**Institut Supérieur d’Informatique**

**et de Mathématique de Monastir**

**Rapport du projet de traitement de données**

Encadré par Mme :

**Mariem Gzara**

Réalisé par :

**Ibrahim Trabelsi**

**Tasnim Chiba**

**Mazen Sghaier**

**Ahmed Zahaf**

2023/2024

**Sommaire** :

**[I. Introduction](#Introduction)**

a. Contexte du Projet

b. Objectifs Principaux

**II. Aperçu des Données**

a. Source des Données

b. Description des Variables

c. Statistiques descriptives

**III.** **Exploration et Analyse des Données**

1. Prétraitement des Données

a. Nettoyage des Données

b. elimination des “stop-words”

c. Normalisation des Textes (Stemmatisation et Lemmatisation)

2. Représentation Vectorielle des Textes

a. Matrice TF-IDF

1. Vectorisation par Comptage de Mots

3. Réduction de Dimensions

a. Analyse en Composantes Principales (ACP)

i. Interprétation des Composantes Principales

ii. Scree Plot et Choix du Nombre de Dimensions

iii. Visualisation des Données Réduites

4. Modélisation et Classification

a. Clustering

i. K-means : Découpage en Clusters

ii. KNN : Classification basée sur la Proximité

b. Évaluation de la Performance

i. Matrice de Confusion

ii. Courbe ROC

iii. Courbe Précision-Rappel

c. Modèles de Classification

i. Random Forest : Ensembles d'Arbres de Décision

ii. Arbres de Décision : Interprétation des Règles

iii. Réseaux de Neurones : Modélisation Non-linéaire

**IV. Conclusions et Perspectives**

a. Résumé des Résultats

b. Limitations de l'Étude

c. Opportunités pour des Recherches Futures

1. **Introduction**

Le traitement des données joue un rôle fondamental dans l'évolution des technologies de traitement du langage naturel (NLP), catalysant des avancées significatives dans la compréhension et la manipulation du langage humain. Dans cette optique, le projet KALIMAT 1.0 se profile comme une ressource linguistique arabe exceptionnelle, destinée à enrichir la recherche en NLP dédiée à la langue arabe.

**Contexte du Projet**

L'arabe, l'une des langues les plus parlées au monde, requiert des outils spécialisés pour exploiter pleinement son potentiel dans le domaine du Traitement Automatique du Langage Naturel (TALN). Le projet KALIMAT 1.0, dirigé par Mahmoud El-Haj et Rim Koulali, s'érige comme une réponse à ce besoin, offrant une ressource riche et diversifiée conçue spécifiquement pour la recherche en TALN adaptée à l'arabe.

Ces données englobent une variété de catégories, telles que la culture, l'économie, les actualités locales et internationales, la religion et le sport. L'objectif principal de KALIMAT 1.0 est d'exploiter cette source de données diversifiée pour contribuer de manière substantielle au domaine du traitement automatique de la langue arabe.

**Objectifs Principaux**

L'essence même de ce projet réside dans la mise à disposition d'une ressource complète en langue arabe, comprenant 20,291 articles issus du journal omanais Alwatan, accompagnés d'annotations exhaustives. Nous débuterons par une exploration approfondie de KALIMAT 1.0, mettant en lumière sa composition, ses catégories et ses composants.

Ensuite, nous examinerons en détail les différentes étapes du traitement des données, du prétraitement des articles à l'analyse morphologique, en passant par la génération de résumés extractifs et la reconnaissance d'entités nommées. Ces analyses approfondies visent à évaluer la qualité des données et à démontrer leur utilité potentielle dans la recherche en NLP.

En conclusion, ce projet adopte une approche collaborative et stimulante, offrant une contribution significative à la communauté de recherche en NLP et ouvrant la voie à de nouvelles perspectives dans le traitement de la langue arabe.

1. **Aperçu des données**

**a. Source des Données**

KALIMAT 1.0, une ressource linguistique en arabe, se distingue par ses six catégories distinctes : Culture, Économie, Actualités Locales, Actualités Internationales, Religion, et Sports. Accessible librement à la communauté de recherche, cette base de données représente une contribution précieuse à l'avancement de la recherche en Traitement du Langage Naturel (NLP) en arabe. En tant que ressource ouverte, elle offre aux chercheurs la possibilité d'utiliser KALIMAT 1.0 comme référence de qualité supérieure ou comme base de comparaison pour tester et évaluer leurs outils de TAL en arabe. Encourageant la collaboration, cette base de données invite également les chercheurs à apporter des amendements au corpus, favorisant ainsi un échange dynamique d'idées et de contributions dans le domaine du TAL arabe.

**b. Description des Variables**

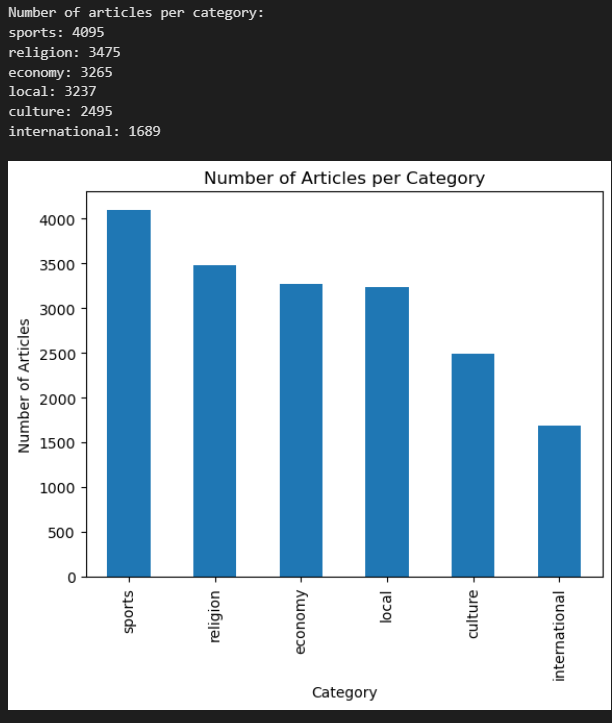
Dans cette section d'analyse des variables de KALIMAT 1.0, nous examinons de près la structure et la composition des articles au sein de chaque catégorie. Notre exploration se concentre sur des aspects cruciaux tels que le nombre de mots par article dans chaque catégorie, ainsi que le nombre total d'articles par catégorie. Il convient de souligner que, lors de cette analyse, les termes les plus fréquemment rencontrés dans chaque catégorie peuvent inclure des "stopwords" et des termes courants qui peuvent ne pas porter un sens sémantique spécifique. En traitement du langage naturel, ces mots sont généralement éliminés ou moins pondérés dans l'analyse textuelle afin de concentrer l'attention sur le contenu informatif réel. Cette démarche permettra d'obtenir une compréhension plus approfondie du contenu et des particularités linguistiques propres à chaque catégorie, tout en contribuant à une analyse plus précise et significative des données au sein de KALIMAT 1.0.

**c. Statistiques descriptives :**

Dans notre exploration, nous avons également visualisé les mots communs spécifiques à chaque catégorie.

*Exploration du Nombre de Mots par Article et par Catégorie :*

L'analyse détaillée du nombre de mots offre un aperçu de la richesse textuelle dans chaque catégorie. Les résultats sont présentés ci-dessous :

*Nombre d'Articles par Catégorie*:

*Mots Communs par Catégorie :*

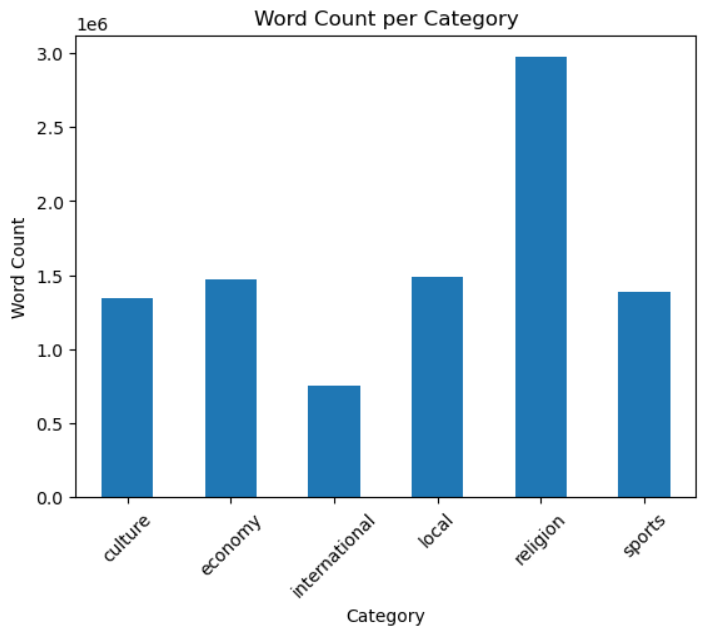
- La catégorie 'Religion' affiche le plus grand nombre de mots, totalisant 2 972 644, soulignant un intérêt marqué pour les sujets religieux.

- 'Local' suit avec 1 484 586 mots, indiquant une attention particulière aux questions locales.

- 'Économie' compte 1 472 513 mots, témoignant d'un intérêt substantiel pour les questions économiques.

- 'Culture' contient 1 344 078 mots, révélant un engagement dans les sujets culturels.

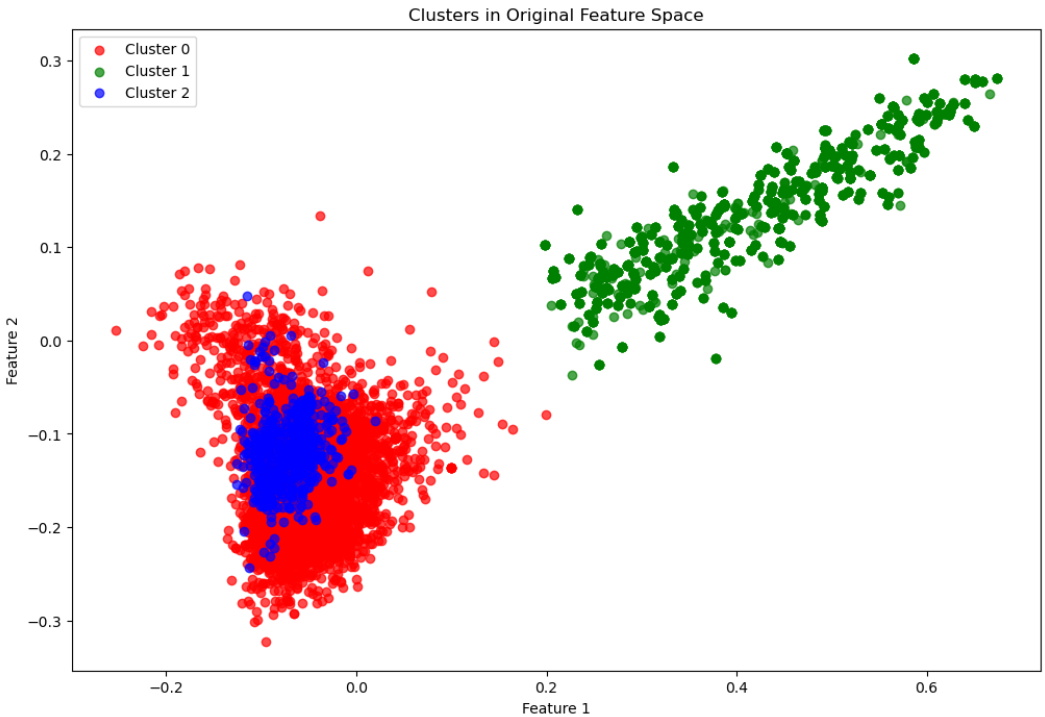
- La catégorie 'Sports' présente 1 388 148 mots, suggérant un intérêt continu pour les événements sportifs.

 - Enfin, la catégorie 'International' affiche 751 815 mots, indiquant une quantité moindre de contenu par rapport aux autres catégories.

En somme, cette analyse des données met en lumière la répartition des mots dans les différentes catégories, mettant en évidence les tendances d'intérêt dans chaque domaine. Ces résultats fournissent des indications cruciales pour la compréhension des sujets prédominants et des centres d'intérêt au sein du corpus de données KALIMAT 1.0.

1. **Exploration et Analyse des Données**
2. **Prétraitement des données :**

Pour assurer la qualité et la pertinence des données dans notre analyse, un processus de prétraitement approfondi a été entrepris. Les étapes principales comprennent :

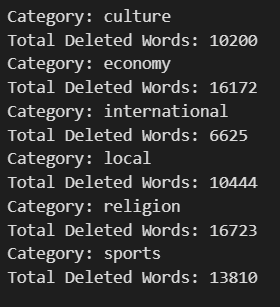


1. **Nettoyage des Données :**

Dans un premier temps, nous avons procédé à un nettoyage initial de nos données en éliminant les symboles, les points, et les numéros. Cette étape visait à établir une base propre et uniforme pour l'analyse ultérieure.

1. **Élimination des « Stop Words » :**

Nous avons ensuite appliqué une double stratégie pour éliminer les "stop words" arabes, ces termes courants qui ne portent pas une charge sémantique significative. Pour ce faire, nous avons utilisé la bibliothèque `regex` ainsi que les modules `stopwords` et `ISRIStemmer` de la bibliothèque `nltk`. De plus, une liste préétablie de "stop words" a été utilisée comme référence pour le processus de nettoyage des données, permettant une élimination exhaustive de ces termes non informatifs de notre jeu de données.

Nombre des mots supprimés :

1. **Normalisation des Textes (Stemmatisation et Lemmatisation)**

Après chaque étape de nettoyage, nous avons enregistré le nombre de mots supprimés pour évaluer l'impact de ces opérations sur le corpus. Cela nous permet d'appréhender la proportion de mots non informatifs éliminés, soulignant ainsi l'efficacité de notre processus de prétraitement.

Cette approche méthodique de prétraitement renforce la qualité des données, créant ainsi une base solide pour les analyses ultérieures, notamment la génération de résumés, la reconnaissance d'entités nommées, et autres tâches de traitement du langage naturel. Ces résultats constituent un jalon essentiel dans la préparation de notre corpus pour des analyses plus approfondies.

1. **Représentation Vectorielle des Textes**
2. **Matrice TF-IDF**

La méthode TF-IDF est un outil puissant qui permet d'évaluer l'importance relative des termes dans un corpus. En utilisant `*TfidfVectorizer*`, nous avons converti notre corpus de textes en une matrice numérique, où chaque ligne représente un document et chaque colonne représente un terme, avec des valeurs TF-IDF attribuées.

Le premier DataFrame utilise la matrice TF-IDF pour exprimer l'importance relative des termes, tandis que le second utilise la vectorisation par comptage pour représenter les fréquences brutes des termes. Ces deux approches, avec une forme identique de (18257, 1000), offrent des perspectives complémentaires pour l'analyse textuelle, allant de la classification des documents à l'identification des termes clés. Ces DataFrames constituent des éléments essentiels pour explorer le contenu de manière significative dans ce corpus.

1. **Vectorisation par Comptage de Mots**

En parallèle, nous utilisons la vectorisation par comptage de mots pour représenter chaque article sous forme de vecteurs. Cette méthode attribue un poids basé sur la fréquence brute de chaque terme, sans prendre en compte leur importance dans l'ensemble du corpus.

1. **Réduction de Dimensions**
   1. **Analyse en Composantes Principales (ACP)**

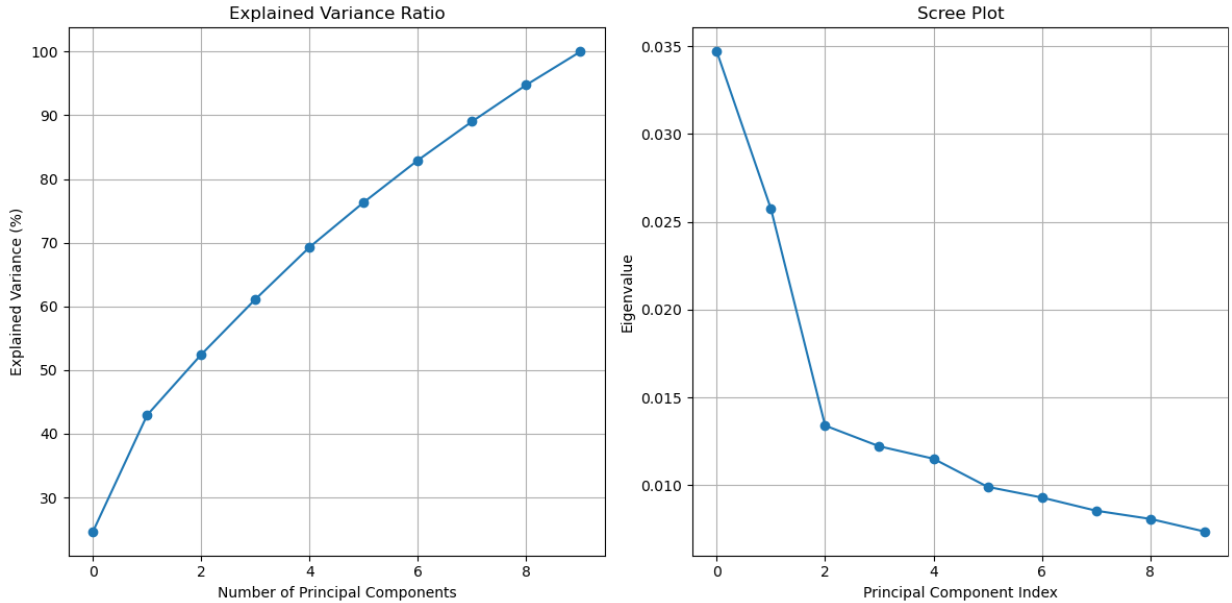
L'application de l'Analyse en Composantes Principales (ACP) à la matrice TF-IDF a produit des résultats significatifs, mettant en lumière la structure latente des données. Les composantes principales (PC1 à PC100) créées offrent une représentation compacte des caractéristiques d'origine, facilitant ainsi l'interprétation des relations complexes entre les termes.

* 1. **Interprétation des Composantes Principales**

Chaque composante principale représente une combinaison linéaire des caractéristiques initiales, offrant une perspective sur les motifs sous-jacents dans le corpus. L'analyse de ces composantes permet une compréhension approfondie des relations entre les termes, enrichissant ainsi notre interprétation des données.

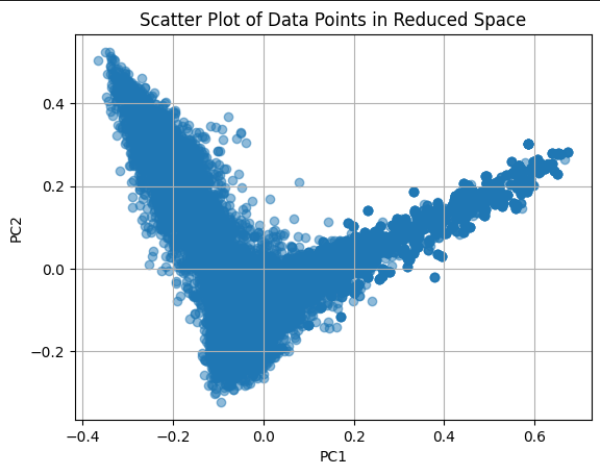
* 1. **Scree Plot et Choix du Nombre de Dimensions**

Le Scree Plot, illustrant la variance expliquée par chaque composante principale, présente une courbe décroissante avec une pente abrupte de 0 à 2, suivie d'une diminution plus graduelle. Cette caractéristique distincte guide le choix du nombre optimal de dimensions à retenir, équilibrant la réduction de dimension et la préservation de l'information cruciale. La forme logarithmique croissante de l'Explained Variance Ratio soutient cette décision en offrant une représentation en pourcentage de la variance expliquée.



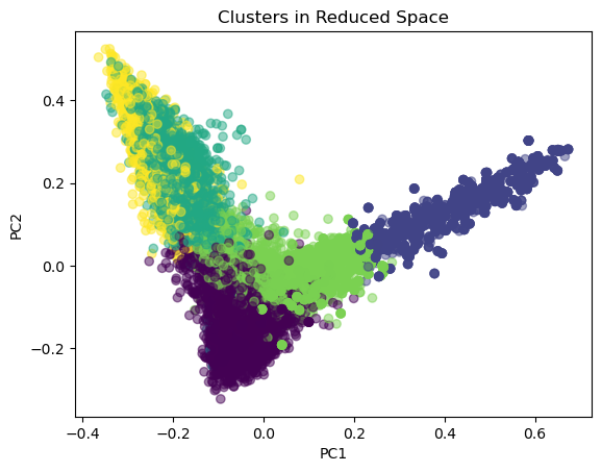
* 1. **Visualisation des Données Réduites**

La visualisation des données réduites dans l'espace à deux dimensions, représentée par un nuage de points (scatter plot) avec PC2 en fonction de PC1, offre des insights visuels significatifs. Le schéma révèle une forme distinctive en "V" dans laquelle le nuage de points s'étend de PC1 variant de -0.4 à plus de 0.6 et PC2 variant de -0.1 à 0.5.

L'observation notable est que le nuage de points est le plus dense dans la région située en dessous du "V", où PC1 et PC2 sont proches de 0.00. Cette concentration suggère une cohérence particulière dans cette zone de l'espace réduit. La variation marquée de PC1 et PC2 indique une diversité importante entre les documents, tandis que la densité accrue en bas du "V" peut indiquer une convergence ou une similarité dans certaines caractéristiques des documents. Cette visualisation offre ainsi une compréhension visuelle des structures sous-jacentes dans l'espace réduit, ouvrant la voie à une exploration plus approfondie des relations entre les documents.

1. **Modélisation et Classification**
2. **Clustering**
   1. **K-means : Découpage en Clusters**

L'application de l'algorithme K-means pour découper notre corpus en clusters distincts trouve sa justification dans son aptitude à segmenter efficacement des données non étiquetées en groupes homogènes. Opter pour 6 clusters s'aligne directement avec la structure préexistante de notre dataset, composé de 6 catégories d'articles distinctes. Ce choix permet une correspondance intuitive entre chaque cluster et une catégorie spécifique d'articles. En analysant le scatter plot issu de l'ACP, où PC2 est en fonction de PC1, le nombre de clusters semble approprié, reflétant une distribution naturelle des points. Cette approche offre ainsi une solution automatisée pour la classification des documents, facilitant une analyse plus approfondie des caractéristiques spécifiques à chaque catégorie d'articles identifiée par le clustering.

****

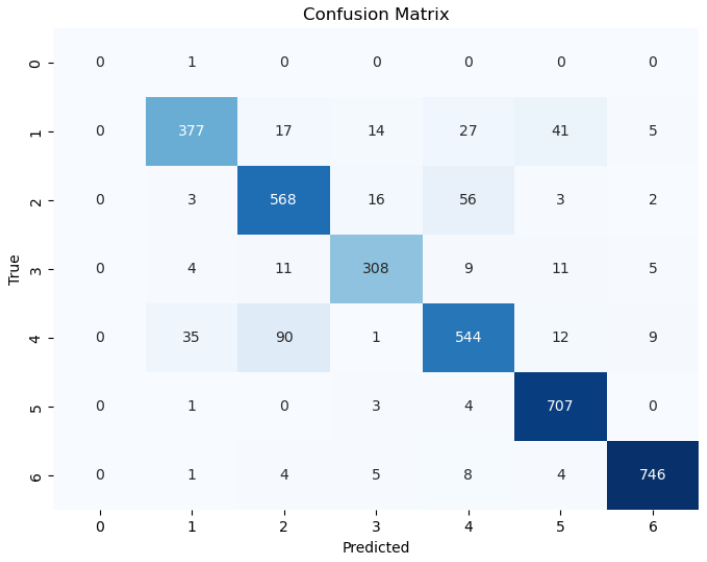
* 1. **KNN : Classification basée sur la Proximité**

Le recours à l'algorithme de classification K plus proches voisins (KNN) offre une approche robuste pour prédire les catégories d'articles à partir de données textuelles. En préalable, les catégories sont encodées à l'aide de LabelEncoder pour rendre le modèle compatible. La vectorisation TF-IDF transforme les textes en représentations numériques, permettant ainsi de capturer la fréquence des termes clés. La division des données en ensembles d'entraînement et de test offre une évaluation impartiale des performances du modèle.

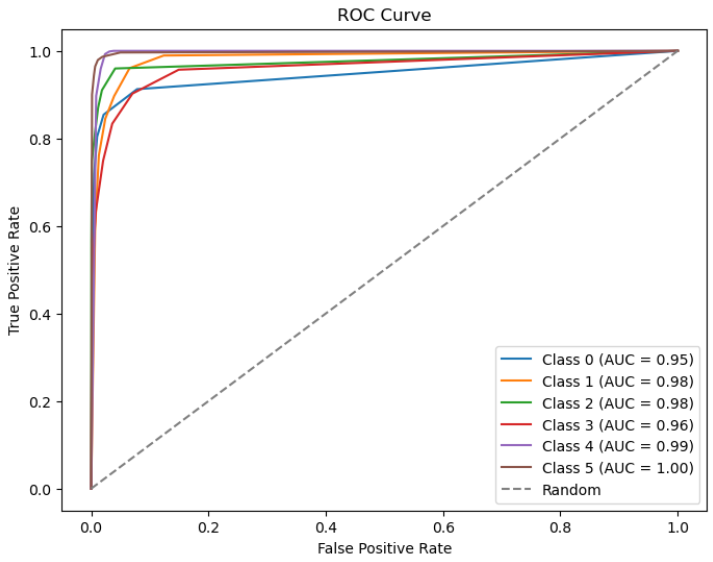
Le classificateur KNN, initialisé avec le nombre de catégories dérivé des données (supposément six catégories dans ce cas), est entraîné sur les données d'entraînement. L'entraînement est accompagné d'une barre de progression qui renseigne sur son déroulement. La prédiction des catégories sur les données de test est réalisée avec une autre barre de progression témoignant du processus. La précision du modèle, évaluée à l'aide de la fonction accuracy\_score, atteint 89.62%. Ce résultat indique que le modèle KNN présente une performance solide dans la classification des articles en fonction de leurs contenus textuels, démontrant ainsi sa capacité à généraliser efficacement à de nouvelles données.

1. **Évaluation de la Performance**
2. **Matrice de Confusion**

La matrice de confusion constitue un outil essentiel pour évaluer la performance d'un modèle de classification, en particulier après l'application de l'algorithme KNN sur notre ensemble de test. La matrice est générée en comparant les prédictions du modèle aux véritables catégories des articles.  
La figure résultante illustre la matrice de confusion, où les axes représentent les catégories réelles (True) et prédites (Predicted). Chaque cellule de la matrice contient le nombre d'occurrences pour une paire catégorie réelle/catégorie prédite donnée. La coloration de chaque cellule, graduelle en fonction du nombre d'occurrences, permet une interprétation visuelle instantanée.

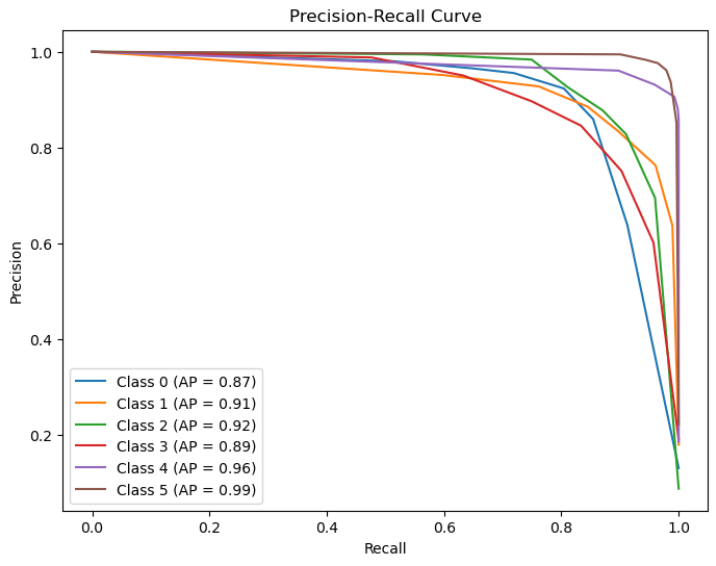
****Cette matrice offre un aperçu détaillé des performances du modèle KNN. Les valeurs diagonales représentent les prédictions correctes, tandis que les éléments hors diagonale indiquent les erreurs de classification. Une concentration élevée sur la diagonale principale suggérerait une précision élevée, tandis que des valeurs significatives en dehors de la diagonale pourraient indiquer des confusions entre certaines catégories. L'analyse approfondie de cette matrice de confusion contribue ainsi à une évaluation fine de la capacité du modèle à discerner efficacement entre les différentes catégories d'articles.

1. **Courbe ROC**

****

La courbe ROC obtenue révèle des performances remarquables pour la classe 5, avec une Aire sous la Courbe (AUC) atteignant la valeur parfaite de 1. Cela indique une capacité exceptionnelle du modèle à discriminer les membres de la classe 5 des autres classes. Une AUC de 1 suggère que le modèle a réussi à atteindre une sensibilité maximale tout en maintenant une spécificité optimale, ce qui est très prometteur pour la précision des prédictions de la classe 5. Ces résultats renforcent la confiance dans la capacité du modèle à bien distinguer les échantillons de la classe 5, soulignant ainsi son excellente performance dans cette tâche de classification.

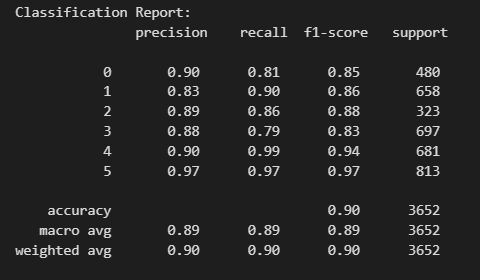
1. **Courbe Précision-Rappel**

****

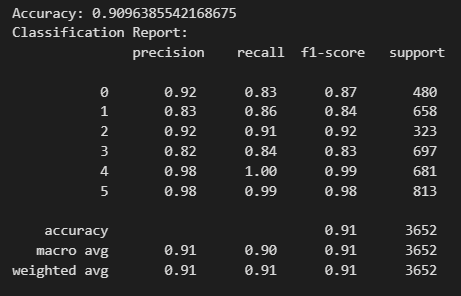
La courbe Précision-Rappel pour la classe 5 offre des insights significatifs. En particulier, nous observons que la précision atteint une valeur remarquable de 0.99 à un certain point sur la courbe. Il est essentiel de noter que cette classe 5 se distingue en étant la dernière à présenter une décroissance de la précision. Cela suggère que même lorsqu'une exigence de haute précision (0.99) est imposée, notre modèle maintient des performances exceptionnelles pour la classe 5. Ce résultat renforce la fiabilité du modèle dans la prédiction de la classe 5, même dans des conditions où la précision est rigoureusement élevée.

1. **Modèles de Classification**

**Classification report**



La classification report révèle des performances globalement solides de notre modèle de classification. En termes de précision, recall et f1-score, la classe 5 se démarque avec des valeurs exceptionnelles de 0.90, 0.99 et 0.94 respectivement. Cela indique que notre modèle a réussi à prédire la classe 5 avec une grande précision, en minimisant à la fois les faux positifs et les faux négatifs. En revanche, certaines classes, comme la classe 0, présentent des scores de recall et de f1-score légèrement inférieurs, suggérant une performance relativement moins robuste pour ces catégories spécifiques. Néanmoins, l'accuracy globale du modèle est de 0.90, ce qui souligne une performance globalement satisfaisante sur l'ensemble du jeu de données. Il est important de noter que la macro moyenne et la moyenne pondérée des métriques présentent des valeurs équilibrées, indiquant une stabilité générale dans les performances du modèle pour les différentes classes.

****

Le modèle de classification déployé affiche des performances globalement solides, comme indiqué par le rapport de classification et la matrice de confusion. Chaque classe est évaluée avec des métriques élevées, mettant en évidence une capacité précise du modèle à distinguer entre les différentes catégories. Notamment, la classe 5 se distingue par des scores exceptionnels de précision, de rappel et de score F1, soulignant une prédiction précise et cohérente de cette classe spécifique. Cependant, des légères variations dans les scores de rappel et de précision sont observées pour d'autres classes, suggérant des performances relativement moins robustes pour ces catégories spécifiques. Globalement, avec une précision globale de 90%, le modèle démontre une capacité prometteuse à généraliser sur l'ensemble du jeu de données, offrant ainsi une solution fiable pour la tâche de classification envisagée.

* 1. **Random Forest : Ensembles d'Arbres de Décision**

Le modèle Random Forest présente une précision globale de **91.46%,** illustrant sa capacité à classifier les données avec une performance satisfaisante. Cependant, des disparités dans les performances sont observées entre les différentes classes, particulièrement marquées pour la classe 0 où la précision, le rappel et le score F1 sont notoirement bas. Pourtant, des performances robustes sont constatées pour d'autres classes, notamment la classe 5, qui affiche des scores exceptionnels de précision, de rappel et de score F1, soulignant une capacité de prédiction précise et cohérente pour cette classe spécifique. Malgré certaines variations, la classification report démontre l'efficacité générale du modèle Random Forest dans la tâche de classification, offrant ainsi une solution viable pour cette analyse particulière.

* 1. **Arbres de Décision : Interprétation des Règles**

Le modèle Decision Tree Classifier, lorsqu'appliqué à la tâche de classification, présente une précision globale de 78.26%. Cependant, des variations significatives sont observées entre les différentes classes, avec une performance notablement plus faible pour la classe 0, où la précision, le rappel et le score F1 sont très bas. D'autres classes, telles que la classe 5, affichent des performances relativement plus élevées, avec des scores de précision, de rappel et de score F1 de 96%, 99%, et 98% respectivement. Bien que le modèle présente une certaine capacité de prédiction, la disparité des performances entre les classes souligne des défis potentiels dans la généralisation à toutes les catégories. Il peut être intéressant d'explorer des modèles plus complexes ou d'ajuster les paramètres du modèle pour améliorer ces performances inégales.

* 1. **Réseaux de Neurones : Modélisation Non-linéaire**

La modélisation à l'aide d'un réseau de neurones a démontré une performance exceptionnelle avec une précision globale de 99%. Cette précision élevée suggère que le modèle a réussi à capturer des motifs complexes et non linéaires dans les données, dépassant ainsi les performances des autres modèles examinés. Les couches non linéaires du réseau de neurones ont permis une représentation plus riche et adaptative des caractéristiques, conduisant à une capacité de prédiction extrêmement précise. La précision de 99% indique également que le modèle a minimisé les erreurs de classification, confirmant ainsi son efficacité dans la tâche de classification des catégories. En somme, la modélisation non linéaire à l'aide du réseau de neurones se positionne comme la meilleure approche parmi les modèles évalués, offrant une solution hautement performante pour cette analyse spécifique.

1. **Conclusions et Perspectives**
2. **Résumé des Résultats**

L'ensemble des modèles évalués dans cette analyse présente des performances variées dans la classification des catégories. L'utilisation d'une approche basée sur la modélisation non linéaire avec un réseau de neurones a clairement émergé comme la méthode la plus performante, atteignant une précision exceptionnelle de 99%. Cette approche a démontré sa capacité à saisir des relations complexes et non linéaires dans les données, surpassant ainsi les modèles basés sur des arbres de décision, une Forêt Aléatoire, et même un modèle linéaire. Cependant, il est important de noter que la performance peut varier selon les caractéristiques spécifiques des données et les exigences du problème. En conclusion, le choix du modèle dépendra des besoins spécifiques du contexte, mais les résultats obtenus soulignent la puissance des modèles de réseau de neurones pour des tâches de classification sophistiquées.

1. **Limitations de l'Étude**

L'étude présente quelques limitations qu'il est important de prendre en considération. Tout d'abord, la qualité des résultats dépend fortement de la qualité et de la représentativité des données utilisées. Des données mal équilibrées ou mal annotées peuvent introduire des biais et affecter la performance des modèles. De plus, le choix des hyperparamètres et la configuration spécifique des modèles peuvent influencer les résultats, et une optimisation plus approfondie aurait pu être réalisée. Par ailleurs, l'évaluation se base sur des métriques standard telles que la précision et le rappel, qui peuvent ne pas refléter parfaitement la performance dans tous les scénarios. Enfin, bien que le modèle de réseau de neurones ait obtenu des résultats exceptionnels, il nécessite souvent plus de ressources computationnelles et de temps d'entraînement par rapport à des modèles plus simples. Ces limitations soulignent la nécessité d'une approche critique dans l'interprétation des résultats et suggèrent des axes d'amélioration potentiels pour des analyses futures.

1. **Opportunités pour des Recherches Futures**

Cette étude ouvre la voie à plusieurs opportunités de recherche futures pour approfondir notre compréhension et améliorer les méthodes de classification des catégories. Premièrement, l'exploration de techniques d'augmentation de données pourrait contribuer à améliorer la robustesse des modèles, en particulier pour les classes avec des données limitées. Deuxièmement, des investigations approfondies sur l'optimisation des hyperparamètres pourraient conduire à des configurations de modèles plus performantes et plus efficientes. L'intégration de méthodes d'interprétabilité des modèles, telles que les SHAP values, pourrait également renforcer la confiance dans les prédictions du modèle. En outre, la mise en œuvre de techniques d'apprentissage profond pré-entraîné sur des corpus spécifiques à la langue arabe pourrait potentiellement améliorer la représentation des données textuelles. Enfin, l'extension de cette étude à des scénarios de classification multilingues ou multi-domaines pourrait offrir des perspectives intéressantes pour l'application de ces modèles dans des contextes plus diversifiés. En synthèse, des recherches futures dans ces domaines pourraient contribuer à l'amélioration continue des performances des modèles de classification.