OPENCLASSROOMS

kaggle

Participez à une compétition Kaggle !

Note méthodologique

Rédacteur : Corentin BONNEFOND

Alternant OpenClassrooms
Cursus Ingénieur Machine Learning

Contact Entreprise
Nom contact : Paul CHARNAY
Email : paul.charnay@abelio.io

Contact OpenClassrooms
Nom contact : Benjamin DEPLUS
Email : benjamin.deplus@sciencespo.fr

Confidentialité : NON

1 La note Méthodologie

La note peut se lire indépendamment du rapport, elle résume la démarche méthodologique adoptée jusqu'à obtention du modèle final soumis sur Kaggle.

1.1 Approche par TF-IDF

La première approche utilisée est basée sur la fréquence des mots, elle se nomme TF-IDF.

Après nettoyage et étude exploratoire, 2000 tokens issus du corpus ont été conservés.

Le feature engineering revient à la concaténation entre discourse type et iscourse text.

Le modèle final est la moyenne arithmétique entre trois classifiers différents (Regression Logistique + RandomForest + GradientBoostingClassifier).
Note : Pour résoudre les problèmes de catégorisation multiclasse, un classifier OneVsRest est associé à chaque modèle ci dessus.

Le score obtenu avec cette méthode est acceptable : 0.784, cependant le classement dans le leaderboard n'est pas satisfaisant.

1.2 Une approche avec DistilBERT

Pour améliorer les prédictions, deux méthodes transfer learning utilisant distilBERT ont été testées.

- Une première approche inspirée de [1].
  Le feature engineering est semblable à celui de la méthode TF-IDF (utilisation du séparateur SEP de DistilBERT en plus). En sortie de réseau, des tests ont été effectués (ajout couche dense et multi sample dropout par exemple) pour limiter l'overfitting.
- Une deuxième approche dont le feature engineering est repris de [2] : discourse type + SEP + discourse text + SEP + Essay text.
  La colonne essay text permet de contextualiser et d'améliorer le score de précision.

La version avec essay text permet de gagner quelques point de précision en comparaison avec la première version sans essay text

Malheureusement, le meilleur score obtenu par transfer learning : 0.808, reste très proche de celui de l'approche TF-IDF qui ne permet pas d'avoir un classement satisfaisant.

1.3 Le modèle de la communauté

Afin d'améliorer au maximum le score obtenu et obtenir un classement satisfaisant, des améliorations ont été effectuées sur un notebook fourni par la communauté [2]. Ce notebook est lui-même basé sur deux notebooks différents utilisant des variantes de BERT. Pour ces deux notebook, le vecteur d'entrée est un dérivé de (discourse type + SEP + Discourse text + SEP + Essay text).

Les étapes principales des modèles repris sont les suivantes :

- roberta-base + validation croisée
- deberta-large + BiLSTM + multi sample drop out + validation croisée

La contribution à la communauté fut de rassembler le modèle TF-IDF développé dans la première partie avec, les deux modèles ci-dessus et créer un modèle ensembliste.

La combinaison pour le modèle final :

- TF-IDF + RoBERTa-base + DeBERTa-large

Avec cette surcouche ensembliste, l'évolution du score n'est pas gigantesque (0.614 à 0.613), cependant elle permet de faire gagner quelques centaines de places. [3]

Approche	Score logloss
ENSEMBLE(TF-IDF + Classifier)	0.784
DistilBERT	0.808
RoBERTA + DeBERTa + ENSEMBLE(TF-IDF + Classifier)	0.613

Table 1 – Score par approche

Le lien du notebook avec un score de 0.613 : `https://www.kaggle.com/code/corentinbonnefond/fork-rob-deberta-woc`.

Note : Tous les codes sont consultables sur le dossier Github suivant : `https://github.com/bonnefco/P8`

Références

[1] Feedback prize : Eda & starter for beginners. disponible en ligne a l'adresse `https://www.kaggle.com/code/iamleonie/feedback-prize-eda-starter-for-beginners`.

[2] Fork+ | ensemble : deberta + roberta. disponible en ligne a l'adresse `https://www.kaggle.com/code/renokan/fork-ensemble-deberta-roberta`.

[3] `p8_fork_ensemble`. disponible en ligne à l'adresse `https://github.com/bonnefco/P8`.