Projet 8 : Compétition Kaggle

Ali-higo Ebo Adou

November 7, 2022

Outline

Outline

Natural Language Inference (NLI)

NLI considers two sentences: a "premise" and a "hypothesis".

Given the premise, the task is to determine whether the hypothesis is :

- true (entailment),
- false (contradiction) or
- neutral.

Sentence A (Premise)	Sentence B (Hypothesis)	Label
A soccer game with multiple males playing.	Some men are playing a sport.	entailment
An older and younger man smiling.	Two men are smiling and laughing at the cats playing on the floor.	neutral
A man inspects the uniform of a figure in some East Asian country.	The man is sleeping.	contradiction

Connections to other tasks

Task	**NLI framing**
Paraphrase	text ≡ paraphrase
Summarization	text ⊏ summary
Information retrieval	query ⊐ document
Question answering	question ⊐ answer
	Who left? ⇒ Someone left
	Someone left ⊐ Sandy left

La plateforme Kaggle

Kaggle est une plateforme web qui accueille la plus grande communauté de Data Science au monde. Sont accessibles gratuitement des GPU et une grande quantité de données et de codes publiés par la communauté.

Les compétitions

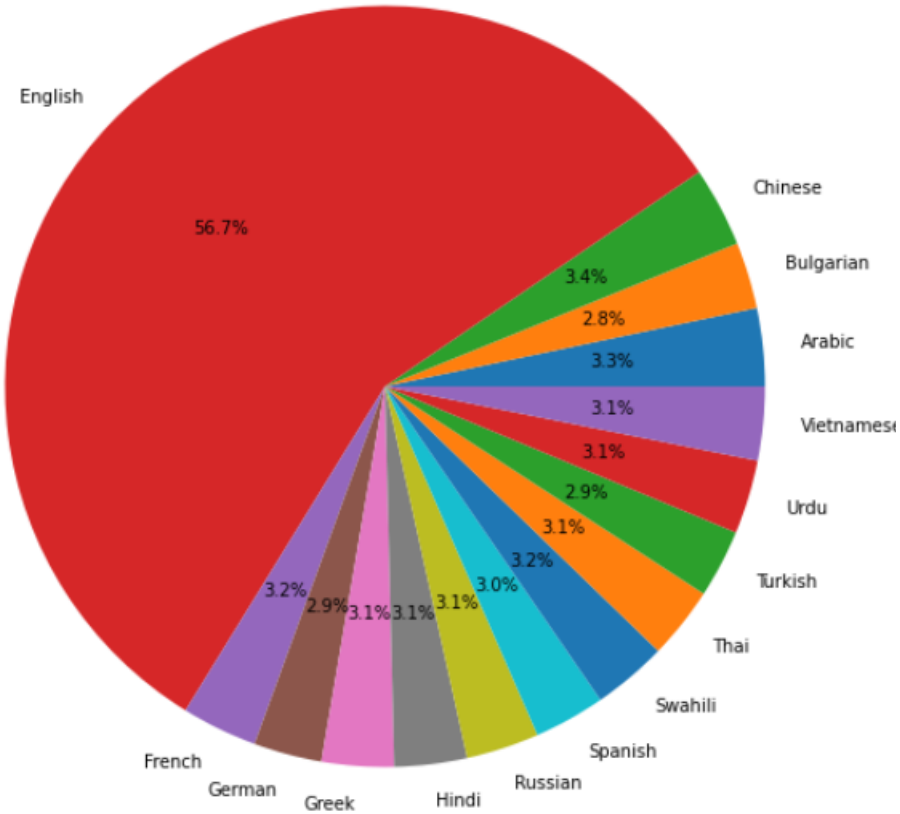
- Sponsorisée : par des entreprises, associations, etc... avec des prix à la clé.
- Recherche : sujets orientées vers la recherche.
- Débutant : parfaitement adaptées pour les nouveaux utilisateurs avec des sujets accessibles et des données facilement interprétables.

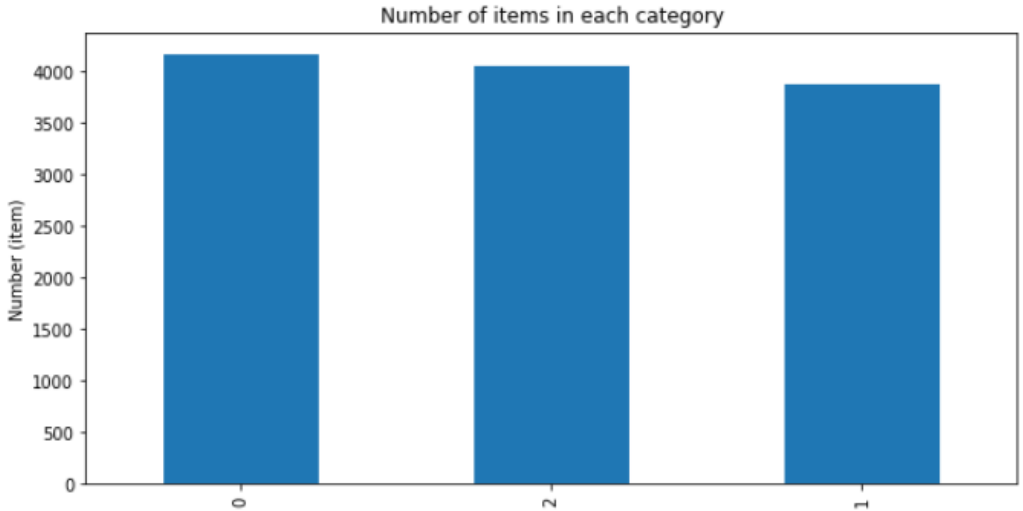
Outline

Contradictory, My Dear Watson

- 15 langues différentes & plus de 12000 lignes dans le *train set*

  Labels : • 0== entailment • 1 == neutral • 2 == contradiction



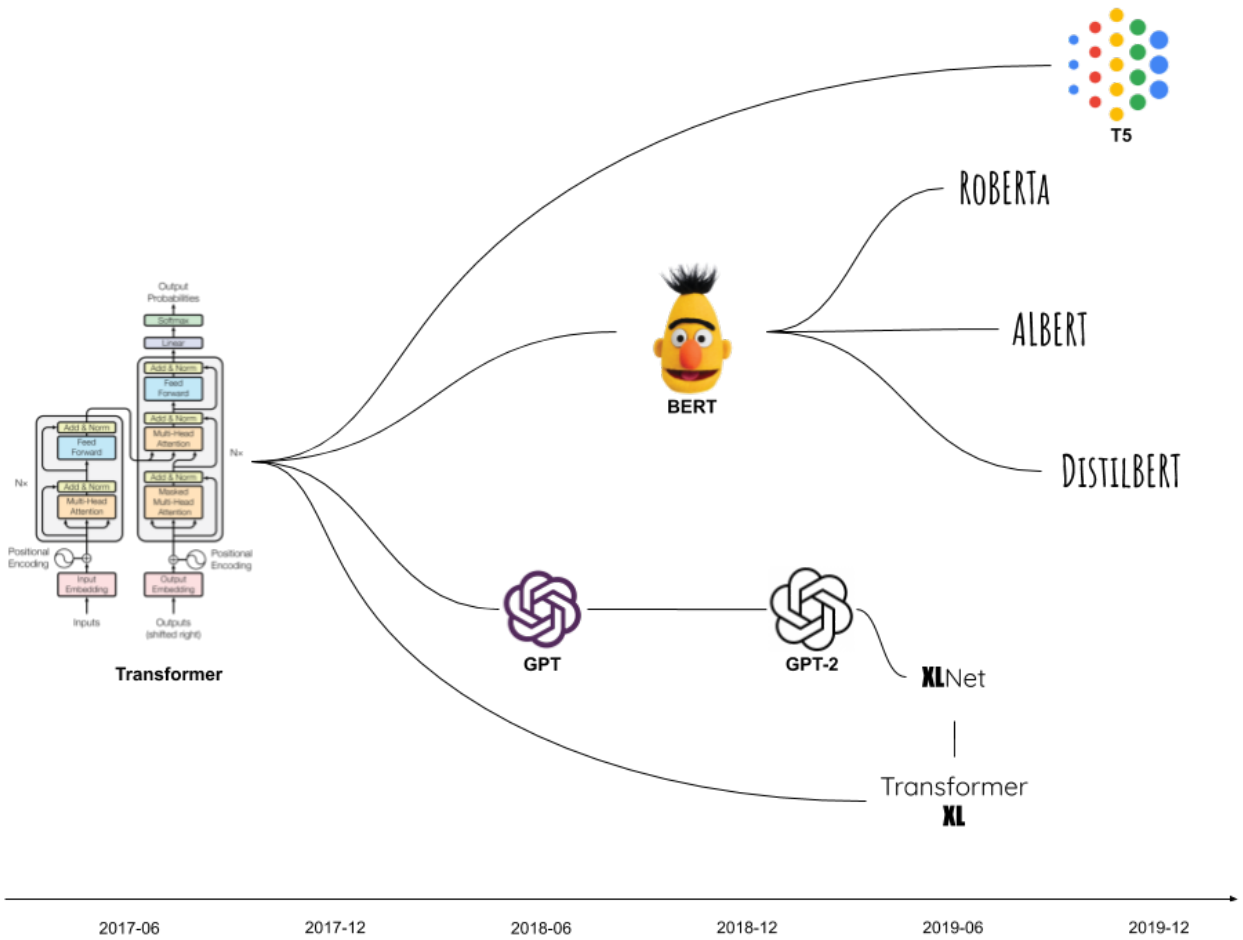


Préparation des données

- *Tokenization*: [CLS] **premise** [SEP] **hypothesis** [SEP]
- *Data Augmentation*: *On-the-fly* pendant la méthode *.fit()*

Outline

La famille des modèles BERT



Pré-entraînement pour les tâches de NLI

The General Language Understanding Evaluation (GLUE) benchmark is a collection of 9 datasets for evaluating natural language understanding systems. Tasks are framed as either single-sentence classification or sentence-pair classification tasks.

- **SNLI** : The Stanford Natural Language Inference corpus is a collection of 570k human-written English sentence pairs manually labeled for balanced classification.
- **MNLI** : The Multi-Genre Natural Language Inference corpus is a crowd-sourced collection of 433k sentence pairs annotated.
- **XNLI** : The Cross-lingual Natural Language Inference corpus is the extension of the Multi-Genre NLI (MultiNLI) corpus to 15 languages.

Modèle final

The RoBERTa model is based on Google's BERT model released in 2018. It modifies key hyperparameters, removing the next-sentence prediction from pretraining objective and training with much larger mini-batches and learning rates.

`'joeddav/xlm-roberta-large-xnli'`

This model takes `xlm-roberta-large` and fine-tunes it on a combination of NLI data in 15 languages.

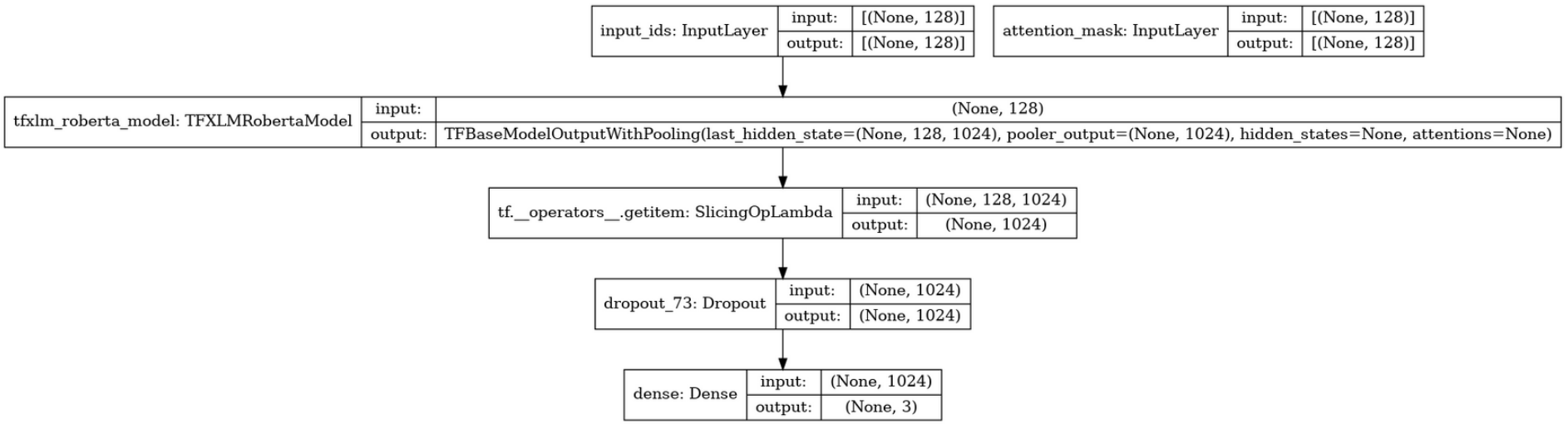
The model is intended to be used for zero-shot text classification, especially in languages other than English.

For English-only classification, it is recommended to use `bart-large-mnli` or a `distilbart-mnli` models.

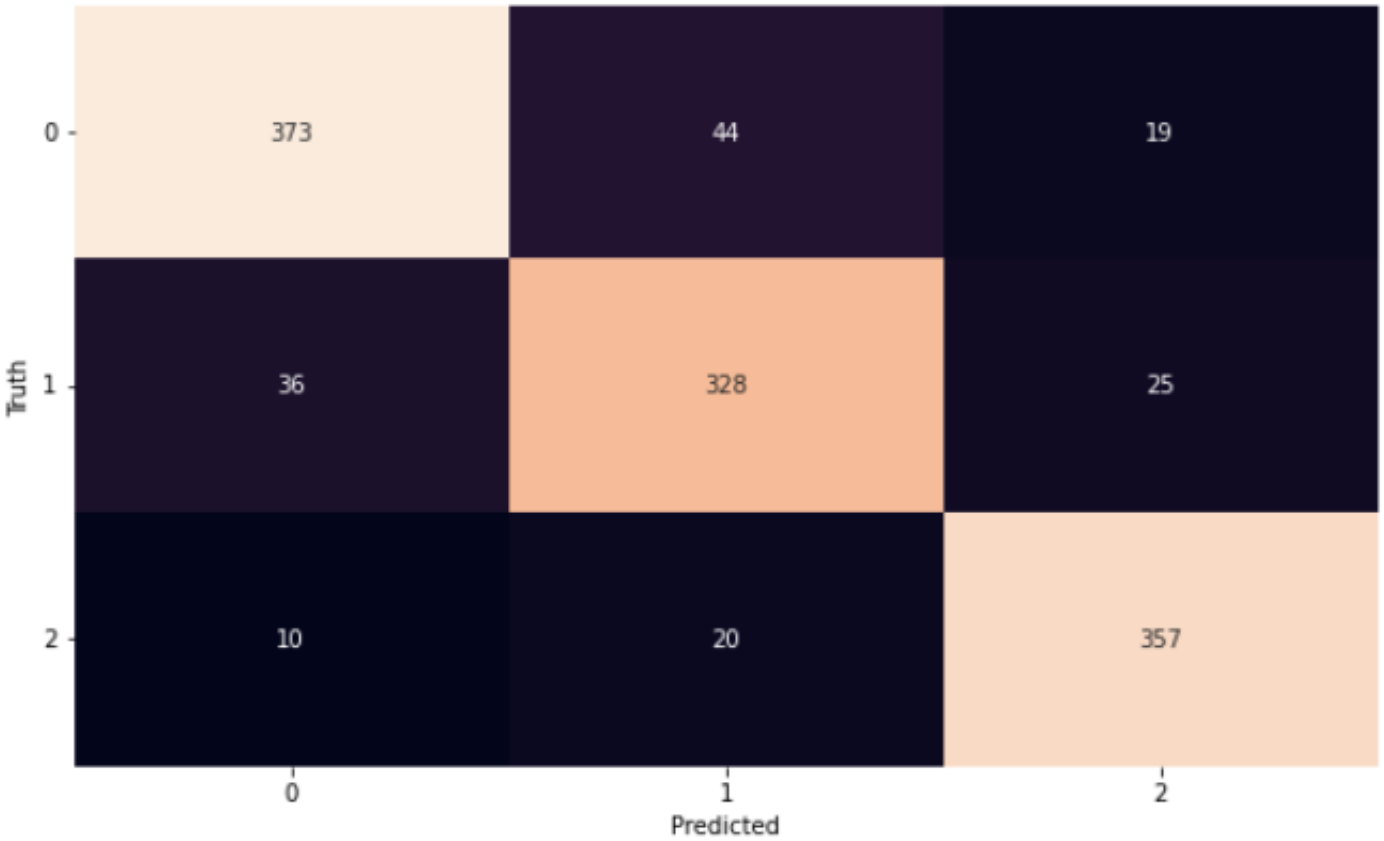
Outline

Tunning des hyperparamètres

- Couche *Dense Fully-Connected & Dropout*
- *Learning rate*



Matrice de Confusion



Score : 0.8733 $\longrightarrow$ Rank: 26/52

Perspectives

- Utiliser l'API `googletrans` pour utiliser modèle monolingue
- Data Augmentation avec *WordNet*

Merci de votre attention.