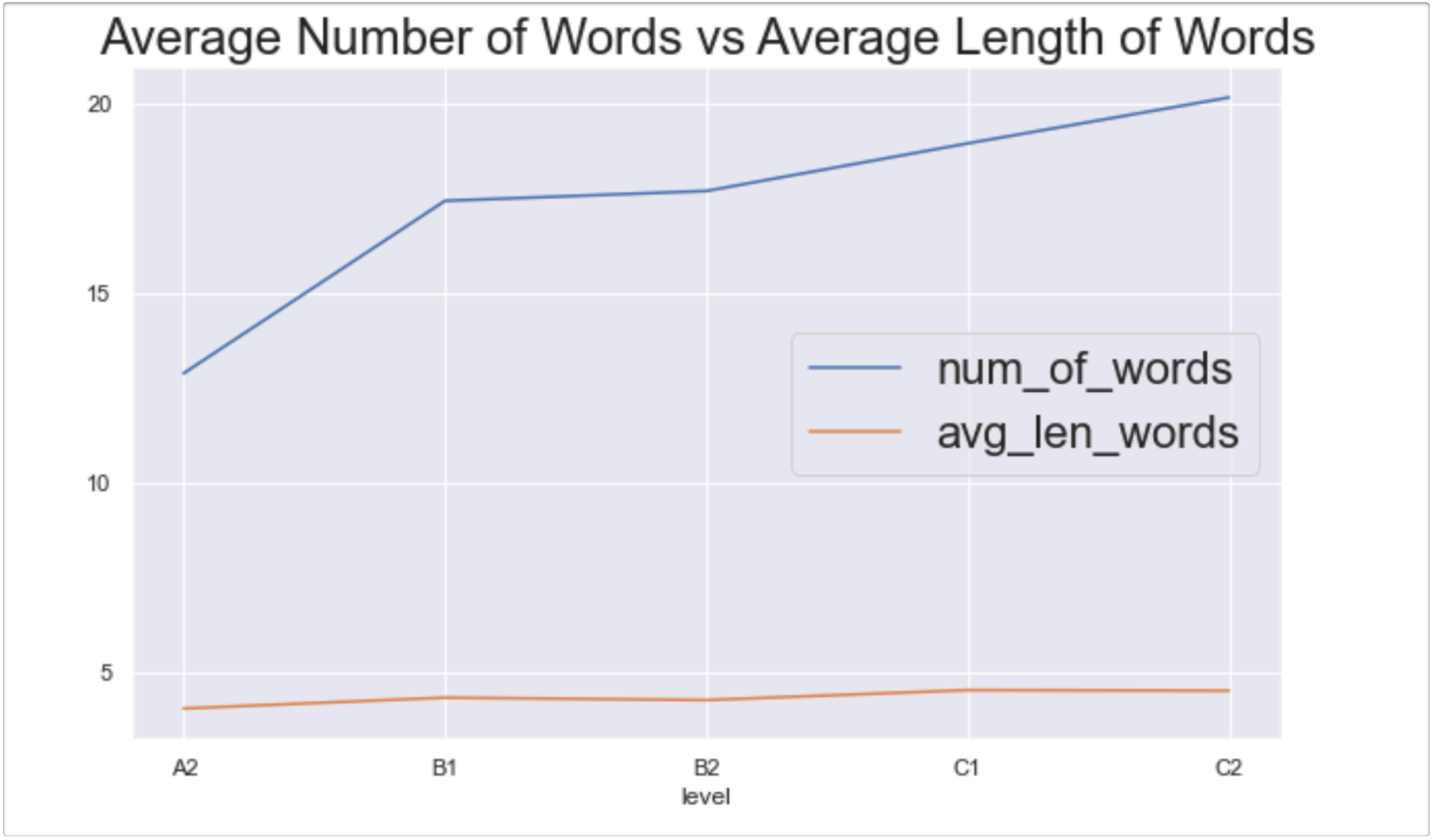
Neural Networks with tokenized vectorization

LSTM Model with texts broken up piece-wise by sentence




Distribution of CEFR Level Readings
Count
0
20
40
60
A2
B1
B2
C1
C2
cefr_level

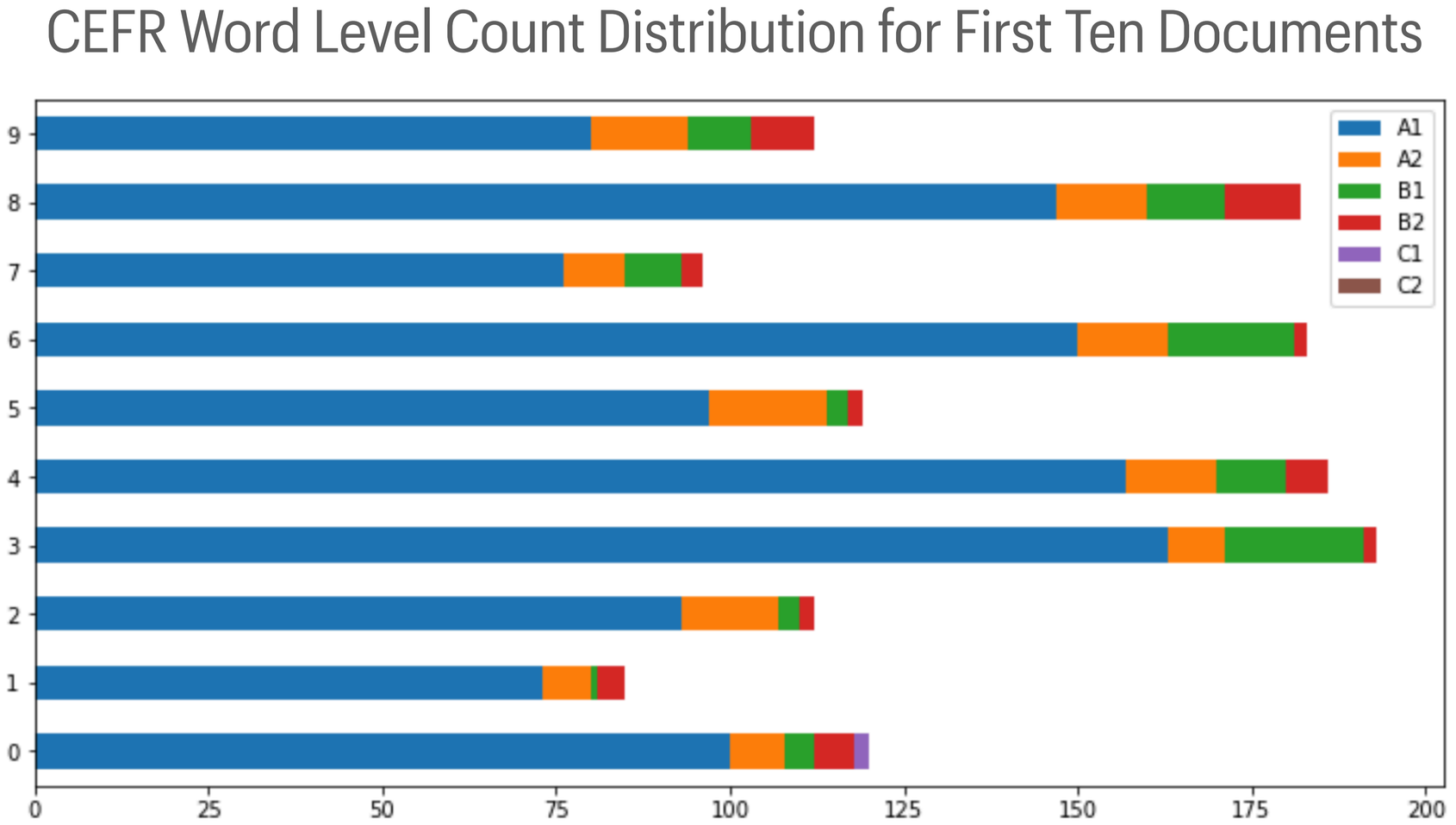

Average Total Number of Sentences per Document
40
30
20
10
0
A2
B1
B2
C1
C2
CEFR Levels


Average Number of Words vs Average Length of Words
20
15
10
5
num_of_words
avg_len_words
A2
B1
B2
C1
C2
level

Two Other Features

CEFR WORD LEVEL COUNTS

PARTS OF SPEECH COUNTS



Models: **Logistic Regression: One versus All /Gradient Boasting Classifier**

Training Accuracy: **0.93**

Validation Accuracy: **0.79**

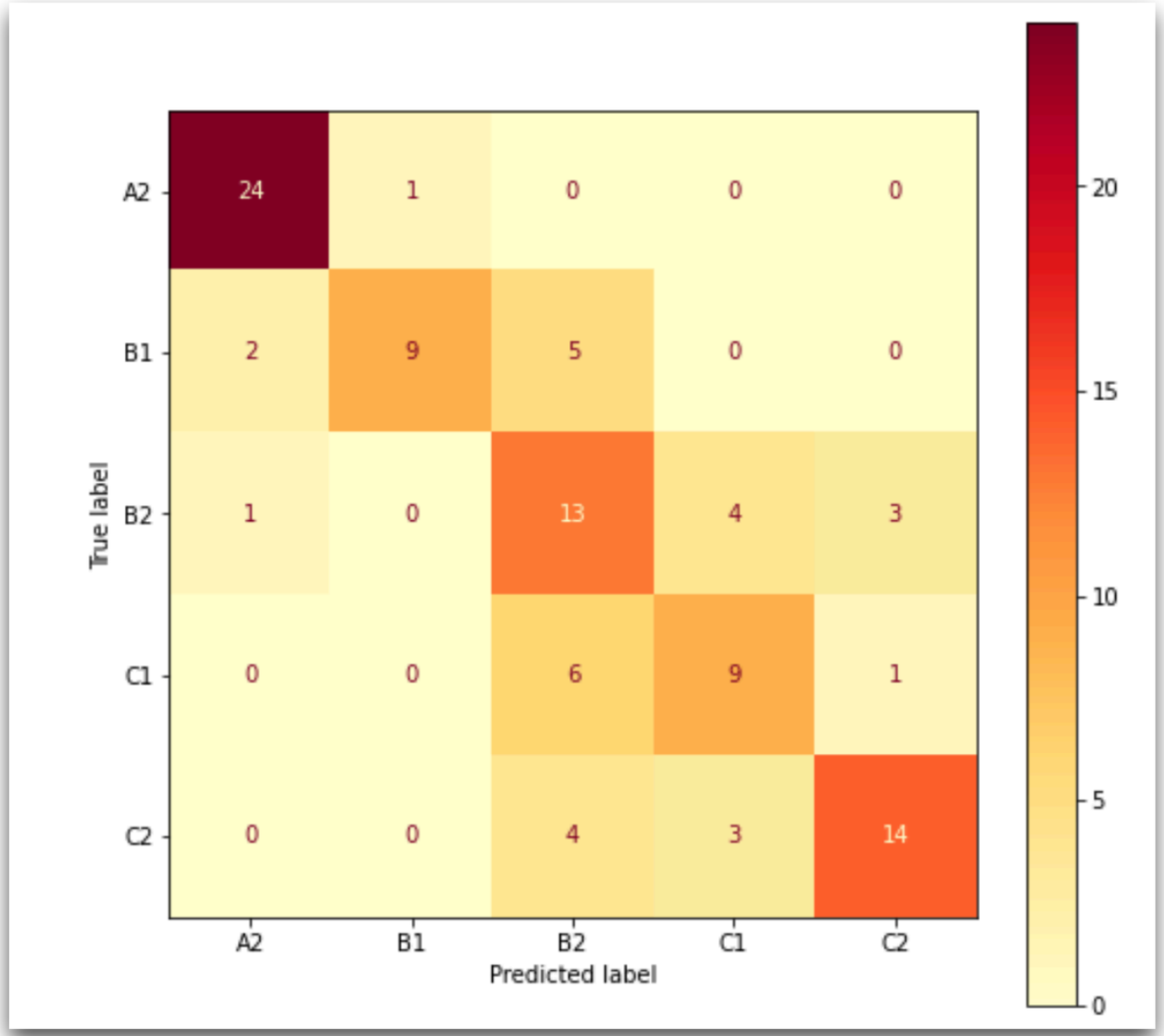
Overfitting

Training Accuracy: **1.0**

Validation Accuracy: **0.8**

Best Model: **Multinomial Naive Bayes**

Training Accuracy:
0.72
Validation Accuracy:
0.70

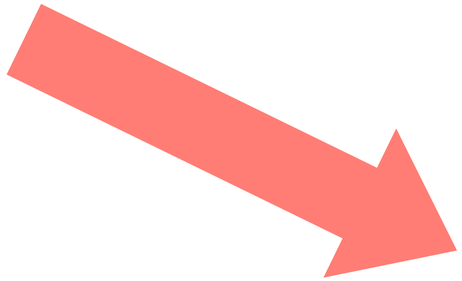


Challenges:

Lack of Data

Use Neural Networks

Build an App



```
pipe_nb.fit(X_train, y_train)
print("Training Acc.: ", pipe_nb.score(X_train, y_train))
print("Valid Acc.: ", pipe_nb.score(X_test, y_test))
```

```
Training Acc.:  0.7205240174672489
Valid Acc.:  0.696969696969697
```

```
test_pred = pipe_nb.predict(X_test)
```

```
list_of_keys = cefr_pos_df.columns.tolist()
twilight_sample = pp.process_sample(twilight, list_of_keys,
```

```
pipe_nb.predict([twilight_sample])
```

```
array(['B2'], dtype='<U2')
```

A2 B1 B2 C1 C2

CEFR Level: CEF Level B2

(Intermediate)
IELTS Level 5-6

Suggested vocabulary:

deep-voiced
aback
implied
more—that
prohibited
manner
toignore

number of words: 73
average sentence length: 15
average word length: 4.6
word complexity: 1538

DEFINITIONS

NEW TEXT

WORD STATS

THANK YOU FOR LISTENING

I HOPE TO HEAR FROM YOU SOON