Ecole Nationale Supérieure d'Informatique et d'Analyse des Systèmes - RABAT

Projet NLP

Filiére : GD

Classification des documents par domaine

Encadré par :

M. TABII Youness

Réalisé par :

Touahri Imane

Meskini Yassir

Année Scolaire 2024/2025

Remerciements

Je tiens à exprimer ma profonde gratitude à Mme Sanae L'Efkihi, chef de filière, pour son soutien constant et ses conseils avisés tout au long de cette année. Son encadrement bienveillant et son engagement envers la réussite des étudiants ont été d'une grande aide et d'une source de motivation.

Je remercie également M. TABII Youness, notre professeur, pour l'excellence de son enseignement et pour nous avoir confié ce projet. Ses connaissances approfondies et son accompagnement pédagogique nous ont permis d'acquérir des compétences essentielles et de mener à bien ce travail avec confiance.

Merci à tous deux pour leur précieux soutien.

Cordialement.

Résumé

Ce projet se concentre sur la classification de documents en utilisant des techniques de traitement automatique du langage naturel (NLP). Nous avons travaillé avec une base de données contenant des documents issus de dix domaines distincts, tels que l'éducation, le sport, l'espace, la gastronomie, etc. La première étape a consisté à entraîner plusieurs modèles de classification, en commençant par le modèle Naïve Bayes, suivi du modèle SVM, puis du modèle BERT et Longformer. .

Ensuite, nous avons procédé à l'extraction de texte à partir de divers formats de fichiers, incluant des documents PDF, des PDF scannés .

Une fois l'extraction de texte réalisée, nous avons testé chacun des quatre modèles et comparé leurs performances respectives en termes de précision et de pertinence pour la classification des documents.

Finalement, nous avons conclu en analysant les résultats obtenus pour chaque modèle et en identifiant le modèle le plus adapté à notre base de données.

Abstract

This project focuses on document classification using natural language processing (NLP) techniques. We worked with a database containing documents from ten distinct fields, such as education, sport, space, gastronomy, etc. The first step consisted of training several classification models, starting with the Naïve Bayes model, followed by the SVM model, then the BERT and Longformer models.

Once the text extraction was carried out, we tested each of the four models and compared their respective performances in terms of accuracy and relevance for document classification.

Finally, we concluded by analyzing the results obtained for each model and identifying the model best suited to our database.

Table de matières

Table des figures

Introduction Générale

Dans un contexte où la gestion et l'organisation de grandes quantités de documents sont cruciales, la classification automatique des documents devient un défi majeur pour de nombreuses industries. Ce projet explore différentes techniques de classification de documents à l'aide de modèles de traitement automatique du langage naturel (NLP).

Nous avons utilisé une base de données composée de documents provenant de dix domaines distincts, tels que l'éducation, le sport, l'espace, la gastronomie, et bien d'autres.

Dans un premier temps, nous avons entamé un processus d'entraînement en testant plusieurs modèles de classification, à commencer par le modèle **Naïve Bayes**, suivi du modèle **SVM**, puis du modèle **BERT** et enfin du modèle **Longformer**.

L'étape suivante a impliqué l'extraction de texte à partir de divers formats de fichiers, incluant des documents PDF, des PDF scannés. Cette extraction a permis de normaliser les données et de les préparer pour les différentes étapes de classification.

Enfin, les performances des quatre modèles ont été évaluées et comparées, ce qui nous a permis d'analyser leur efficacité respective dans la classification des documents selon les différents domaines. Cette analyse nous a conduit à identifier le modèle le plus performant et à formuler des recommandations pour de futures applications.

Chapitre 1

Description et visualisation des données

Ce chapitre se concentre sur la description et la visualisation du dataset utilisé pour l'entraînement des modèles de classification de documents. L'objectif était de comprendre la structure et les caractéristiques des données afin d'optimiser les étapes suivantes du projet. Nous avons créé un dataset équilibré, composé de documents provenant de dix domaines distincts . Chaque domaine contient exactement 40 documents, garantissant ainsi une distribution homogène des données.

Ce chapitre présente d'abord les détails de la création de ce dataset, puis explore diverses visualisations qui permettent d'analyser la distribution et les tendances des données à travers des outils tels que des nuages de mots, des histogrammes et des graphiques de fréquence des termes.s.

1.1 Présentation des données initiales

Initialement, nous avons utilisé un dataset créé par la concaténation de deux ensembles distincts, "10-data" et "bbc". Cependant, cette approche a engendré un déséquilibre dans la distribution des classes, certaines contenant plus de fichiers que d'autres, ce qui a rendu le dataset non équilibré.

```
Label Counts:
9    611
0    610
7    517
2    486
1    100
3    100
5    100
4    100
6    100
8    100
Name: count, dtype: int64
```

FIGURE 1.1 – Le désiquilibre de la data

Pour remédier à cela, nous avons équilibré le nombre de fichiers par domaine en utilisant des techniques de data augmentation. Plus précisément, nous avons utilisé **nltk** pour enrichir les données. Cet outil ne se contente pas de dupliquer les fichiers existants, mais remplace certains mots par leurs synonymes, créant ainsi de nouvelles versions augmentées des documents.

```
Label
0.0    611
1.0    611
2.0    611
3.0    611
4.0    611
5.0    611
6.0    611
7.0    611
8.0    611
9.0    611
Name: count, dtype: int64
```

FIGURE 1.2 – L'équilibre de la data

Bien que nous ayons équilibré le nombre de fichiers dans chaque domaine, nous avons observé un déséquilibre significatif dans le nombre de mots par document selon les domaines. Cette inégalité est mise en évidence grâce à la visualisation suivante (voir l'histogramme ci-dessus), qui montre la distribution des longueurs de documents par classe. On peut voir que certaines classes ont des documents beaucoup plus longs que d'autres, ce qui peut entraîner des défis supplémentaires pour l'entraînement du modèle.

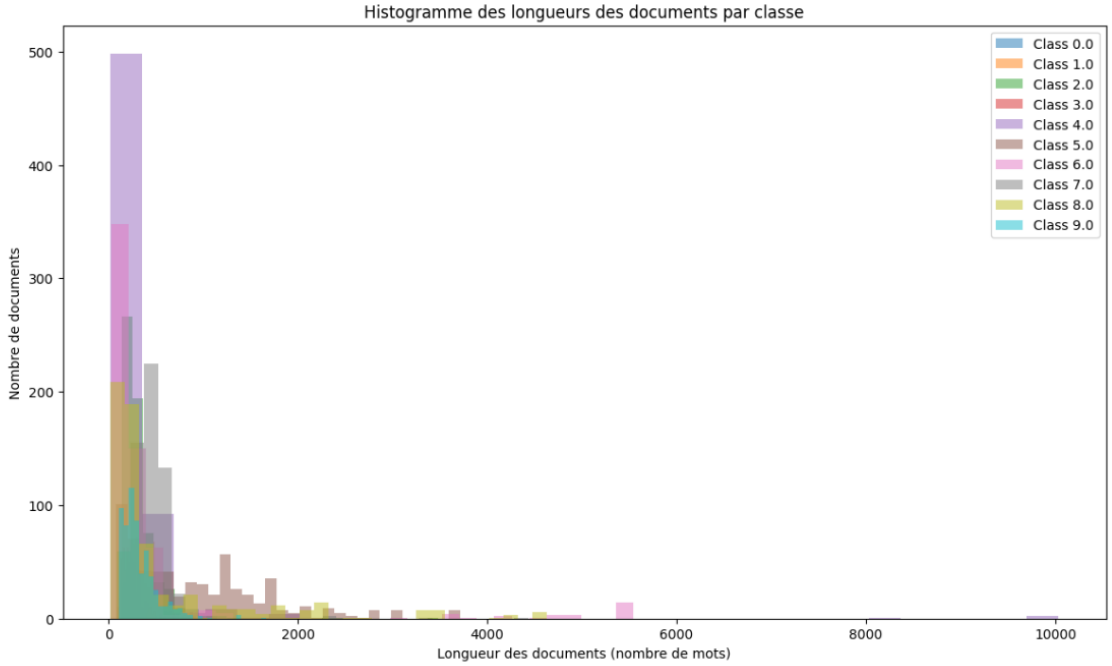

FIGURE 1.3 – Histogramme des longueurs des documents de la data initiale

1.2 Génération de la nouvelle base de données

La création d'un dataset personnalisé a été motivée par la nécessité d'obtenir un équilibre optimal entre les différentes catégories ou domaines présents dans le modèle de classification. En générant un nouveau dataset comprenant 10 domaines distincts, chacun contenant exactement 40 documents, l'objectif était de garantir une distribution homogène des données. Cette approche permet de réduire les risques de biais dans l'apprentissage automatique, en s'assurant que chaque domaine est représenté de manière égale. Cela permet aussi d'éviter que certains domaines ne dominent le processus de classification en raison d'une surreprésentation, contribuant ainsi à améliorer la robustesse et la précision du modèle.

1.2.1 Description de notre base de données

Détails sur les domaines :

Les domaines choisis pour ce projet sont variés et couvrent des secteurs d'importance dans le contexte actuel. Les 10 domaines sélectionnés sont : **Education, Entertainment, Fashion, Finance, Food, Healthcare, Law, Real Estate, Sport, Technology**. Ces domaines ont été choisis en raison de leur diversité et de leur pertinence dans la classification de documents textuels. Chacun de ces domaines représente une catégorie distincte et bien définie, permettant d'entraîner le modèle à effectuer des distinctions fines entre des sujets très différents, tout en restant en lien avec des sujets d'actualité et des enjeux sociétaux.

```python
import pandas as pd

# Supposons que df est votre DataFrame
valeurs_uniques = df['Domain_Label'].unique()

# Afficher les valeurs uniques
print(valeurs_uniques)
```

```
['Education' 'Entertainment' 'Fashion' 'Finance' 'Food' 'Healthcare' 'Law'
 'Real Estate' 'Sport' 'Technology']
```

FIGURE 1.4 – Les différents domaines de notre data

Structure de la nouvelle base de données :

Les documents ont été générés à partir de sources fiables et vérifiés pour assurer leur authenticité et leur pertinence. Pour chaque domaine, 40 documents ont été collectés, garantissant ainsi un total de 400 documents dans le dataset. Chaque document a été soigneusement examiné pour s'assurer qu'il correspondait bien à la catégorie à laquelle il appartenait, et des mesures ont été prises pour maintenir une qualité de contenu homogène. Cette méthode de génération et de vérification a permis d'obtenir une base de données structurée, équilibrée et représentative des différents domaines.

1.2.2 Visualisation des données

Pour mieux comprendre la structure du dataset et explorer les caractéristiques des documents, plusieurs visualisations ont été réalisées :

Nuage de mots (Word Cloud) : Un nuage de mots a été généré pour chaque domaine afin de visualiser les termes les plus fréquents présents dans les documents. Cette représentation graphique permet de repérer rapidement les mots-clés et les thèmes dominants dans chaque catégorie.

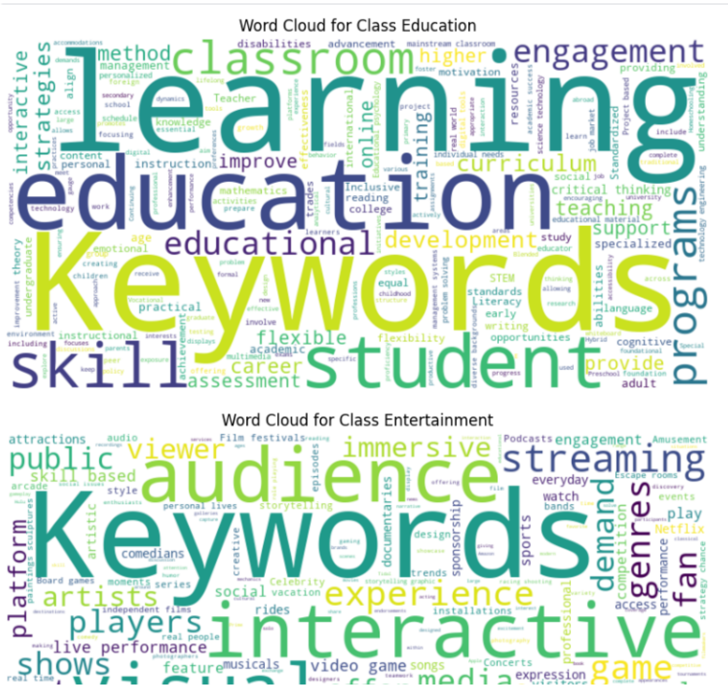

FIGURE 1.5 – Nuage de mots

Histogramme des longueurs des documents : Un histogramme a été tracé pour illustrer la distribution des longueurs des documents, en termes de nombre de mots. Cela permet de vérifier la cohérence du dataset et d'identifier d'éventuelles anomalies, comme des documents trop courts ou trop longs.

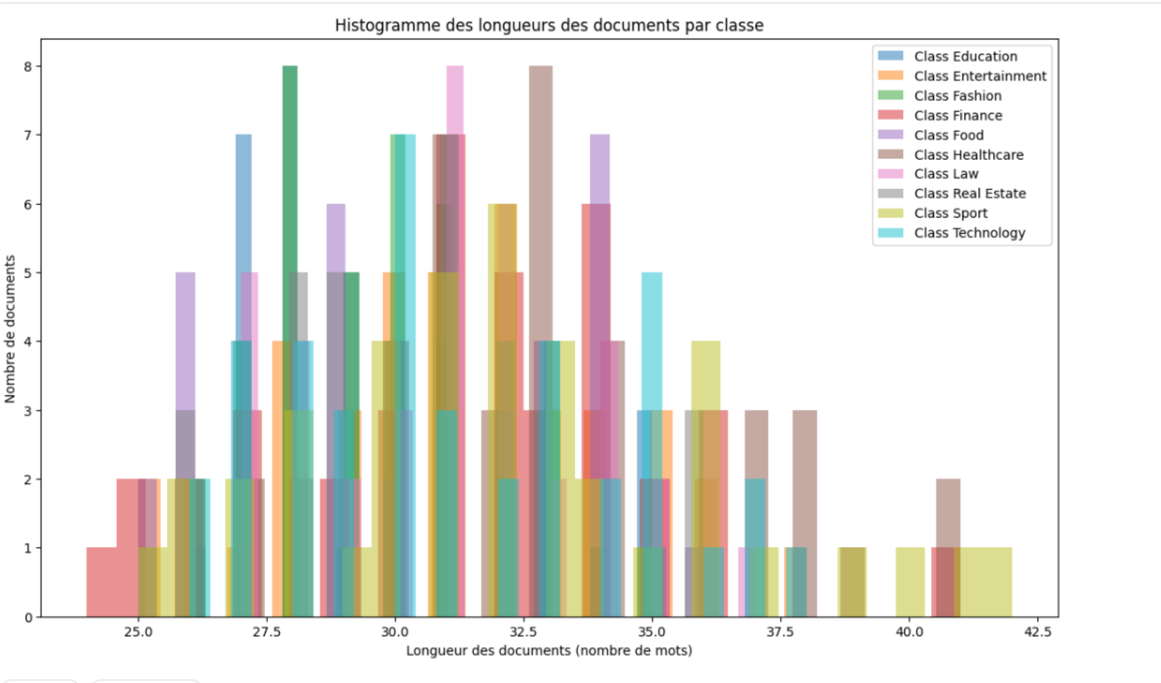

Figure 1.6 – Histogramme des longueurs des documents

Graphique de fréquence des termes (TF) : Un graphique de la fréquence des termes (Term Frequency) a été créé pour analyser la distribution des mots les plus utilisés à travers les documents. Cela aide à comprendre quels termes apparaissent fréquemment dans l'ensemble du dataset.

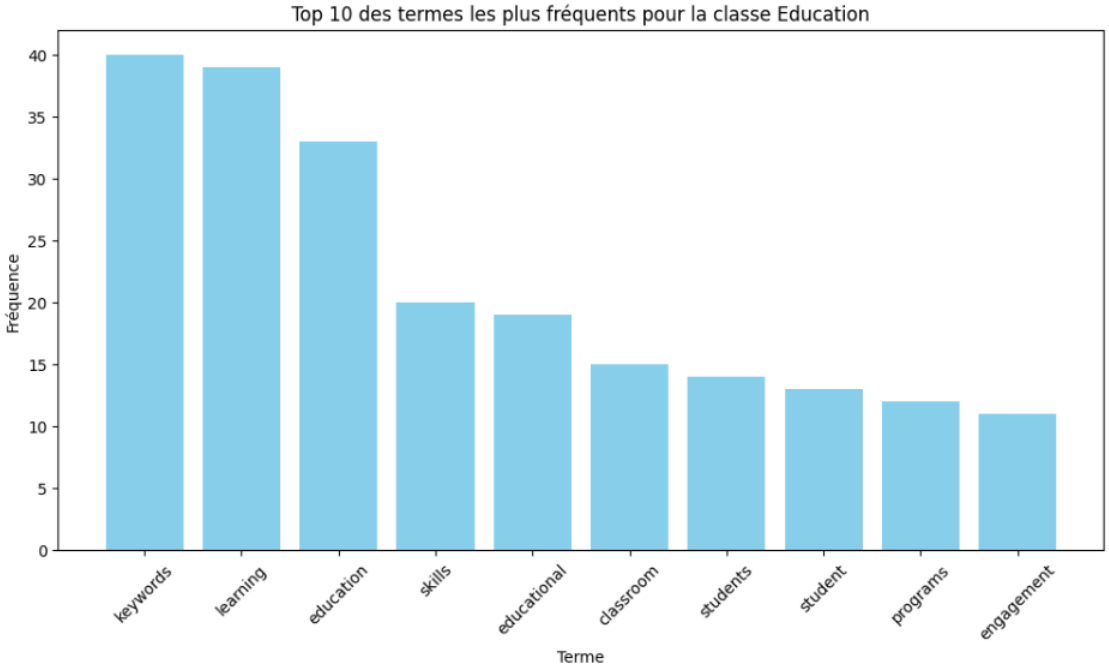

FIGURE 1.7 – Graphique de fréquence des termes (TF)

1.3 Conclusion

Ce chapitre décrit la création d'un dataset équilibré composé de 10 domaines (Education, Entertainment, Fashion, Finance, Food, Healthcare, Law, Real Estate, Sport, Technology), chacun avec 40 documents. L'objectif était d'assurer une distribution homogène des données. Des visualisations, telles qu'un nuage de mots, un histogramme des longueurs de documents, un graphique de fréquence des termes et une matrice de corrélation, ont été utilisées pour analyser les caractéristiques du dataset et préparer le modèle de classification.

Chapitre 2

Entrainement des modèles

2.1 Modèle Naïve Bayes

2.1.1 Dédinition du modèle

Le modèle Naïve Bayes est un classifieur probabiliste basé sur le théorème de Bayes avec l'hypothèse d'indépendance entre les caractéristiques. En d'autres termes, il suppose que la présence ou l'absence d'une caractéristique particulière dans une classe est indépendante de la présence ou de l'absence de toute autre caractéristique. Cette simplicité le rend efficace pour les tâches de classification de texte.

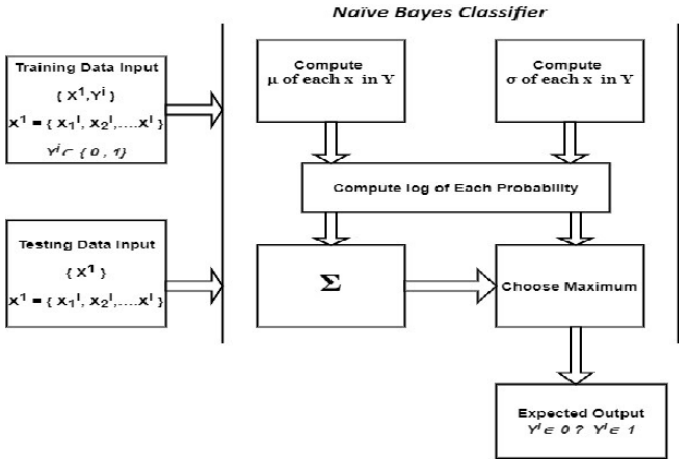

FIGURE 2.1 – architecture du Naïve Bayes

2.1.2 Entrainement du modèle

Pour entraîner le modèle Naïve Bayes, nous avons d'abord transformé les textes en vecteurs numériques en utilisant la méthode TF-IDF (Term Frequency-Inverse Document Frequency). Cette méthode permet de quantifier l'importance de chaque mot dans un document par rapport à l'ensemble du corpus. Ensuite, le classifieur Multinomial Naïve Bayes a été ajusté sur les données d'entraînement, apprenant ainsi les probabilités conditionnelles des mots pour chaque classe.

2.1.3 Resultat du modèle

Le modèle Naïve Bayes a obtenu une précision de test de 0.97 . Le rapport de classification montre des scores élevés de précision, rappel et F1 pour toutes les classes, ce qui indique une performance globale solide. Malgré cette précision élevée, le modèle n'est pas en surapprentissage (overfitting). En effet, il a été testé sur des données non vues (unseen data) et a maintenu une performance élevée. Cela signifie que le modèle a réussi à généraliser les connaissances acquises lors de l'entraînement pour faire des prédictions précises sur de nouvelles données.

2.2 Modèle SVM

2.2.1 Dédinition du modèle

Le modèle SVM est un algorithme d'apprentissage supervisé qui analyse les données pour la classification et la régression. Il fonctionne en trouvant l'hyperplan optimal qui sépare les données de différentes classes avec la plus grande marge possible. Les SVM sont efficaces dans les espaces de haute dimension et sont polyvalents grâce à l'utilisation de différentes fonctions noyau.

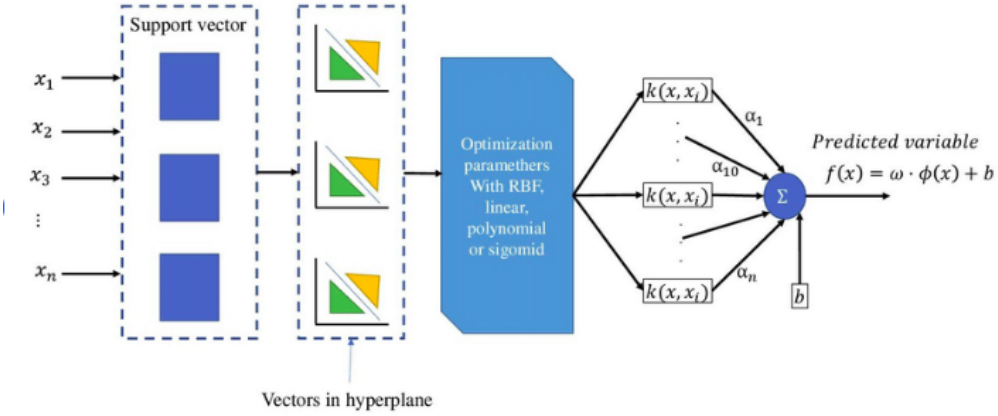

FIGURE 2.2 – architecture d'SVM

2.2.2 Entrainement du modèle

Comme pour le modèle Naïve Bayes, les textes ont été vectorisés en utilisant la méthode TF-IDF. Le classifieur SVM a ensuite été entraîné sur ces vecteurs pour trouver l'hyperplan qui sépare au mieux les différentes classes de textes. Nous avons utilisé une SVM linéaire pour cette tâche.

2.2.3 Resultat du modèle

Le modèle SVM a atteint une précision de test de 1.0 . Tous les scores de précision, rappel et F1 sont parfaits pour chaque classe dans le rapport de classification. Cette performance exceptionnelle sur les données de test indique que le modèle a appris les caractéristiques distinctives de chaque classe de manière efficace. Malgré la précision de 1.0, le modèle n'est pas en surapprentissage, car il a été capable de généraliser avec succès sur des données non vues. Cela suggère que le modèle SVM a capturé les structures sous-jacentes des données sans mémoriser les exemples spécifiques de l'ensemble d'entraînement.

2.3 Modèle Bert

2.3.1 Dédinition du modèle

BERT (Bidirectional Encoder Representations from Transformers) est un modèle de langage développé par Google qui utilise une architecture Transformer bidirectionnelle. Il est pré-entraîné sur un vaste corpus de textes et est capable de comprendre le contexte des mots en considérant les deux sens (gauche et droite) simultanément. Cela le rend particulièrement puissant pour diverses tâches de traitement du langage naturel, y compris la classification de texte.

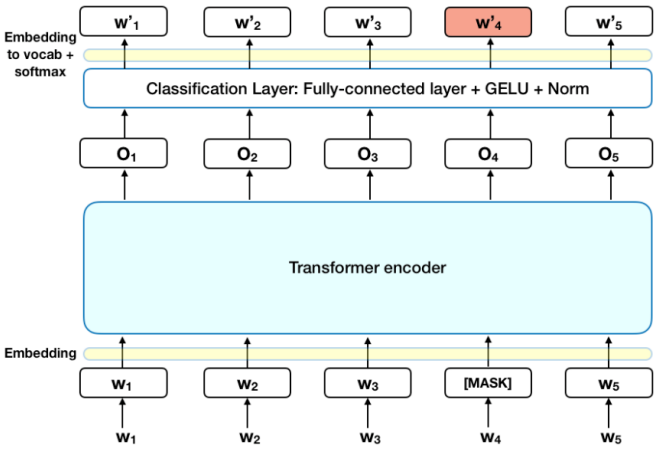

FIGURE 2.3 – architecture de Bert

2.3.2 Entrainement du modèle

Le modèle BERT a été fine-tuné sur notre jeu de données de classification. Les textes ont été tokenisés en utilisant le tokenizer spécifique de BERT, qui gère les sous-mots pour mieux capturer le sens. Le modèle a été entraîné sur trois époques, avec une optimisation de la fonction de perte de classification pour ajuster les poids du réseau neuronal.

2.3.3 Resultat du modèle

Au cours de l'entraînement, le modèle BERT a montré une amélioration constante de la perte et de la précision. Après la troisième époque, il a atteint une précision de validation de 0.92 . Cette haute précision sur des données non vues indique que le modèle généralise bien et n'est pas en surapprentissage. Bien qu'il n'atteigne pas les 1.0 comme le modèle SVM, BERT offre une compréhension plus profonde du contexte linguistique, ce qui peut être avantageux pour des textes plus complexes ou ambiguës.

2.4 Modèle Longformer

2.4.1 Dédinition du modèle

Longformer est une extension de l'architecture Transformer conçue pour gérer efficacement de longues séquences de texte. Contrairement à BERT, qui est limité à des séquences de 512 tokens, Longformer peut traiter des séquences beaucoup plus longues en utilisant une attention locale et globale, ce qui le rend adapté pour des documents volumineux.

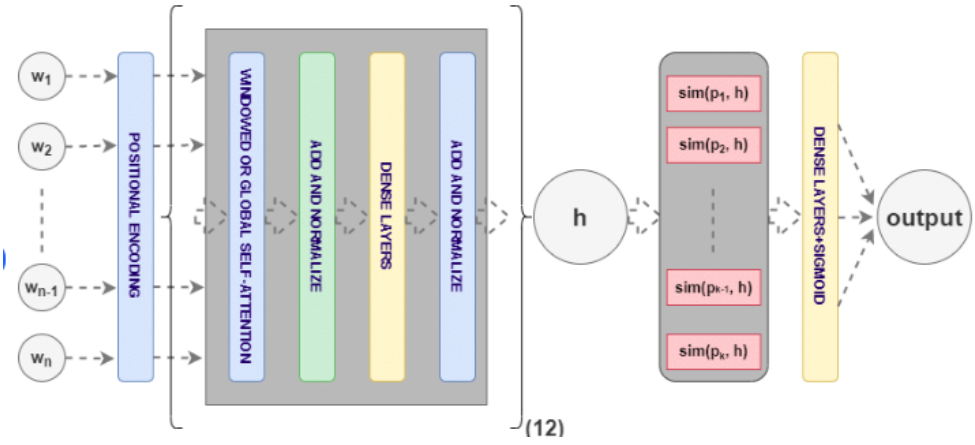

FIGURE 2.4 – architecture de Longformer

2.4.2 Entrainement du modèle

Le modèle Longformer a été entraîné de manière similaire à BERT, avec une tokenisation adaptée pour gérer de longues séquences de texte. Nous avons entraîné le modèle sur trois époques, en veillant à optimiser la fonction de perte de classification tout en gérant efficacement les ressources computationnelles grâce à l'attention optimisée de Longformer.

2.4.3 Resultat du modèle

Le modèle Longformer a atteint une précision de validation de 1.0 après la troisième époque. Cette performance impeccable sur des données non vues témoigne de sa capacité à généraliser efficacement sans surapprentissage. Le modèle a su capturer les relations complexes dans de longs textes, ce qui lui a permis de faire des prédictions précises sur de nouvelles données. L'utilisation de l'attention locale et globale a permis au modèle de se concentrer sur les informations pertinentes tout en traitant de grandes quantités de données.

Chapitre 3

Extraction du text

3.1 Préparation du Texte Extrait

Dans cette partie du projet, nous avons procédé à l'extraction du texte à partir de documents PDF cibles, stockant le contenu obtenu dans la variable input-text. Cette étape est essentielle pour simuler l'utilisation réelle des modèles sur des données non vues auparavant. Le texte extrait représente des documents authentiques susceptibles d'être classifiés dans une application pratique. La préparation du texte implique également un prétraitement adapté pour chaque modèle, afin d'assurer une compatibilité avec les processus de tokenisation ou de vectorisation utilisés lors de l'entraînement.

3.2 Prédictions des Modèles

Une fois le texte extrait prêt, nous l'avons soumis aux quatre modèles de classification entraînés : Naïve Bayes, SVM, BERT et Longformer. Pour les modèles Naïve Bayes et SVM, le texte a été transformé en vecteurs TF-IDF en utilisant le même vectoriseur que lors de l'entraînement, garantissant ainsi la cohérence des caractéristiques utilisées. Les modèles BERT et Longformer ont nécessité une tokenisation spécifique, adaptée à leurs architectures respectives et respectant les limites de longueur maximale (512 tokens pour BERT et 1024 pour Longformer). Chaque modèle a ensuite effectué une prédiction sur le texte, identifiant le domaine correspondant. Cette démarche

a permis une comparaison directe des prédictions fournies par chaque modèle sur le même jeu de données, offrant une évaluation concrète de leurs performances en conditions réelles et mettant en évidence leurs points forts et limites dans le traitement de textes extraits de documents PDF.

Chapitre 4

Evaluation des Modèles de Classification et Analyse Comparée

4.1 Méthodologie d'Évaluation

Pour évaluer les performances des modèles de classification entraînés (Naïve Bayes, SVM, BERT et Longformer), nous avons testé chacun d'eux sur plusieurs documents PDF non vus auparavant. Ces documents représentent des données réelles et variées, permettant de mesurer la capacité des modèles à généraliser et à maintenir leur précision en dehors de l'ensemble de données d'entraînement. Les textes extraits des PDF ont été soumis à chaque modèle, et les prédictions ont été comparées aux étiquettes attendues pour évaluer leur exactitude.

4.2 Résultats et Interprétation

Les tests sur les documents PDF ont révélé des différences notables dans les performances des modèles :

- **Modèle BERT** :
  - **Observations** : BERT a souvent fourni des résultats incorrects lors de la classification des documents PDF non vus.
  - **Interprétation** : Bien que BERT soit performant sur des données d'entraînement,

sa limitation à des séquences de 512 tokens peut entraîner une perte d'informations cruciales dans des documents plus longs. De plus, si les textes contiennent des structures ou des vocabulaires différents de ceux présents dans l'ensemble d'entraînement, BERT peut avoir du mal à généraliser correctement.

- **Modèles SVM et Naïve Bayes** :
  - **Observations** : Ces modèles ont affiché des performances plus robustes que BERT, avec des résultats majoritairement corrects sur les mêmes documents.
  - **Interprétation** : SVM et Naïve Bayes, basés sur des représentations vectorielles du texte (comme TF-IDF), sont moins sensibles à la longueur du document. Leur simplicité les rend moins susceptibles de surapprendre les données d'entraînement, ce qui améliore leur capacité à généraliser sur des données non vues.
- **Modèle Longformer** :
  - **Observations** : Le Longformer a constamment produit des résultats corrects sur les données non vues.
  - **Interprétation** : Conçu pour gérer efficacement de longues séquences de texte (jusqu'à 4096 tokens ou plus), le Longformer capture davantage de contexte que BERT. Sa capacité à traiter intégralement de longs documents lui permet de saisir des informations clés dispersées dans le texte, améliorant ainsi la précision de ses prédictions sur des documents volumineux.

4.3 Discussion et Conclusion

Les différences de performance entre les modèles peuvent être attribuées à plusieurs facteurs scientifiques :

- **Capacité à Gérer la Longueur du Texte** : *BERT vs. Longformer* : BERT est limité à des séquences de 512 tokens, ce qui le rend moins adapté aux longs documents. Le Longformer, en revanche, est spécifiquement conçu pour traiter de longues séquences grâce à son mécanisme d'attention optimisé. Cette différence architecturale explique

pourquoi le Longformer surpasse BERT sur des documents PDF volumineux.

— **Approche de Représentation du Texte** :
  — *SVM et Naïve Bayes* : Ces modèles utilisent des représentations basées sur la fréquence des mots (comme TF-IDF), ce qui capture les informations globales du texte sans dépendre de la séquence des mots. Cela les rend plus robustes face à la variabilité des documents et moins sensibles aux limites de longueur.
  — *BERT* : Basé sur l'apprentissage profond, BERT apprend des représentations contextuelles des mots, ce qui nécessite des données d'entraînement volumineuses et diversifiées pour bien généraliser. Sur des données non vues avec des structures ou des vocabulaires différents, BERT peut avoir des difficultés.
— **Généralisation et Surapprentissage** :
  — *SVM et Naïve Bayes* : Leur simplicité réduit le risque de surapprentissage. Ils généralisent mieux sur des données non vues car ils ne capturent pas les moindres détails des données d'entraînement, mais plutôt des tendances globales.
  — *BERT* : Avec sa grande capacité de modélisation, BERT peut surapprendre si l'ensemble d'entraînement n'est pas suffisamment diversifié ou volumineux, ce qui affecte sa performance sur des données non vues.
  — *Longformer* : Alliant la puissance des modèles Transformers et la capacité à traiter de longues séquences, le Longformer capture efficacement les informations essentielles sans perdre le contexte, ce qui améliore sa généralisation.

Conclusion générale

Ce rapport a exploré la classification automatique de documents, en détaillant chaque étape, de la préparation des données à l'évaluation des modèles. Le premier chapitre a permis de créer un dataset équilibré en utilisant des techniques d'augmentation des données, et des visualisations ont aidé à comprendre la structure des documents. Ensuite, plusieurs modèles de classification (Naïve Bayes, SVM, BERT, Longformer) ont été testés et comparés pour déterminer leur performance. Les résultats ont montré l'importance d'une préparation soignée des données et d'une évaluation rigoureuse des modèles, fournissant ainsi des bases solides pour de futures applications.