# **Partie 1 :** modèle pour identifier les noms de personne dans du texte

**Résumé des Expérimentations :**

Dans ce projet, nous avons exploré deux modèles puissants : BERT (bert-base-multilingual-cased) et Camembert (camembert-base). L'expérimentation a impliqué des ajustements, notamment le gel sélectif des couches des modèles, afin d'équilibrer l'apprentissage pré-établi avec l'adaptation spécifique à nos données. Nous avons testé plusieurs solutions pour améliorer les performances du modèle, en modifiant les hyperparamètres sur les données d'entraînement. Ensuite, nous avons effectué une évaluation sur nos données de développement et de test pour déterminer la performance de notre modèle, et en particulier sa capacité à trouver la bonne étiquette pour nos données. Nous avons utilisé de petits échantillons pour les tests, car le modèle consomme beaucoup de ressources.

**Démarche, solutions et problèmes :**

Nous avons réalisé une exploration approfondie et un ajustement fin du modèle de traitement du langage naturel basé d'abord sur Camembert, puis sur BERT, ce dernier ayant été écarté pour son manque d'efficacité. Nous avons utilisé des techniques d'apprentissage par transfert, gelant une partie significative des couches du modèle pour préserver les connaissances pré-entraînées. Nous avons également expérimenté le gel de toutes les couches, tout en permettant l'adaptation aux spécificités de notre tâche via les couches de classification, dans le but d'optimiser l'équilibre entre l'exploitation des connaissances générales du modèle et son ajustement à nos besoins spécifiques.

Nous avons testé plusieurs approches dans le but d'améliorer le modèle, telles que l'ajustement des poids sur les étiquettes ou la mise en œuvre de la validation croisée, mais sans succès. En conséquence, nous avons décidé de conserver notre premier modèle de base. Voici les résultats obtenus après plusieurs tests sur différents paramètres.

**Analyses des résultats (ce qui a marché ou non)**

Chaque itération d'entraînement permet d’affiner la précision du modèle, avec une attention particulière portée à l'impact du gel sélectif des couches sur les performances globales. Les résultats ont montré une évolution de la capacité du modèle à traiter efficacement les prompts. (Voir annexe pour plus de résultat)

Une image contenant texte, diagramme, ligne, Tracé

Description générée automatiquement{'train\_runtime': 467.3929, 'train\_samples\_per\_second': 37.763, 'train\_steps\_per\_second': 2.375, 'train\_loss': 0.08679021843918809, 'epoch': 10.0}

Notre modèle a été entraîné pendant environ 7,8 minutes, traitant en moyenne 37,76 échantillons par seconde, avec une performance moyenne de 2,38 étapes par seconde. La perte d'entraînement faible à 0,08679 indique une bonne capacité d'apprentissage sur les données fournies, en 10 époques.

Le modèle affiche une haute précision, un rappel et un score F1 pour la classe 0 (97%), mais des performances nettement inférieures pour la classe 1, avec une précision de 20%, un rappel de 17% et un score F1 de 18%. Malgré cela, la précision globale est de 94%, indiquant une bonne performance générale, principalement due à la domination de la classe 0 dans l'ensemble de données (65562 contre 2681 pour la classe 1).

**Conclusion**  
  
Nous pouvons d'améliorer le modèle en modifiant les hyperparamètres. Cependant, en raison des performances limitées de nos ordinateurs, nous pourrions rencontrer des problèmes de temps de calcul. Par exemple, si nous utilisons l'intégralité de la base de données, un seul cycle d'entraînement peut durer jusqu'à trois jours. De plus, le poids élevé attribué à la classe 1 a tendance à diminuer les performances de notre modèle qui pourrai être réduit en modifiant les poids.

# Partie 2 Prompt chatgpt

# Nous avons d’abord défini et expliquer le sujet a chatgpt : nous avons une base/phrase contenant le titre d’une vidéo qui peut inclure les noms de plusieurs personnes comique ou non. La tâche est d'identifier tous les noms présents dans les phrases. Pour chaque il faudrait identifier si sais le nom d'un comédien ou non. Je voudrais que tu me donne le Nom, Comédien : Oui/Non. Je t'enverrai des titres/bases de vidéo tu me diras s’il y a un comédien ou non

**Expériences :**

Nous avons commencé avec une base de données contenant des titres de vidéos pour identifier les noms de comédiens. Les données comprenaient des titres de vidéos, des indicateurs de noms, et si le titre mentionnait un comédien. L'approche initiale basée sur les indicateurs de présence de noms dans les titres s'est avérée infructueuse, ne parvenant pas à extraire correctement les noms. Une révision de la méthode a été faite en utilisant directement les noms de comédiens fournis dans la base de données, ce qui a permis une extraction et un formatage réussis. Nous avons tenté avec un second fichier beaucoup plus complexe mais chagpt a eu un problème de formatage qui ont entraîné un échec. J’ai aussi essayé avec des vidéos pris sur internet et chat gpt se base sur les connaissances préexistantes et son entraînement pour reconnaître des noms connus dans le domaine de la comédie ou du spectacle

**Prompts, Itérations et Raisonnement :**

Le prompt initial demandait l'identification des noms de comédiens dans les titres de vidéos, avec un format spécifique de réponse. Après le premier essai infructueux, j'ai ajusté l'approche en me basant sur les noms de comédiens déjà identifiés dans la base de données, plutôt que sur l'analyse de chaque mot du titre. Cette itération a permis d'obtenir des résultats précis, conformes à la demande initiale. L'incapacité de traiter le second fichier a souligné l'importance de la qualité des données et de leur formatage pour une analyse réussie. Pour finir la dernière expérience montre que chatgpt se refaire aussi de ces informations préexistantes et de son entraînement

**Résultat :**

Une image contenant texte, capture d’écran, Police

Description générée automatiquement

Figure : Présence de nom d’un/e comédien/ne dans les titres de film (base de données)

Une image contenant texte, capture d’écran, Police, document

Description générée automatiquement

Figure : présence du nom d’un/e comédien/ne dans des vidéos données

# Annexe

**Model Bert :2 epoch**

Une image contenant texte, capture d’écran, Police, menu

Description générée automatiquement

**Model Camembert 64 batch**

Une image contenant texte, ligne, diagramme, capture d’écran

Description générée automatiquement

{'train\_runtime': 4334.0497, 'train\_samples\_per\_second': 24.669, 'train\_steps\_per\_second': 1.542, 'train\_loss': 0.01033675684000046, 'epoch': 2.0}

Une image contenant texte, capture d’écran, Police, noir

Description générée automatiquement