

Rapport 26.02.2025

Structure des fichier

J'ai restructuré les fichiers du projet pour faciliter la navigation. Désormais, il y a trois dossiers principaux : dataSources, testModel et variables.

- dataSources contient les différentes sources de données utilisées, ainsi que le fichier Python permettant de récupérer les données via leurs API. Il inclut également des fichiers Python contenant différents tests effectués sur les sources de données (par exemple, PubMed/data_num.py, qui récupère le nombre d'articles publiés sur PubMed). Pour l'instant, seule la source PubMed est incluse.
- testModel contient tous les scripts Python de test des modèles, ainsi que les résultats et les datasets utilisés.
- variables regroupe tous les fichiers Python contenant des variables réutilisées à travers le projet.

Tous les autres dossiers contiennent des fichiers utiles ou utilisés dans le cadre du projet.

Test des models

Longeur des textes classifier

Après discussion avec Monsieur Glück, nous avons décidé d'examiner si la longueur du texte influence la performance des modèles. L'objectif est également d'observer si certains modèles sont plus performants sur des textes longs mais moins précis sur des textes courts, et inversement.

J'ai commencé par analyser la longueur des textes dans mon dataset :

```
Longuest: 3863
Shortest: 31
Mean: 823.4525
Median: 538.5
```

Sur la base de ces résultats, j'ai décidé de séparer les textes en quatre catégories : SHORT, MEDIUM, LONG et VERY LONG:

- Short: 0-300 caractères
- Medium: 301-600 caractères
- Long: 601-900 caractères
- Very Long: 901-inf caractères

Cela donne la répartition suivante:

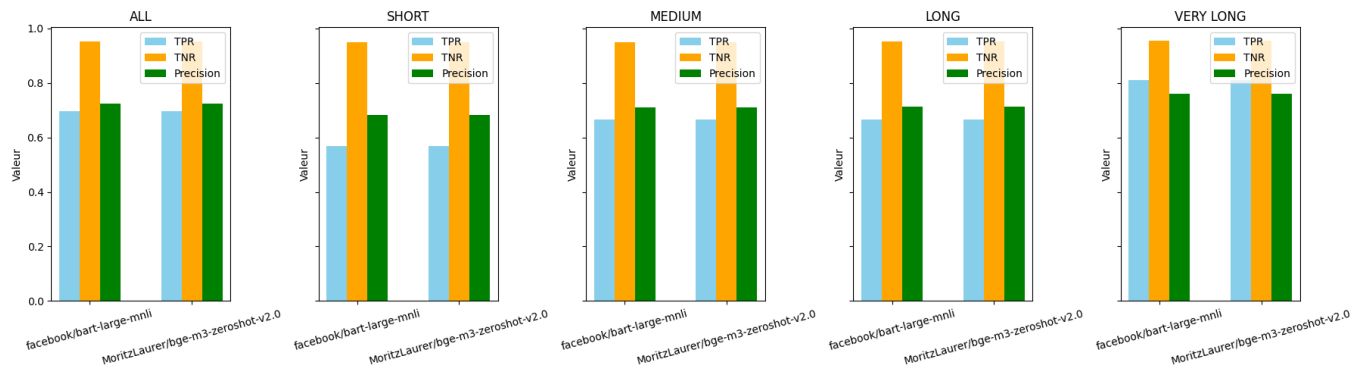
```
SHORT: 144
MEDIUM: 300
```

LONG: 75
VERY LONG: 281

Par la suite, il faudra recréer le dataset afin d'obtenir une répartition plus équilibrée des longueurs d'articles.

J'ai également modifié mon modèle de test pour classer les résultats obtenus en fonction des quatre catégories de longueur. Cela permettra d'effectuer des comparaisons et ainsi de faciliter mon choix de modèle.

En parallèle, j'ai réfléchi à l'affichage des résultats. Actuellement, la version que j'ai retenue est la suivante:



Cependant, je trouve que cette représentation manque de lisibilité. Je vais donc poursuivre mes recherches pour trouver un affichage plus clair et pertinent.

Dans l'image ci-dessus, seules les valeurs du modèle facebook/bart-large-mnli sont correctes, car je n'ai pu retrouver l'accès au serveur Baobab de l'UNIGE que ce matin (en raison de problèmes avec mes clés SSH).

Je n'ai pas pu exécuter les tests sur mon PC personnel, car chaque exécution posait des problèmes et prenait plus d'une heure par modèle. J'ai donc préféré attendre l'accès à Baobab.

Toutefois, j'ai tout de même tenté d'obtenir des résultats pour un modèle, afin d'avoir une première idée de la direction que je prenais. Voici les résultats pour facebook/bart-large-mnli:

```
-----  
Result confusion matrix: [[103, 48], [78, 923]]  
True Positive Rate (TPR): 0.569060773480663  
True Negative Rate (TNR): 0.9505664263645726  
Precision: 0.6821192052980133  
-----  
Result confusion matrix: [[247, 100], [124, 1929]]  
True Positive Rate (TPR): 0.6657681940700808  
True Negative Rate (TNR): 0.9507146377525875  
Precision: 0.7118155619596542  
-----  
Result confusion matrix: [[60, 24], [30, 486]]  
True Positive Rate (TPR): 0.6666666666666666  
True Negative Rate (TNR): 0.9529411764705882  
Precision: 0.7142857142857143  
-----  
Result confusion matrix: [[265, 83], [62, 1838]]  
True Positive Rate (TPR): 0.8103975535168195  
True Negative Rate (TNR): 0.956793336803748  
Precision: 0.7614942528735632  
-----  
Time to classify all articles: 6689.5230884552 seconds  
Result confusion matrix: [[675, 255], [294, 5176]]  
True Positive Rate (TPR): 0.6965944272445821  
True Negative Rate (TNR): 0.953047320935371  
Precision: 0.7258064516129032
```

L'ordre des résultats, de haut en bas sur l'image, est le suivant : SHORT, MEDIUM, LONG, VERY LONG et ALL. On observe que ce modèle est plus performant sur les textes VERY LONG.

Suite pour semaine prochaine

Je suis conscient de ne pas avoir avancé autant que prévu, mais voici les tâches que je compte accomplir d'ici vendredi:

- Refaire tourner mes testes sur les serveurs Baobab
- Afficher tous les résultats d'une façon lisible
- Regarder ce qu'est Mistral AI (une LLM dont un amis m'a parler)
- Essayer Ollama
- Mettre au propre tous les résultats pour les LLM et les modèles de HuggingFace