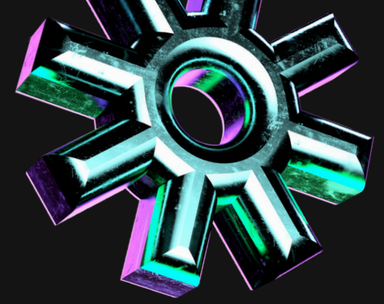
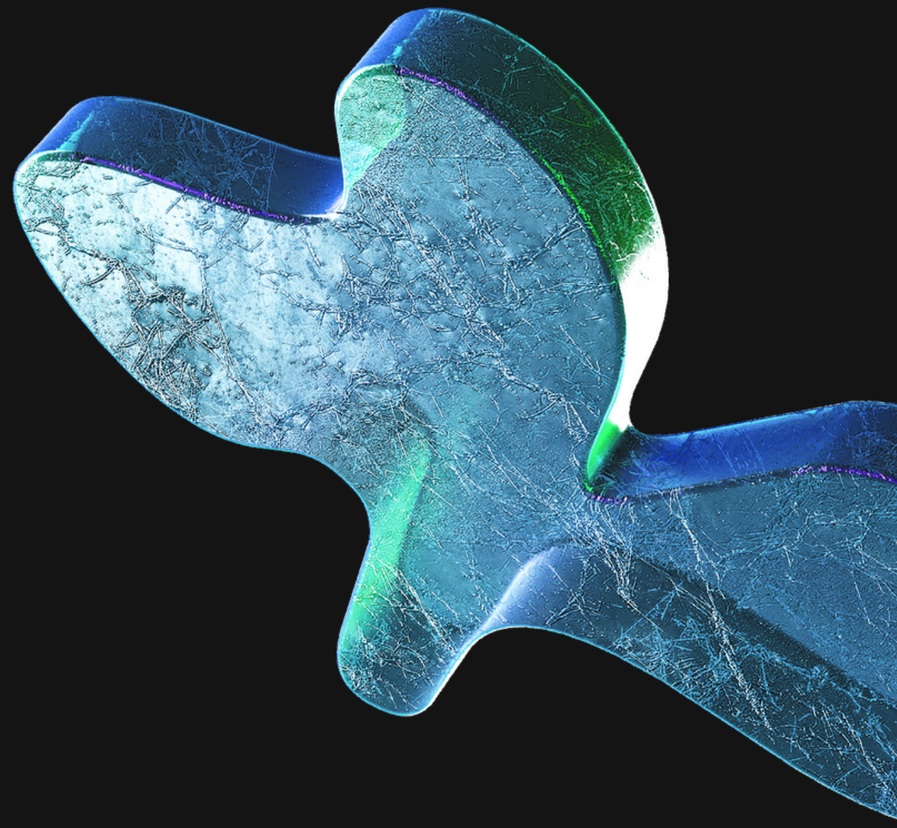


Detecting the difficulty level of French texts



Leo Andrade
Thomas Anthoine Milhomme

The problem

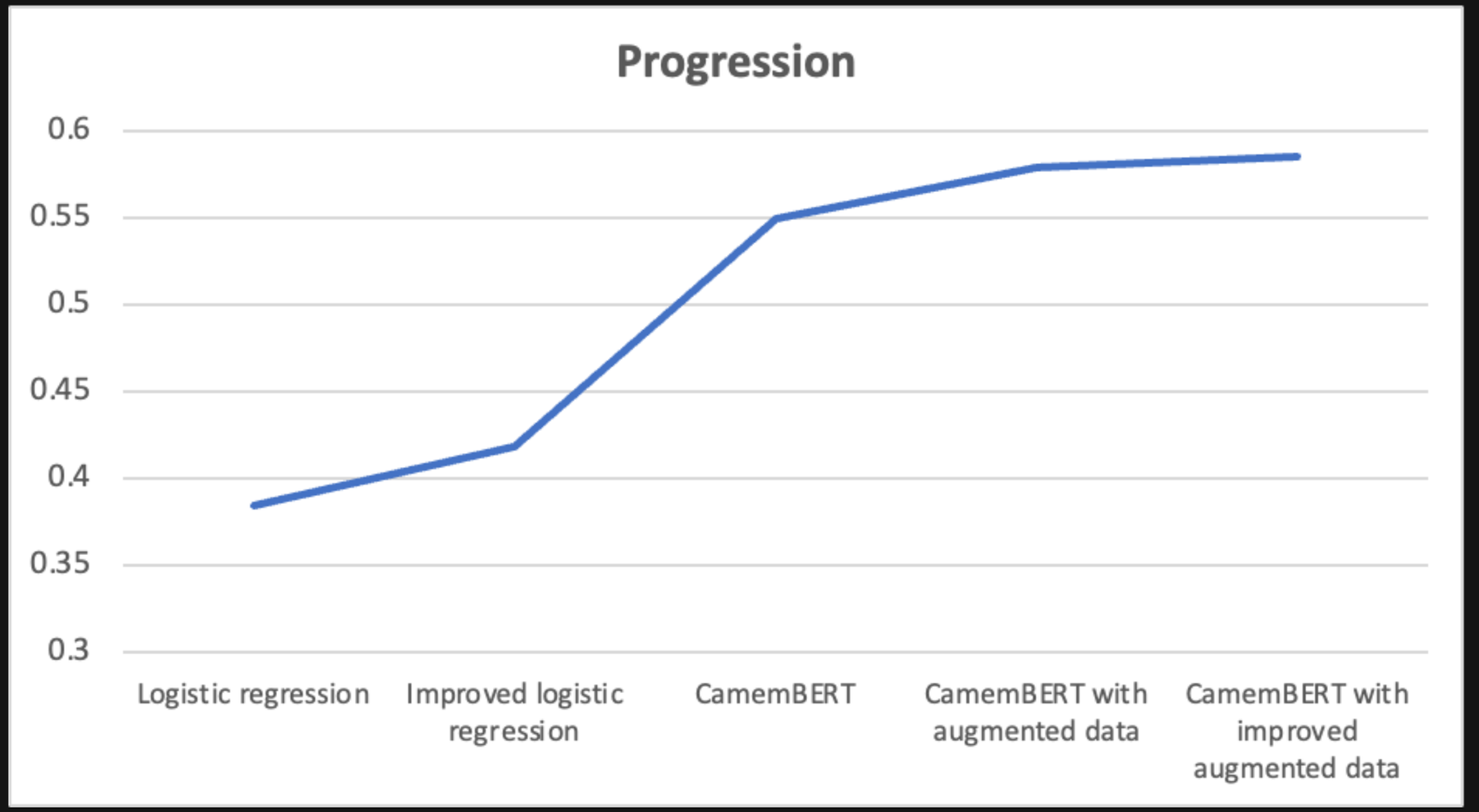
Labelled data

id	sentence	difficulty
0	Les coûts kilométriques réels peuvent diverger sensiblement des valeurs moyennes en fonction du moyen de transport utilisé, du taux d'occupation ou du taux de remplissage, de l'infrastructure utilisée, de la topographie de	C1
1	Le bleu, c'est ma couleur préférée mais je n'aime pas le vert!	A1
2	Le test de niveau en français est sur le site Internet de l'école.	A1
3	Est-ce que ton mari est aussi de Boston?	A1
4	Dans les écoles de commerce, dans les couloirs de places financières, il arrive aujourd'hui de croiser de jeunes adultes de 20 ou 25 ans qui prévoient d'ouvrir une maison d'hôtes "dans une quinzaine d'années".	B1
5	voilà une autre histoire que j'ai beaucoup aimée.	A2
6	Les médecins disent souvent qu'on doit boire un verre de vin rouge après les repas.	A2
7	Il est particulièrement observé chez les personnes ayant un besoin de popularité développé, qui considèrent dès lors le moindre signal du smartphone comme un possible indicateur de leur degré de popularité.	B2
8	J'ai retrouvé le plaisir de manger un oeuf à la coque	A2

Unlabelled data

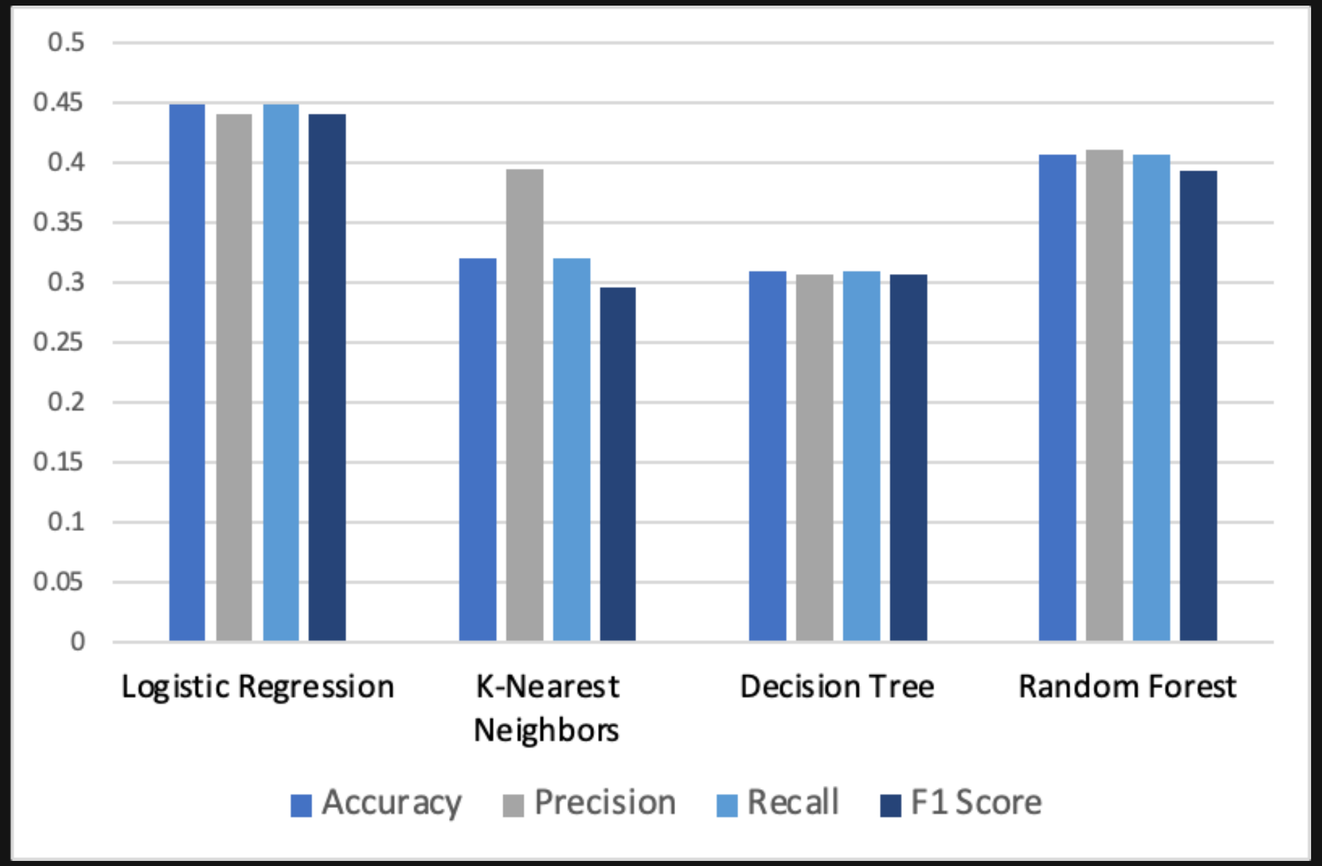
id	sentence
0	Nous dûmes nous excuser des propos que nous eûmes prononcés
1	Vous ne pouvez pas savoir le plaisir que j'ai de recevoir cette bonne nouvelle.
2	Et, paradoxalement, boire froid n'est pas la bonne parade.
3	Ce n'est pas étonnant, car c'est une saison mystérieuse
4	Le corps de Golo lui-même, d'une essence aussi surnaturelle que celui de sa monture, s'arrangeait de tout obstacle matériel, de tout objet gênant qu'il rencontrait en le prenant comme ossature et en se le rendant intérieur, f
5	Elle jeta un cri, un petit cri, voulut se dresser, se débattre, le repousser ; puis elle céda, comme si la force lui eût manqué pour résister plus longtemps.
6	Madame, Monsieur, Votre fils Léo arrive tous les jours en retard à l'école.
7	Comment tu as trouvé le repas de ce midi
8	Mais la racine du mal est bel est bien cette façon de penser tendant vers un manichéisme néfaste, néfaste car le respect des opinions divergentes et le débat s'estompent au profit de l'extrémisme et l'uniformisme de la pens

Progression



Final best score: 0.585

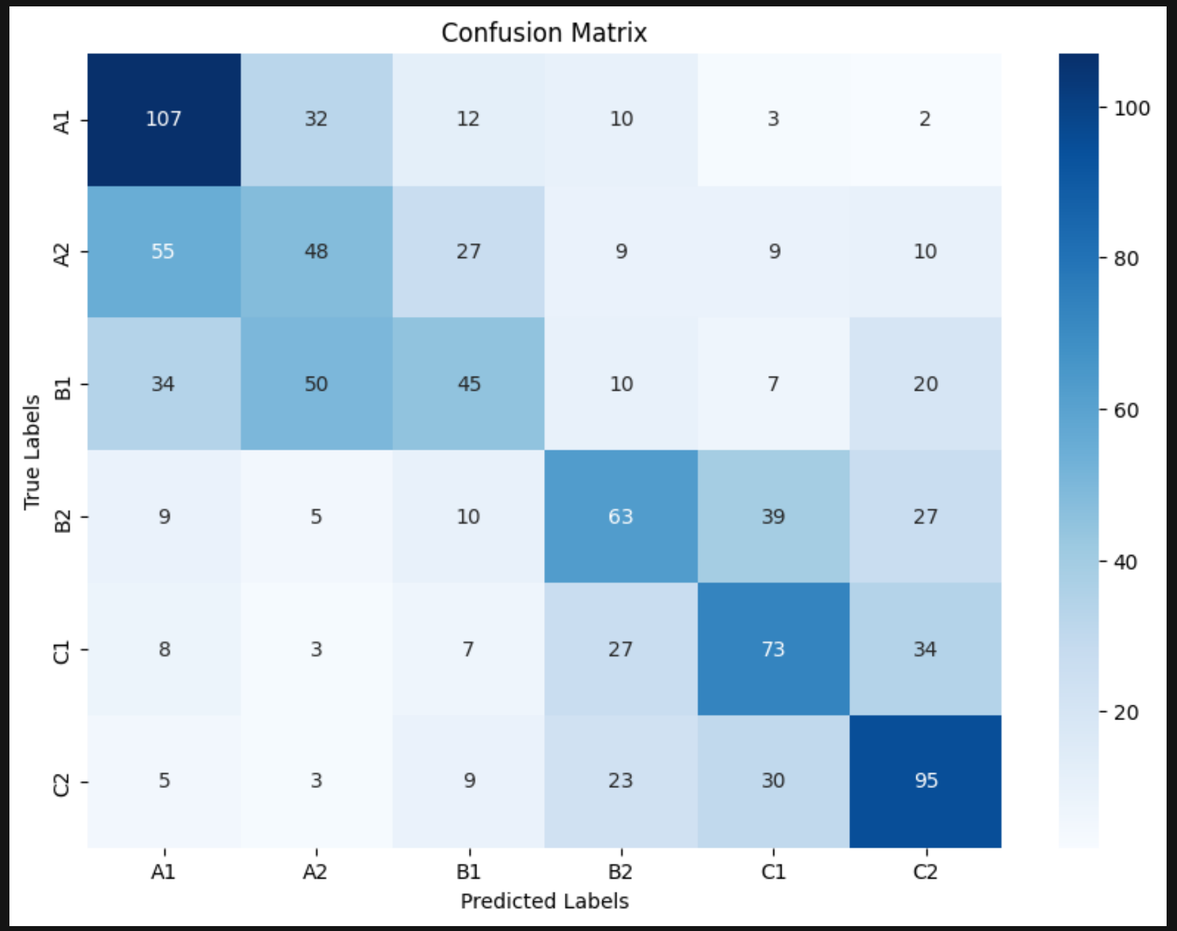
Methods without cleaning data



Logistic regression show the best results

Accuracy	0.449
Precision	0.441
Recall	0.449
F1 Score	0.44

Logistic regression’s confusion matrix

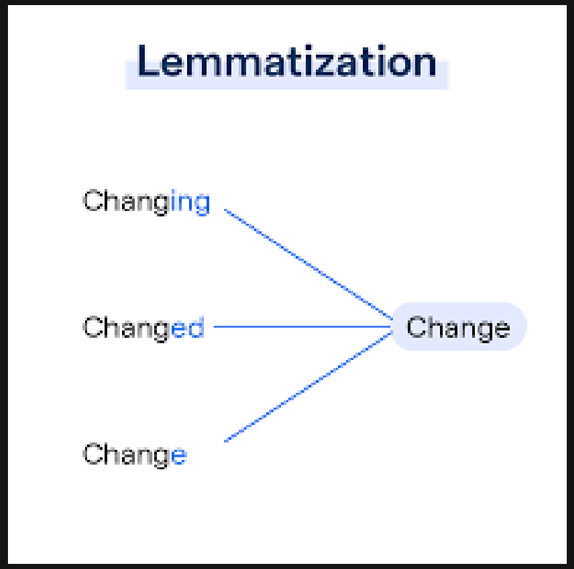


- Quite well distributed
- Concentrated in the diagonal
- Only few big mistakes (ex confuse C2 with A1)
- Most errors on neighbouring difficulty levels

Data cleaning, preprocessing, ...

We have tried to use different techniques to improve our analysis, for example:

- POS tagging
- Sentiment analysis
- Tokenization
- Bigrams
- Stopwords
- Lemmatization



But, instead of improving the predictions, they were even worse

Models specialized for French

BERTS MODELS

The best results were provided by the CamemBERT model

Analyse of hyper-parameters

- num_train_epochs= #
- per_device_train_batch_size= #
- per_device_eval_batch_size= #

The best result that we had for a while was 0.55, obtained with 6 epochs, a train batch size of 16 and an evaluation batch size of 24

We also tried to clean the data, lemmatize, etc, but the results were even worse

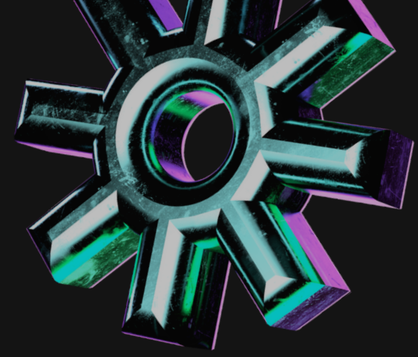
Others methods tried

We have tried different methods and algorithm

- Scikit-learn
- ULMFiT
- Support Vector Machine algorithm(SVM)

But best results were still provided by the CamemBERT model

Data Augmentation

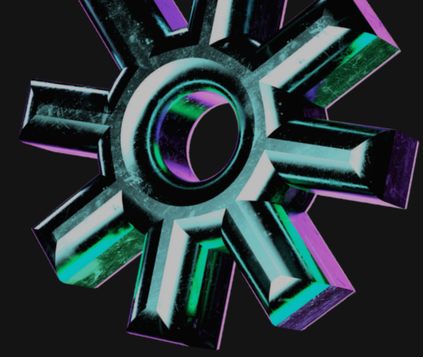


We tried different combinaisons of techniques (cognates, POS, translations, paraphrase, ...) and verified either, manually and by testing the most consistent ones.

	b
	i
	e
	n
	En chemin, il a pu admirer la magnifique fontaine Bartholdi sur la grande Place
	des Terreaux.
	T
	h
	e
	w

Example of problems encountered

Data Augmentation



- Finally, the technique we selected is a combination of paraphrasing using a language masking model (CamemBERT) and synonym substitution (using NLTK).

- The level assigned to the new sentence is the level of the sentence on which the data increase is based

- The training dataset created contains around 13,000 observations.

Discussion

When we added pre-processing, the results were often worse

Models like Camembert are already highly trained and complete, adding pre-processing steps can lead to a loss of information.
It may therefore compromise the ability to capture the complexity of the language.

Best model

Best predictions obtained by:

Finding a good augmentation of the data

Train the model with CamemBERT on the augmented data

This method gave us the following results

Kaggle score	0.585
Accuracy	0.773
Precision	0.786
Recall	0.773
F1 Score	0.775