# Detecting the difficulty level of French texts

Joris Booij-Liewes
HEC Lausanne
Data Mining
Ahmed M’sabou

# Presentation Outline

ISFA

UNIL | Université de Lausanne

Maastricht University

# Introduction

01

## Problematic

Given a text in French, can you predict its difficulty level (A1, A2, B1, B2, C1, C2)?

02

## Why this project?

We have noticed that to improve one's skills in a new foreign language, it is vital to read texts in a suitable difficulty level.

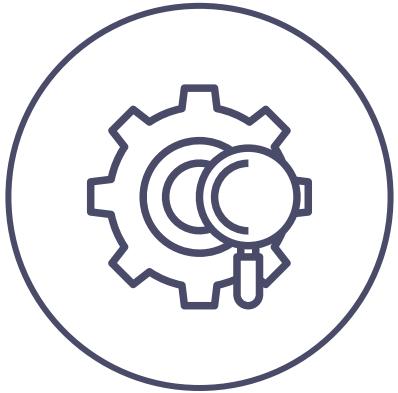
ISFA

UNIL | Université de Lausanne

Maastricht University

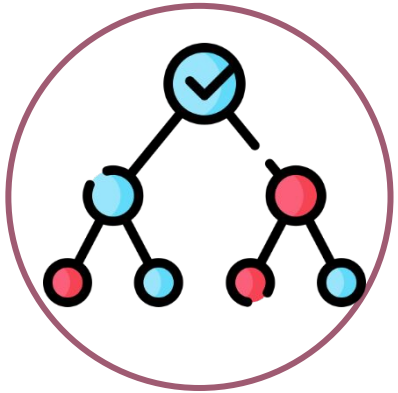
# Models

Classic models saw in course



**Logistic regression**

**kNN**



**Decision Tree**

**Random Forest**

ISFA

UNIL | Université de Lausanne

Maastricht University

# Results

Using a validation set that is 20% of the data

| | Logistic regression | kNN | Decision Tree | Random Forests | Camem BERT |
|---|---|---|---|---|---|
| Precision | 0.4582 | 0.3997 | 02964 | 0.4187 | 0.5795 |
| Recall | 0.4604 | 0.3677 | 0.3021 | 0.4208 | 0.5740 |
| F1-score | 0.4561 | 0.3530 | 0.2975 | 0.4099 | 0.5717 |
| **Accuracy** | **0.4604** | **0.3677** | **0.3021** | **0.4208** | **0.5740** |

ISFA

UNIL | Université de Lausanne

Maastricht University

# Techniques to improve our models?

What we did to improve the accuracy of our models

## Ensemble

Ensemble models are machine learning models that combine the predictions of multiple individual models to make a final prediction.

## Embedding

embeddings are a way of representing data such as words, phrases, or other discrete items as continuous, low-dimensional numerical vectors

## Word2vec

Word2vec is a method for learning vector representations of words

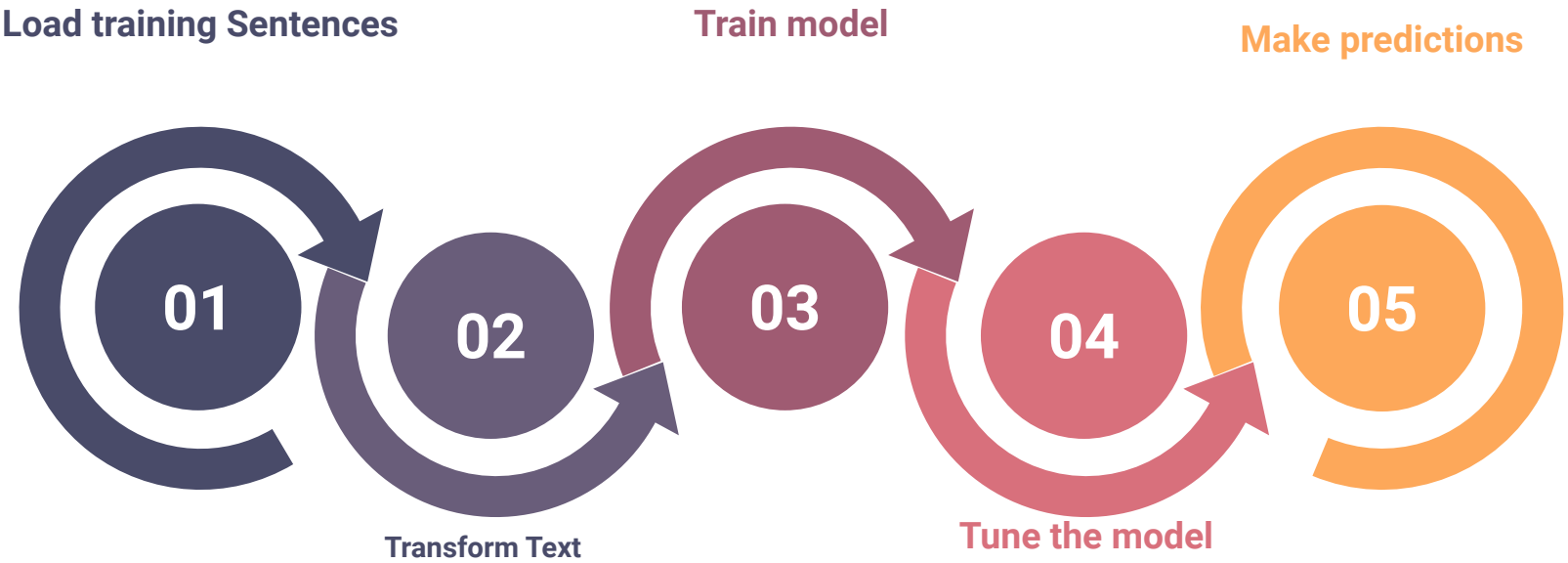
ISFA

UNIL | Université de Lausanne

Maastricht University

# CamemBERT model

French version of the BERT model for NLP classification




ISFA

UNIL | Université de Lausanne

Maastricht University

# Explaining CamemBERT & Code

The process of preparing the data

```
# train data
df_train = pd.read_csv('training_data.csv')

# Rename Labeling
df_train['difficulty'] = df_train['difficulty'].replace(['A1','A2','B1', 'B2', 'C1', 'C2'],[0,1,2,3,4,5])

# test data
df_test = pd.read_csv('unlabelled_test_data.csv')
```

```
epochs = 6
MAX_LEN = 64
batch_size = 16
device = torch.device('cuda' if torch.cuda.is_available() else 'cpu')

# Initialize CamemBERT tokenizer
tokenizer = CamembertTokenizer.from_pretrained('camembert/camembert-large', do_lower_case=True)
```

ISFA

UNIL | Université de Lausanne

Maastricht University

# Explaining CamemBERT & Code

Embedding

```
#user tokenizer to convert sentences into tokenizer'
input_ids  = [tokenizer.encode(sent,add_special_tokens=True,max_length=MAX_LEN) for sent in text]

# Pad our input tokens
input_ids = pad_sequences(input_ids, maxlen=MAX_LEN, dtype="long", truncating="post", padding="post")

attention_masks = []

# Create a mask of 1s for each token followed by 0s for padding
for seq in input_ids:
    seq_mask = [float(i>0) for i in seq]
    attention_masks.append(seq_mask)
```

| Words | cycle | car | road | tree | root | hotel | river |
|---|---|---|---|---|---|---|---|
| Indices | 1 | 2 | 3 | 4 | 5 | 6 | 7 |
| d | 0.2 | 0.1 | 0.1 | 0.2 | 0.3 | 0.4 | 0.3 |
| | 0.1 | 0.7 | 0.4 | 0.9 | 0.2 | 0.8 | 0.7 |
| | 0.5 | 0.8 | 0.6 | 0.5 | 0.8 | 0.2 | 0.8 |
| | 0.6 | 0.2 | 0.9 | 0.3 | 0.3 | 0.6 | 0.1 |

WORD EMBEDDINGS

| input_ids | | |
|---|---|---|
| 101 | 7592 | 2088 |
| 999 | 102 | 0 |

- real tokens
- [PAD] tokens

attention_mask

| | | |
|---|---|---|
| 1 | 1 | 1 |
| 1 | 1 | 0 |

ISFA

UNIL | Université de Lausanne

Maastricht University

# Explaining CamemBERT & Code

Train/val Split and Training the model

```
# Use train_test_split to split our data into train and validation sets for training
train_inputs, validation_inputs, train_labels, validation_labels, train_masks, validation_masks = train_test_split(input_ids, labels, attention_masks,
                                                            random_state=42, test_size=0.2)
```

```
model = CamembertForSequenceClassification.from_pretrained("camembert/camembert-large", num_labels=6)
```

CONFUSION MATRIX:

```
[[139  26   1   0   0   0]
 [ 41  75  39   1   2   0]
 [ 24  45  86   9   1   1]
 [  0   4  32  85  26   6]
 [  0   0   8  42  79  23]
 [  0   1   3  37  37  87]]
```

ISFA

UNIL | Université de Lausanne

Maastricht University

# Conclusion

**What was our best model?**

**If we had more time?**

**What did we learn from this project?**

ISFA

UNIL | Université de Lausanne

Maastricht University

# THANK YOU !