Organisation	:	Université Virtuelle du Burkina
Filière	:	Fouilles de Données et Intelligence Artificielle
Niveau	:	Master I
Promotion	:	2022-2023

CONSTRUCTION DE MODÈLES ET LEUR DÉPLOIEMENT

Modèle de Détection de Tweet Suspect

Novembre 2024

Membres du Groupe

Nom	**Prénom**
KIEMA	Wend-denda Arsène
OUEDRAOGO	Mahomed

Chargé du cours : Dr KABORE Kader

1. Introduction

Dans un monde où les interactions en ligne se multiplient, il est de plus en plus important de surveiller et d'identifier les discours potentiellement dangereux, tels que les menaces, le terrorisme et l'intimidation. Les réseaux sociaux, forums et autres plateformes de communication peuvent devenir des vecteurs de discours haineux ou de menaces. Afin de protéger les utilisateurs et de garantir un environnement en ligne sûr, la détection automatisée de discours suspect est essentielle.

L'objectif de ce projet est de développer un modèle de machine learning capable de classer un texte comme "Suspect" ou "Non Suspect" en analysant son contenu. En plus de construire un modèle de classification robuste, ce projet vise à explorer différents algorithmes de classification, et à déployer une API pour une utilisation pratique.

2. Méthodologie

2.1. Exploration et Prétraitement des Données

Les étapes suivantes ont été réalisées pour préparer les données :

- **Chargement des données** : Un dataset contenant des messages textuels étiquetés comme "Suspect" (1) ou "Non Suspect" (0) a été utilisé.
- **Analyse exploratoire** : Visualisation de la répartition des classes pour identifier un fort déséquilibre (environ 85% de la classe "Suspect" contre 15% pour "Non Suspect").
- **Prétraitement du texte** : Suppression des caractères spéciaux, conversion en minuscules, suppression des stop words et tokenisation des mots. Cette étape est cruciale pour réduire le bruit dans les données et améliorer la performance des modèles.

2.2. Embeddings et Représentation des Données

Deux types de représentations de texte ont été explorés :

- **TF-IDF** : Utilisé pour les modèles traditionnels comme la régression logistique, le Random Forest et le SVM. TF-IDF permet de transformer le texte en vecteurs numériques basés sur la fréquence et la pertinence des mots.
- **BERT** : Un modèle de deep learning pré-entraîné qui génère des embeddings contextuels pour chaque mot, permettant une meilleure compréhension des relations et des nuances dans le texte.

2.3. Gestion de l'Équilibre des Classes

Le dataset présentait un déséquilibre marqué entre les classes suspectes et non suspectes. Pour pallier cela, un sous-échantillonnage de la classe majoritaire a été effectué, de manière à équilibrer les classes. Cette méthode permet de réduire le biais induit par la classe majoritaire lors de l'entraînement.

2.4. Sélection des Modèles et Entraînement

Les modèles suivants ont été sélectionnés et entraînés :

- **Régression Logistique** : Modèle simple et rapide, efficace pour les données textuelles basiques.
- **Random Forest** : Modèle d'ensemble basé sur des arbres de décision, adapté aux données déséquilibrées.
- **Support Vector Machine (SVM)** : Modèle de classification efficace pour les données textuelles avec un séparateur maximal.
- **BERT** : Modèle de deep learning puissant pour le traitement du langage naturel, capable de capturer les nuances contextuelles du texte.

Chaque modèle a été entraîné en utilisant la validation croisée pour éviter le surapprentissage et pour évaluer la robustesse.

2.5. Évaluation des Modèles

Les métriques utilisées pour évaluer les modèles incluent :

- **Accuracy** : Proportion d'échantillons correctement classifiés.
- **F1-score** : Moyenne harmonique de la précision et du rappel, robuste dans les cas de déséquilibre.
- **Matrice de confusion** : Visualisation des classifications correctes et incorrectes pour chaque classe.
- **AUC-ROC** : Mesure de la capacité du modèle à distinguer les classes.

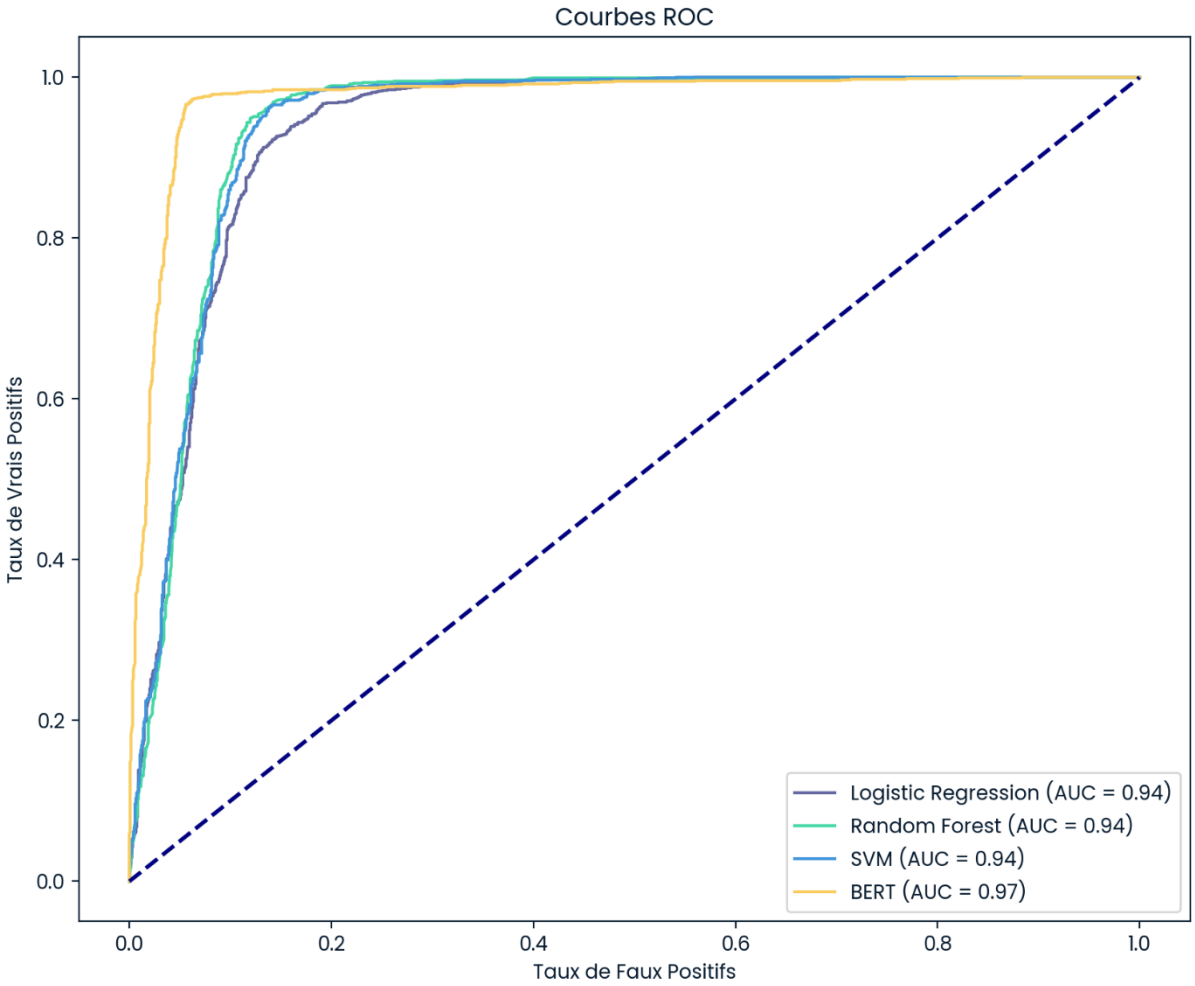
Les modèles ont été optimisés vie la recherche de grille et de la technique de cross validation.

3. Résultats

Performance des Modèles

Les performances des différents modèles sont résumées ci-dessous :

Modèle	Accuracy	Précision (Classe 0)	Rappel (Classe 0)	F1-Score (Classe 0)	Précision (Classe 1)	Rappel (Classe 1)	F1-Score (Classe 1)	AUC
Logistic Regression	0.89	0.93	0.84	0.88	0.86	0.93	0.89	0.94
Random Forest	0.91	0.97	0.84	0.9	0.86	0.98	0.91	0.94
SVM	0.9	0.98	0.81	0.89	0.84	0.98	0.91	0.94
BERT	0.95	0.96	0.95	0.95	0.95	0.96	0.95	0.97



Les résultats montrent que le modèle BERT offre les meilleures performances en termes de précision, rappel, F1-score, et AUC (0.97), surpassant les modèles traditionnels dans la classification des tweets suspects. La courbe ROC, illustrée ci-dessus, montre que BERT est capable de capturer davantage de vrais positifs avec un faible taux de faux positifs par rapport aux autres modèles, ce qui le rend idéal pour cette tâche de détection de discours suspect.

Les modèles Random Forest et SVM ont également démontré de bonnes performances, avec des AUC de 0.94, mais leur capacité à capturer la classe minoritaire est inférieure à celle de BERT. La régression logistique est légèrement en retrait, bien qu'elle reste performante avec une AUC de 0.94.

4. Discussion sur les Limites

- Bien que le sous-échantillonnage ait permis de résoudre le problème de déséquilibre des classes, cette méthode a réduit le nombre d'exemples de la classe majoritaire, ce qui peut entraîner une perte d'information. Une alternative pourrait être la collecte de données pour augmenter la classe minoritaire.

- Variabilité des Données Textuelles : Les messages textuels peuvent varier en longueur et en style, ce qui complique la tâche de détection.
- Modèles Traditionnels : Les modèles comme la régression logistique et le SVM sont limités en termes de capacité de représentation du contexte.

5. Suggestions d'Améliorations

- **Augmentation des Données** : Obtenir davantage de données annotées pourrait améliorer les performances des modèles.
- **Amélioration des performances :** Une approche ensembliste combinant les prédictions de plusieurs modèles, tels que BERT et SVM, pourrait offrir des performances encore plus robustes.
- **Utilisation d'un Serveur GPU** : Entraîner BERT sur un serveur GPU pourrait accélérer le processus et permettre d'expérimenter davantage d'hyperparamètres.

Conclusion

Ce projet a permis de développer un modèle de machine learning performant pour la détection de discours suspect, avec des performances très satisfaisantes.

Le modèle BERT s'avère être le plus performant pour la détection des tweets suspects, en obtenant un score AUC de 0.97. Bien que les autres modèles traditionnels, comme Random Forest et SVM, soient également efficaces, BERT offre une meilleure capacité à distinguer les tweets suspects, ce qui en fait un choix optimal pour cette application.