Plan prévisionnel

Dataset retenu
Ayant le choix de se baser sur un projet précédent dans le parcours de formation, j'ai opté pour le projet 5 et l'utilisation du dataset de Stackoverflow, riche en milliers de questions, réponses et commentaires. Dans ce projet, j'avais choisi l'algorithme du Multinomial Naïve Bayes pour réaliser une prédiction multiclasse avec l'aide de OneVsRestClassifier. L'objectif étant de pouvoir prédire les catégories / tags associés aux questions.

Modèle envisagé
Dans ce nouveau projet, j'ai décidé d'adopter le modèle XLNet pour sa capacité à comprendre les relations complexes entre les mots dans un texte grâce à son architecture de Transformer bidirectionnel. Des études récentes ont montré que XLNet surpasse souvent d'autres modèles comme BERT, dans diverses tâches de compréhension de texte.

Son objectif est de modéliser la probabilité conditionnelle de séquences de mots, en prenant en compte toutes les permutations possibles de leur ordre. Dans notre projet, visant à classifier des questions de Stackoverflow, l'utilisation de XLNet devrait permettre une identification plus précise des intentions et des catégories des questions posées, grâce à sa capacité à saisir la complexité du langage naturel.

Références bibliographiques
Article de recherche publié sur le site Arxiv:

- Attention Is All You Need (premier papier sur les transformers)
    - source: https://arxiv.org/pdf/1706.03762.pdf
- XLNet: Generalized Autoregressive Pretraining (papier expliquant le fonctionnement de XLNET)
    - source: https://arxiv.org/pdf/1906.08237.pdf

Explication de votre démarche de test du nouvel algorithme (votre preuve de concept)
Je vais d'abord établir une méthode de baseline qui sera l'algorithme Multinomial Naïve Bayes, déjà choisi dans un projet précédent. Ensuite, je nettoierai le jeu de données (lemmatisation, tokenisation, suppression des stopwords et outliers, vectorisation).

Après, je procéderai à une analyse exploratoire et je mettrai en œuvre MultiLabelBinarizer sur les 100 tags les plus courants. Étant donné que Stackoverflow comprend des milliers de catégories, chacune ayant une représentativité variable, cette étape permettra de convertir les étiquettes multiples en une représentation binaire.

Puis diviserait le dataset de Stackoverflow en ensembles d'entraînement et de test. Après avoir entraîné et testé le modèle Multinomial Naïve Bayes, je mettrai en œuvre le modèle XLNet et le testerai de manière similaire.

J'utiliserai les métriques suivantes pour la comparaison: **précision, recall, F1-Score, temps d'entraînement.** L'accuracy n'est pas adaptée dans ce cas, car il s'agit d'une prédiction multi-classe avec cinq classes. Ainsi, si une seule des classes est prédite de manière incorrecte, cela entraînerait une accuracy de 0. Enfin, je créerai une interface graphique simple pour interroger les deux modèles et comparer leurs prédictions.