Note méthodologique : preuve de concept

Dataset retenu

Le jeu de données utilisé dans cette preuve de concept est l'IMDB Movie Reviews Dataset, une référence pour la tâche de classification de sentiments. Il contient 50 000 critiques de films équilibrées entre étiquettes "positive" et "negative". Chaque ligne représente un avis exprimé librement par un internaute.

Pour des raisons de temps de calcul, un sous-échantillon de 2 000 critiques a été utilisé (1 000 pour l'entraînement, 1 000 pour le test), tiré aléatoirement. Les labels ont été encodés en binaire (1 pour "positive", 0 pour "negative"). Ce dataset est bien adapté à la tâche de classification binaire et permet une comparaison robuste entre modèles simples et avancés.

Les concepts de l'algorithme récent

L'algorithme moderne utilisé dans cette expérimentation est DistilBERT(source 1), une version compressée et optimisée du modèle BERT (Bidirectional Encoder Representations from Transformers), développé par Hugging Face.

DistilBERT repose sur l'architecture Transformer, introduite par Vaswani et al. (2017), qui repose sur un mécanisme d'attention multi-têtes permettant au modèle de capter les relations complexes entre les mots dans une séquence. DistilBERT réduit la taille de BERT tout en conservant 97 % de ses performances sur des tâches standards, avec une vitesse d'exécution accrue (60 % plus rapide).

Le fine-tuning consiste à réentraîner DistilBERT sur le dataset IMDB avec une tête de classification binaire ajoutée. Ce processus d'adaptation permet de spécialiser un

modèle généraliste aux spécificités d'une tâche cible, ici l'analyse de sentiment.

Par rapport aux méthodes classiques, ce type de modèle offre une compréhension contextuelle du langage, sensible à la structure grammaticale, aux négations, et aux nuances de sens.

Source 1: https://arxiv.org/abs/1910.01108

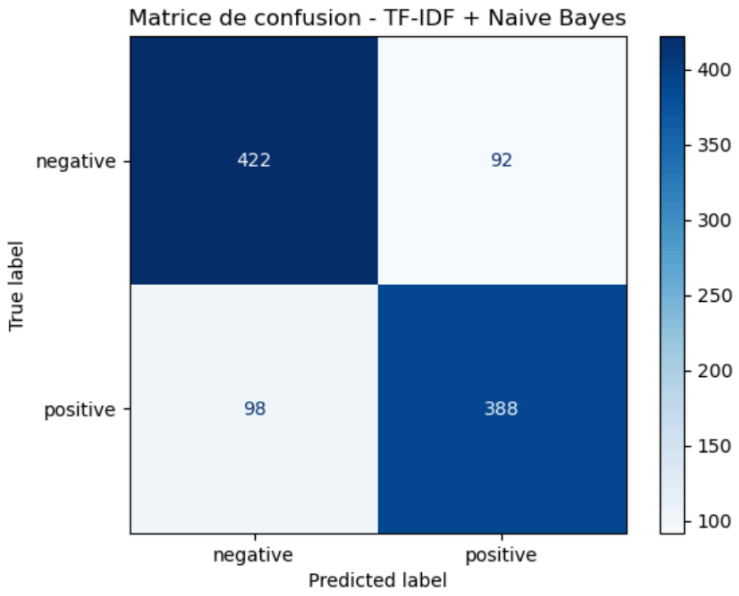
La modélisation

Deux pipelines de modélisation ont été construits et comparés :

1. TF-IDF + Naive Bayes :

- Texte vectorisé avec TF-IDF.

- Classifieur Naive Bayes entraîné sur ces vecteurs.



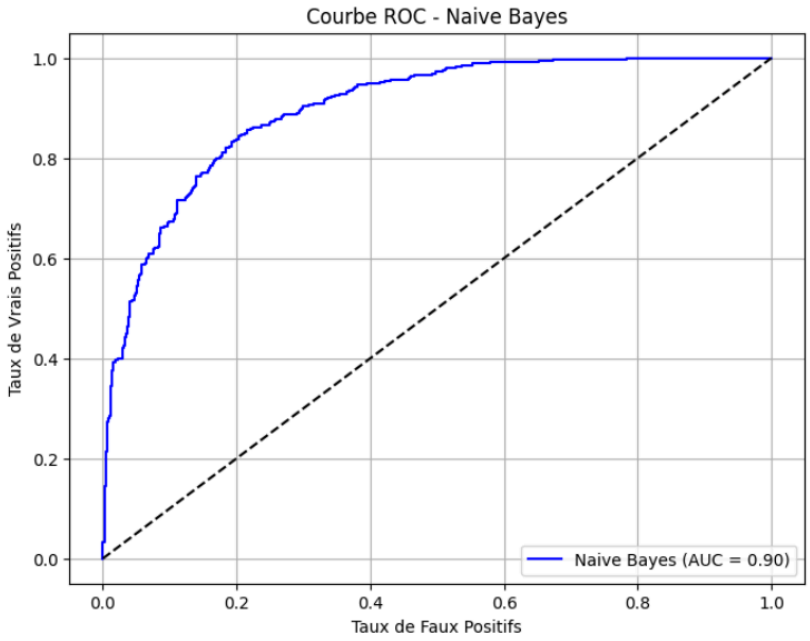
- Évaluation via précision, rappel, F1-score et AUC (courbe ROC tracée).

```
Résultats du modèle classique (TF-IDF + Naive Bayes) :
              precision    recall  f1-score   support

    negative       0.81      0.82      0.82       514
    positive       0.81      0.80      0.80       486

    accuracy                           0.81      1000
   macro avg       0.81      0.81      0.81      1000
weighted avg       0.81      0.81      0.81      1000

Accuracy : 81.00%
```

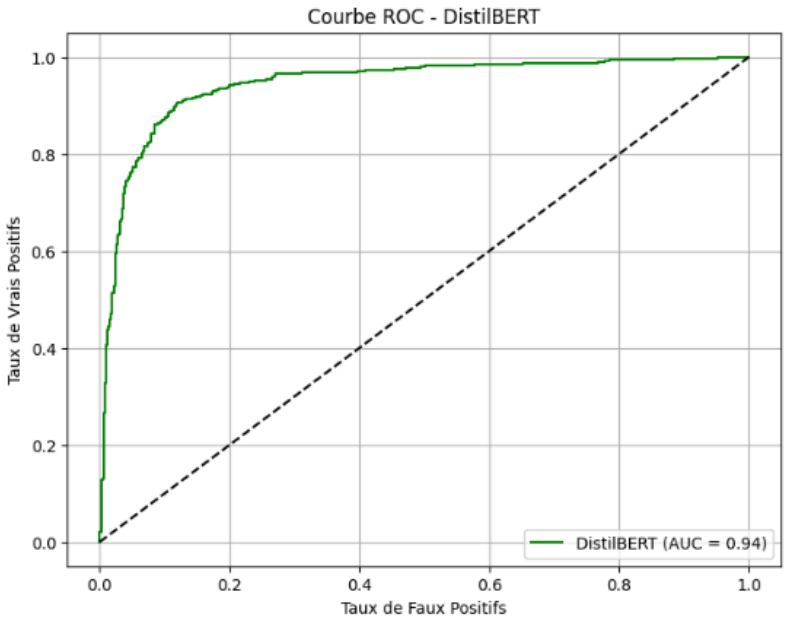


2. DistilBERT fine-tuné

- Tokenisation via `DistilBertTokenizer`.
- Fine-tuning complet du modèle `DistilBertForSequenceClassification` avec la librairie Hugging Face Transformers.
- Évaluation par la méthode `Trainer` avec calcul automatique de l'accuracy et du F1-score.
- AUC obtenu à partir des probabilités prédites.

La métrique principale choisie est l'AUC, permettant de juger la capacité du modèle à séparer les classes sans dépendre d'un seuil de décision fixe. Elle est complétée par la précision et le F1-score pour mesurer la performance globale du modèle.

Le fine-tuning de DistilBERT a été effectué avec un taux d'apprentissage de 2e-5, un batch size de 16, et sur 3 epochs, conformément aux bonnes pratiques courantes.



Une synthèse des résultats

Les performances observées sur le jeu de test sont :

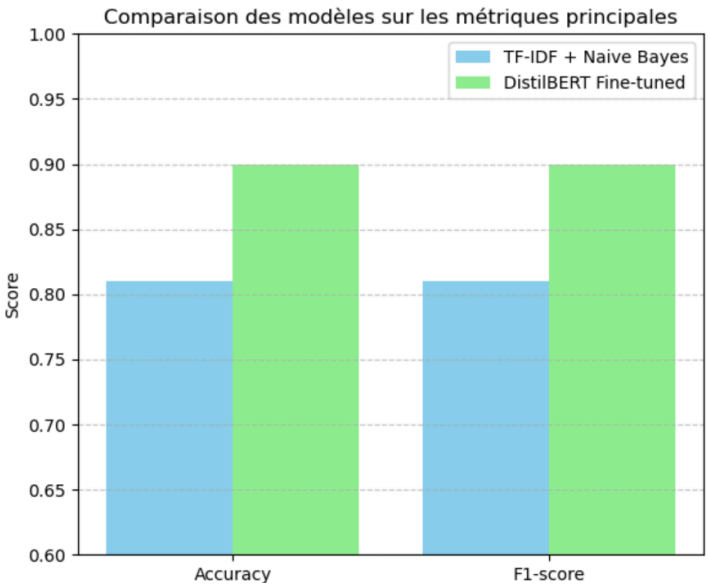
TF-IDF + Naive Bayes

- Accuracy : ~81 %
- F1-score : ~0.81
- AUC : ~0.88

DistilBERT Fine-tuné

- Accuracy : ~90 %

- F1-score : ~0.90
- AUC : ~0.95



La courbe ROC du modèle DistilBERT est significativement supérieure à celle du modèle classique, avec une meilleure séparation entre classes positives et négatives.

En résumé :

- L'approche TF-IDF + Naive Bayes reste rapide et efficace pour des tâches simples, mais elle ignore le contexte des mots.
- L'approche DistilBERT, bien que plus coûteuse, capte les nuances sémantiques du texte et offre des performances supérieures, justifiant son usage dans des cas critiques.

L'analyse de la feature importance globale et locale du nouveau modèle

Les modèles de type Transformer, comme DistilBERT, ne s'appuient pas directement sur des features explicites. Cependant, des méthodes d'explicabilité peuvent être appliquées.

- **Globale** : Une analyse des poids d'attention du modèle ou l'usage d'agrégateurs comme SHAP permet d'identifier des tokens souvent décisifs ("excellent", "boring", "not good", etc.).
- **Locale** : Sur une prédiction donnée, on peut visualiser l'impact de chaque mot via LIME ou SHAP. Ces méthodes permettent d'interpréter le raisonnement du modèle sur un exemple spécifique.

Ces approches d'explicabilité sont essentielles pour évaluer le comportement du modèle sur des cas ambigus ou critiques. Elles renforcent la confiance utilisateur, notamment dans un contexte de décision automatisée.

Les limites et les améliorations possibles

Limites

- Le fine-tuning a été réalisé sur un petit sous-échantillon (2 000 exemples), ce qui limite la généralisation.
- DistilBERT, bien que plus léger que BERT, reste coûteux en ressources pour l'inférence à grande échelle.
- L'interprétabilité reste partielle malgré les outils d'analyse locale.

Améliorations

- Étendre l'entraînement sur le dataset complet pour stabiliser les performances.
- Appliquer une validation croisée pour estimer la robustesse.
- Utiliser des techniques comme LoRA(source2) ou QLoRA pour rendre le fine-tuning plus léger et accessible.
- Intégrer un tableau de bord explicatif pour visualiser les décisions et faciliter l'auditabilité du modèle.

Source 2: https://arxiv.org/abs/2106.09685