RÉPUBLIQUE ALGÉRIENNE DÉMOCRATIQUE ET POPULAIRE

Ministère de l'Enseignement Supérieur et de la Recherche Scientifique

Université de Mohamed El-Bachir El-Ibrahimi - BBA

Faculté des Mathématiques et Informatiques

MÉMOIRE

Présenté en vue de l'obtention du diplôme

Master en Informatique

Spécialité : Réseaux & Multimédias

Thème

Classification thématique des textes multilingue

Etude de cas dans le domaine de sport

Présenté par:

- ABERKANE Ayoub
- ATTIA Asma

Soutenu le: 10/06/2025, Devant le jury composé de:

Pr. Zouache Djaafar	Président
Dr. Boutouhami Sara	Examinateur
Dr. Lynda SAIFI	Encadrant

Promotion 2024 / 2025

Remerciements

Ce mémoire a été réalisé dans le cadre de l'obtention du diplôme d'ingénieur en informatique. Nous tenons tout d'abord à exprimer notre profonde gratitude envers Dieu Tout-Puissant, pour nous avoir donné la force, la patience et la santé nécessaires pour mener à bien ce travail.

Nous remercions sincèrement nos parents pour leur amour inconditionnel, leurs sacrifices constants, leur soutien moral et matériel, ainsi que leur confiance qui nous a toujours motivés à avancer.

Nos remerciements les plus chaleureux vont à ****Mme Lynda Saïfi****, notre encadrante, pour son accompagnement rigoureux, ses conseils avisés, sa disponibilité constante et la qualité de son encadrement tout au long de ce projet. Son implication a grandement contribué à l'aboutissement de ce travail.

Nous exprimons également notre reconnaissance aux membres du jury pour avoir accepté d'évaluer notre travail, et pour l'intérêt qu'ils portent à notre recherche.

Nos remerciements s'adressent aussi à l'ensemble des enseignants, du personnel technique et administratif du département d'informatique de **l'Université de Bordj Bou Arréridj**, pour leur encadrement tout au long de notre formation et pour leur bienveillance.

Enfin, nous tenons à remercier toutes les personnes qui, de près ou de loin, ont contribué à la réalisation de ce mémoire. Merci à toutes et à tous pour votre soutien précieux.

Dédicace

بسم الله الرحمن الرحيم

Louange et gratitude à Allah pour tous les bienfaits dont Il m'a comblé, et pour m'avoir accordé Sa grâce afin d'achever ce travail après tant d'efforts et de fatigue. Je Lui demande de faire de cette étape un début vers un succès encore plus grand, et de me guider dans la suite de mon parcours.

Je dédie ce travail modeste :

À celui qui m'a appris le sens de l'ambition et de la responsabilité, à celui dont je porte le nom avec fierté, et à qui je souhaite longue vie, santé et bonheur, mon cher ***père****, toute ma reconnaissance et mon respect.*

À celle dont les prières ont été mon arme de réussite, dont le soutien a été la lumière sur mon chemin, à mon âme et à mon refuge, ma chère ***mère****, aucun mot ne suffit à exprimer ma gratitude. Je te dois tout mon amour et toute ma reconnaissance.*

À mes chers ***frères et sœurs****, qui ont toujours été à mes côtés à chaque étape. Votre amour et votre force ont été mon véritable soutien.*

À mon compagnon de route depuis l'enfance, ***Yacine*** *mon frère de cœur avant même les mots, merci pour ta loyauté et ta présence constante à mes côtés.*

À mes chers amis ***Hani, Amine et Amir****, qui ont partagé avec moi les moments d'étude, de fatigue et de joie. Merci pour votre belle énergie.*

Et je n'oublie pas ***Asma Attia****, binôme de mémoire , qui a marché avec moi pas à pas. Merci pour ta collaboration, ta patience et ton engagement.*

⋆ ABERKANE Ayoub ⋆

Dédicace

Le chemin n'a pas été court, ni parsemé de facilités, mais je l'ai fait. Louange à Allah qui a facilité les débuts et nous a permis d'atteindre la fin par Sa grâce et Sa générosité.

J'offre ce succès avant tout à moi-même, à mon âme ambitieuse qui a commencé avec un rêve et s'est terminée par une réussite. Ensuite, à tous ceux qui m'ont soutenue pour achever mon parcours universitaire.

À celle qui m'a portée en son sein malgré la fatigue et qui continue à me porter avec son amour et sa tendresse, à l'incarnation de la patience et du sacrifice, à celle qui, par ses veilles et ses prières, ouvre toutes les portes...
Ma chère mère, toute ma gratitude et mon amour t'appartiennent.

À celui qui a quitté ce monde mais jamais mon cœur, à mon cher père — qu'Allah lui fasse miséricorde — mon pilier et modèle. J'offre cette réussite à son âme pure, en espérant qu'il soit fier de moi, comme j'ai toujours été fière de lui.

À mon frère et ma sœur, mes compagnons de route et soutiens inébranlables. Vous avez été et resterez toujours ma force.

À *mes précieuses amies,* ***Boukhalfa Boutheyna****,* ***Salik Soundous*** *et* ***Manel Gedrouh****, qui ont partagé avec moi chaque instant de ce voyage. Votre amitié est un trésor inestimable.*

À mes chères amies ***Belmoumen Houda "Assoltii"*** *et* ***Atamna rim****, merci pour votre soutien et votre présence sincère.*

*Un immense merci à **Mourad Mihoubi**, qui a toujours été là pour me remonter le moral, me motiver à continuer ce mémoire, croire en moi et m'aider dans chaque étape de mon parcours. Ta présence et ton soutien ont été essentiels.*

*Et je n'oublie pas **Ayoub Aberkane**, binôme de mémoire, qui a marché à mes côtés dans ce projet du début à la fin. Merci pour ta collaboration et ton engagement.*

Et enfin, à tous ceux qui m'ont soutenue, encouragée et accompagnée, recevez toute ma gratitude et ma reconnaissance.

Ce grand jour est enfin arrivé, celui dont j'ai tant rêvé après des années d'efforts et de sacrifices. Par la grâce d'Allah, mon rêve s'est réalisé.

Louange à Allah, source de tout espoir et de tout bien.

⋆ ATTIA Asma ⋆

Résumé

Avec le développement numérique et l'augmentation du volume de contenu textuel publié quotidiennement, notamment dans le domaine du sport, le besoin d'organiser ces informations devient de plus en plus crucial. Cette étude vise à traiter des textes sportifs rédigés en plusieurs langues en utilisant des techniques de traitement automatique du langage naturel et d'apprentissage machine, afin de les classer selon les thématiques abordées. Pour unifier le traitement linguistique des textes multilingues, le modèle de traduction automatique **NLLB** a été utilisé pour traduire les contenus en anglais, ce qui a contribué à améliorer la segmentation thématique des textes.

Plusieurs algorithmes supervisés ont été appliqués, notamment **Naive Bayes**, la machine à vecteurs de support **SVM** et le perceptron multicouche **MLP**, sur un jeu de données sportifs extrait de la plateforme **Kaggle**. Après des étapes de nettoyage des données et de vectorisation des textes à l'aide de l'algorithme **TF-IDF**, les modèles ont été entraînés et comparés. Les résultats ont montré que les modèles **SVM** et **MLP** ont obtenu les meilleures performances en termes de précision, tandis que le modèle **Naive Bayes** s'est distingué par sa rapidité d'exécution. Cette étude démontre l'efficacité de la classification thématique des textes multilingues dans le domaine sportif et ouvre la voie à des améliorations futures grâce à des modèles linguistiques plus avancés.

Mots clés : traitement automatique des langues, classification de textes, sport, TF-IDF, SVM, Naive Bayes, MLP,mBERT NLLB.

Abstract

With the rise of digital development and the growing volume of textual content published daily, particularly in the sports domain, the need to organize such content has become increasingly important. This study aims to process multilingual sports texts using natural language processing and machine learning techniques, in order to classify them according to the topics they address. To standardize the linguistic processing of multilingual texts, the automatic translation model **NLLB** was used to translate the content into English, which contributed to improving the thematic segmentation of the texts.

Several supervised algorithms were applied, including **Naive Bayes**, Support Vector Machine **SVM**, and Multilayer Perceptron **MLP**, on a sports dataset collected from the **Kaggle** platform. After data cleaning and converting the texts into numerical representations using the **TF-IDF** algorithm, the models were trained and compared. Results showed that **SVM** and **MLP** achieved the best performance in terms of accuracy, while the **Naive Bayes** model stood out for its execution speed. This study demonstrates the effectiveness of multilingual thematic classification in the sports domain and paves the way for future improvements using more advanced language models.

Keywords : natural language processing, text classification, sport, TF-IDF, SVM, Naive Bayes, MLP,mBERT NLLB.

ملخص

في ظل التطور الرقمي وتزايد حجم المحتوى النصي المنشور يوميا، خاصة في المجال الرياضي، أصبحت الحاجة إلى تنظيم هذا المحتوى أكثر أهمية. تهدف هذه الدراسة إلى معالجة النصوص الرياضية المكتوبة بلغات متعددة من خلال استخدام تقنيات المعالجة الآلية للغة وتقنيات تعلم الآلة، وذلك لتقسيمها حسب المواضيع التي تتناولها. من أجل توحيد المعالجة اللغوية للنصوص متعددة اللغات، تم اعتماد نموذج الترجمة الآلية NLLB لترجمة المحتوى إلى اللغة الإنجليزية، مما ساهم في تحسين عملية التقسيم الموضوعي للنصوص.

تم تطبيق مجموعة من الخوارزميات الإشرافية مثل Naive Bayes، وآلة الدعم الناقل (SVM)، والشبكة العصبية متعددة الطبقات (MLP) على مجموعة بيانات رياضية مأخوذة من منصة Kaggle. بعد مراحل تنظيف البيانات وتحويل النصوص إلى تمثيلات عددية باستخدام خوارزمية TF-IDF، تم تدريب النماذج ومقارنتها. أظهرت النتائج أن نماذج SVM وMLP حققت أفضل أداء من حيث الدقة، في حين تميز نموذج Naive Bayes بسرعة التنفيذ. توضح هذه الدراسة فعالية التصنيف الموضوعي للنصوص متعددة اللغات في المجال الرياضي، وتفتح المجال أمام تحسينات مستقبلية باستخدام نماذج لغوية أكثر تطورا.

الكلمات الرئيسية : معالجة اللغة الطبيعية، تصنيف النصوص، الرياضة، TF-IDF، SVM، Naive Bayes، MLP، mBERT، NLLB.

Table of contents

List of Figures

List of Tables

Abréviations

TAL :	Traitement Automatique des Langues.
TF-IDF:	Term Frequency - Inverse Document Frequency
TF:	Term Frequency
IDF:	Inverse Document Frequency
LDA:	Latent Dirichlet Allocation
BERT:	Bidirectional Encoder Representations from Transformers
SVM:	Support Vector Machine
MLP:	Multilayer Perceptron
NLTK:	Natural Language Toolkit
NLP:	Natural Language Processing
F1-score:	Score F1 (moyenne harmonique de précision et rappel)
ML:	Machine Learning
TALN:	traitement automatique du langage nature
NLLB:	No Language Left Behind
mBART:	Multilingual Bidirectional and Auto-Regressive Transformer

Introduction Générale

Contexte

Le monde du sport génère une quantité massive de textes dans différentes langues, que ce soit sur les performances des athlètes, les événements sportifs, les analyses des matchs ou les tendances en matière de santé et de nutrition. Ces textes sont souvent difficiles à analyser en raison de la diversité des langues et des contextes culturels associés à chaque événement. La segmentation thématique, qui consiste à identifier et à classer les principaux sujets abordés dans un texte, devient donc essentielle pour organiser efficacement ces informations. Ce projet se concentre sur la segmentation thématique des textes multilingues dans le domaine du sport.

Problématique

La diversité linguistique et le caractère non structuré des articles sportifs constituent un défi pour une classification automatisée précise. Il s'avère donc indispensable de créer un système qui peut segmenter et classer par thèmes ces textes, en respectant leur diversité tant linguistique que thématique. Comment peut-on, en utilisant des méthodes d'apprentissage automatique, optimiser la catégorisation thématique des textes multilingues dans le secteur sportif ?

Objectifs

En tenant compte de ce contexte, nous avons déterminé les objectifs suivants pour notre projet :

- Implémenter un système de classification thématique pour les textes en plusieurs langues.
- Évaluer les performances de divers algorithmes de classification (Naive Bayes, SVM, MLP) en comparaison.
- Suggérer une solution efficace et automatisée pour la gestion des articles sportifs.

Plan du mémoire

- **Chapitre 1 :** Dans ce chapitre, nous exprimons un examen théorique de la segmentation thématique, des méthodes traditionnelles et des enjeux associés au multilinguisme.
- **Chapitre 2 :** Dans ce chapitre, nous exprimons un exposition du projet, approche méthodologique, instruments employés et modèles mis en œuvre.
- **Chapitre 3 :** Dans ce chapitre, nous exprimons un mise en œuvre pratique, cadre technique, représentations visuelles et données brutes.
- **Chapitre 4 :** Dans ce chapitre, nous exprimons un évaluation des performances des modèles et discussion des résultats obtenus.

Chapter 1

Segmentation thématique: définitions et concepts de base

1.1 Introduction

Avec la croissance des contenus numériques, le domaine du sport génère une grande quantité de textes dans différentes langues, allant des articles de presse aux analyses tactiques en passant par les discussions sur les réseaux sociaux.Ces textes sont souvent non structurés et abordent plusieurs sujets à la fois, rendant leur analyse difficile.

La segmentation thématique permet d'identifier et de séparer ces différentes thématiques au sein d'un même texte, facilitant ainsi son exploitation.

1.2 Définition et importance de la segmentation thématique

1.2.1 Définition de la segmentation thématique

La segmentation thématique est une tâche qui consiste à diviser un texte non structuré en segments thématiquement cohérents, c'est-à-dire en parties traitant du même sujet. Cette segmentation permet d'organiser le texte en unités homogènes sur le plan thématique. Elle joue un rôle fondamental dans le domaine du traitement automatique des langues (TAL). L'objectif est de structurer les textes en segments dits « thématiques », qui se forment autour d'une identité de sujet. [1] [2]

1.2.2 Importance de la ségmentation thématique

La segmentation thématique de texte est cruciale dans le traitement automatique des langues (TAL) et l'analyse de l'information. Plusieurs éléments essentiels sont à l'origine de son importance :

- **Amélioration des systèmes de recherche d'information :**
  Au lieu de renvoyer un document complet en réponse à une demande, la segmentation permet d'identifier directement les passages pertinents. [3]
- **Amélioration de la lecture :**
  Elle facilite la structuration des données en fonction des sujets traités, ce qui rend les textes plus clairs et compréhensibles. [4]
- **Structuration et organisation de l'information :**
  La segmentation est nécessaire pour toute organisation discursive, en permettant à la fois la division et le regroupement du contenu en fonction d'un critère organisationnel.C'est essentiel en linguistique du discours et en traitement automatique du langage (TAL). [5]

1.3 Les approches traditionnelles de la segmentation thématique

1.3.1 Approches basées sur les règles : Analyse lexicale et syntaxique

Dans le domaine du traitement automatique des langues (TAL), les méthodes fondées sur des règles linguistiques sont fréquemment mises en œuvre pour l'analyse lexicale et syntaxique des documents. Ces dernières s'appuient sur des modèles grammaticaux formels, établissant des normes précises pour la détection et l'organisation des composantes linguistiques.

1.3.1.1 Analyse lexicale

L'analyse lexicale constitue la première phase du traitement d'un texte. Elle a pour objectif de diviser et de reconnaître les éléments fondamentaux du langage, tels

que les mots et les morphèmes. Cette étude se base sur :

- **Des dictionnaires et lexiques annotés**, qui permettent de classer les mots en fonction de leur catégorie grammaticale.
- **Des règles morphologiques**, employées pour identifier les variations d'un terme (par exemple : conjugaison, déclinaison).
- **Des outils d'analyse lexicale**, tels que les analyseurs morphologiques, qui décomposent les mots en éléments de signification.

[6]

1.3.1.2 Analyse syntaxique

L'analyse syntaxique vise à identifier les liens entre les termes d'une phrase pour saisir sa composition grammaticale. Elle est effectuée selon diverses méthodes :

- **Les grammaires basées sur des règles**, où chaque phrase est décortiquée grâce à un ensemble de règles syntaxiques définies manuellement.
- **L'analyse syntaxique dépendante**, qui représente la structure d'une phrase sous la forme d'un arbre de dépendances reliant les mots les uns aux autres.
- **Les analyseurs syntaxiques automatiques**, qui utilisent ces règles pour identifier les groupes nominaux, verbaux et prépositionnels ainsi que structurer le texte.

[6]

1.4 Vectorisation

1.4.1 TF-IDF (Term Frequency - Inverse Document Frequency)

1.4.1.1 Définition de la TF-IDF

TF-IDF est une technique statistique employée pour déterminer la pertinence d'un terme dans un document par rapport à un ensemble de documents (corpus). Elle est constituée de deux sections [7]:

1. **TF (Term Frequency):** La fréquence du mot dans un document donné:

$$tf_{i,j} = \frac{n_{i,j}}{\sum n_{k,j}} \tag{1.1}$$

où $n_{i,j}$ est le nombre d'occurrences du mot t_i dans le document d_j et $\sum n_{k,j}$ est le nombre total de mots dans ce document.

2. **IDF (Inverse Document Frequency):**Une mesure de l'importance générale du mot dans le corpus :

$$idf_i = \log \frac{|D|}{|\{j : t_i \in d_j\}|} \tag{1.2}$$

où $|D|$ est le nombre total de documents et $|\{j : t_i \in d_j\}|$ est le nombre de documents contenant le mot t_i.

3. **TF-IDF final:**

$$tfidf_{i,j} = \frac{tf_{i,j} idf_i}{\sqrt{\sum (tf_{i,j} idf_i)^2}} \tag{1.3}$$

1.4.2 LDA (Latent Dirichlet Allocation)

L'Allocation de Dirichlet Latente (ADL) est une méthode de modélisation thématique qui a pour but de déceler des thèmes dissimulés au sein d'un corpus de documents. Il est basé sur un modèle bayésien probabiliste, mis en place par Blei, Ng et Jordan en 2003. [8]

1.4.3 TextTiling

TextTiling est une méthode de segmentation thématique de textes proposée par Marti A. Hearst en 1997. Elle repose sur la cohésion lexicale et divise un texte en segments thématiquement cohérents. [9]

1.5 Comparaison entre les trois méthodes de segmentation thématique

Les approches **TF-IDF**, **LDA** et **TextTiling** sont parmi les plus utilisées pour la segmentation thématique des textes.

Méthode	Avantages	Limites
TF-IDF	- Simple à implémenter - Efficace pour les textes courts - Compatible avec les modèles supervisés **(SVM, Naive Bayes, MLP)**	- Ne tient pas compte du sens profond **(sémantique)** - Sensible aux synonymes et à la variation lexicale
LDA	- Capacité à identifier des thématiques cachées - Représentation probabiliste des sujets	- Moins performant pour les petits corpus - Résultats parfois difficiles à interpréter
TextTiling	- Bonne segmentation basée sur la cohésion lexicale - Convient aux textes structurés	- Moins adaptée aux textes courts - Peu efficace avec des textes multilingues ou très variés

Table 1.1: Comparaison entre les méthodes de segmentation thématique

• Dans le cadre de notre projet, nous avons privilégié la méthode **TF-IDF** en raison de sa simplicité, rapidité et performance pour la classification supervisée, notamment lorsque les données sont structurées sous forme de titres d'articles. Contrairement à **LDA** ou **TextTiling**, **TF-IDF** s'intègre parfaitement avec les classificateurs tels que **SVM**, **Naive Bayes** ou **MLP**, permettant ainsi une segmentation thématique fine et efficace dans un environnement multilingue.

1.6 Défis liés au multilinguisme

L'analyse automatique de textes en plusieurs langues représente un défi de taille à cause des disparités linguistiques, du déficit de corpus annotés et de l'exigence de modèles sur mesure. Ces enjeux ont un impact direct sur la qualité des résultats obtenus lors des tâches de segmentation et d'analyse thématique de textes dans plusieurs langues.

1.6.1 Variantes linguistiques : divergences grammaticales et lexicales entre les différentes langues.

Le traitement automatique des langues est compliqué par leurs structures grammaticales, leur vocabulaire et leurs règles de syntaxe distinctifs. [10]Les variations de la langue peuvent être de nature phonétique, morphologique ou sémantique et sont façonnées par des éléments sociaux, culturels et géographiques. Ces variations compliquent l'usage homogène des méthodes de traitement du langage naturel (NLP) à

diverses langues. De plus, ces disparités influencent considérablement la compréhension et l'étude des langues dans un contexte multilingue, ce qui rend la normalisation difficile pour les modèles d'apprentissage automatique. [11]

1.6.2 Insuffisance de corpus annotés : difficulté à obtenir des bases de données d'entraînement multilingues.

Le traitement du langage naturel par apprentissage automatique repose fortement sur des corpus annotés de haute qualité. [12] Cependant, dans un cadre multilingue, ces ressources demeurent insuffisantes. L'utilisation efficace par les modèles des textes informels et non structurés nécessite un travail considérable de marquage et d'arrangement. De plus, les corpus parallèles et multilingues sont essentiels pour les travaux de traduction et d'analyse interlinguistique. Cependant, leur rareté et le coût élevé de leur production limitent leur accessibilité, ce qui impacte l'efficacité des modèles dans un contexte multilingue. [13]

1.6.3 Besoin de modèles adaptés : utilisation de modèles capables de gérer plusieurs langues simultanément

Afin de relever ces défis, des modèles de traitement du langage naturel capables de gérer plusieurs langues ont été élaborés. XLM-RoBERTa, l'un d'eux, s'appuie sur un apprentissage non supervisé basé sur un large corpus multilingue, ce qui améliore la compréhension interlangue. [14] De la même manière, mBERT (BERT multilingue) est élaboré pour acquérir des représentations linguistiques communes, favorisant ainsi le transfert de savoir entre diverses langues. Ces méthodes favorisent la généralisation des tâches de traitement du langage naturel dans un cadre multilingue, même si des avancées sont encore requises pour une meilleure prise en charge des langues sous-représentées dans les ensembles de données d'apprentissage. [15]

1.7 Présentation du domaine sportif et justification du choix

Le sport est une pratique universelle qui joue un rôle majeur dans la société. Cela couvre une vaste gamme de disciplines (football, basketball, tennis, athlétisme, etc.) et occupe une place centrale dans les domaines culturels, économiques et sociaux. Des millions de spectateurs sont captivés par des événements sportifs mondiaux, comme la Coupe du Monde de la FIFA, les Jeux Olympiques ou les championnats nationaux, produisant ainsi une quantité substantielle de contenu médiatique.

En raison de la croissance du numérique et des médias numériques, le domaine du sport s'est transformé en un secteur où l'information est constamment diffusée, que ce soit par le biais d'articles de presse, d'analyses, de commentaires en temps réel ou d'échanges sur les réseaux sociaux.

La diversité des sources et l'utilisation de plusieurs langues rendent difficile la structuration et l'organisation des contenus. Voilà pourquoi il est nécessaire d'utiliser des techniques de segmentation thématique sophistiquées pour organiser ces informations de manière efficace.

1.8 Conclusion

Ce chapitre a exposé les bases théoriques et techniques de la segmentation thématique des textes sportifs. Nous avons souligné la pertinence de cette démarche et examiné diverses techniques employées pour catégoriser et classifier les documents selon les sujets traités.

Dans le prochain chapitre, nous exposerons la méthode mise en œuvre, en faisant une comparaison entre diverses stratégies et en précisant les sélections techniques effectuées pour notre projet.

Chapter 2

Méthodologie: Modèles, Algorithmes et Fonctions Utilisés dans la Segmentation Thématique

2.1 Introduction

Ce chapitre se consacre à l'exploration des diverses méthodes de segmentation thématique des textes, tout en examinant les modèles théoriques et algorithmiques qui leur sont associés. Nous effectuerons une comparaison entre plusieurs techniques différentes, décrirons la structure de notre service et préciserons les fonctions et algorithmes utilisés afin d'assurer un découpage efficace et exact.

2.2 Présentation du projet

Ce projet se positionne dans le champ du traitement automatique du langage naturel dédié aux textes sportifs. L'objectif consiste à évaluer et comparer les performances de diverses méthodes de classification thématique sur un ensemble multilingue d'articles sportifs.

À cet effet, nous avons eu recours à des modèles classiques de machine learning tels que **SVM**, **Naive Bayes** et **MLP**.
La base de données employée provient de la plateforme **Kaggle** et comprend des articles sportifs rassemblés via **Google News**.

2.3 Méthodologie adoptée

L'exécution du projet suit une méthodologie organisée en différentes étapes complémentaires, qui vont de l'analyse des données à la valorisation finale des performances des modèles de classification. Voici le processus principal suivi :

1. Collecte des données
2. Prétraitement linguistique
3. Vectorisation des titres
4. Classification thématique
5. Évaluation des performances
6. Optimisation

2.3.1 Collecte des données

Nous avons fait appel au jeu de données **Google News Sports** qui est accessible sur la plateforme Kaggle. Il regroupe des articles en langue anglaise traitant de diverses disciplines sportives telles que **le football**, **le tennis**, **le cricket**, **le rugby**, ...etc. Nous avons procédé à un nettoyage et un filtrage des données, suivis d'une structuration, pour ne garder que les colonnes pertinentes **(titre, contenu, catégorie)**.voici le lien de la base de données:

https://www.kaggle.com/datasets/shivamtaneja2304/google-news sports/data.

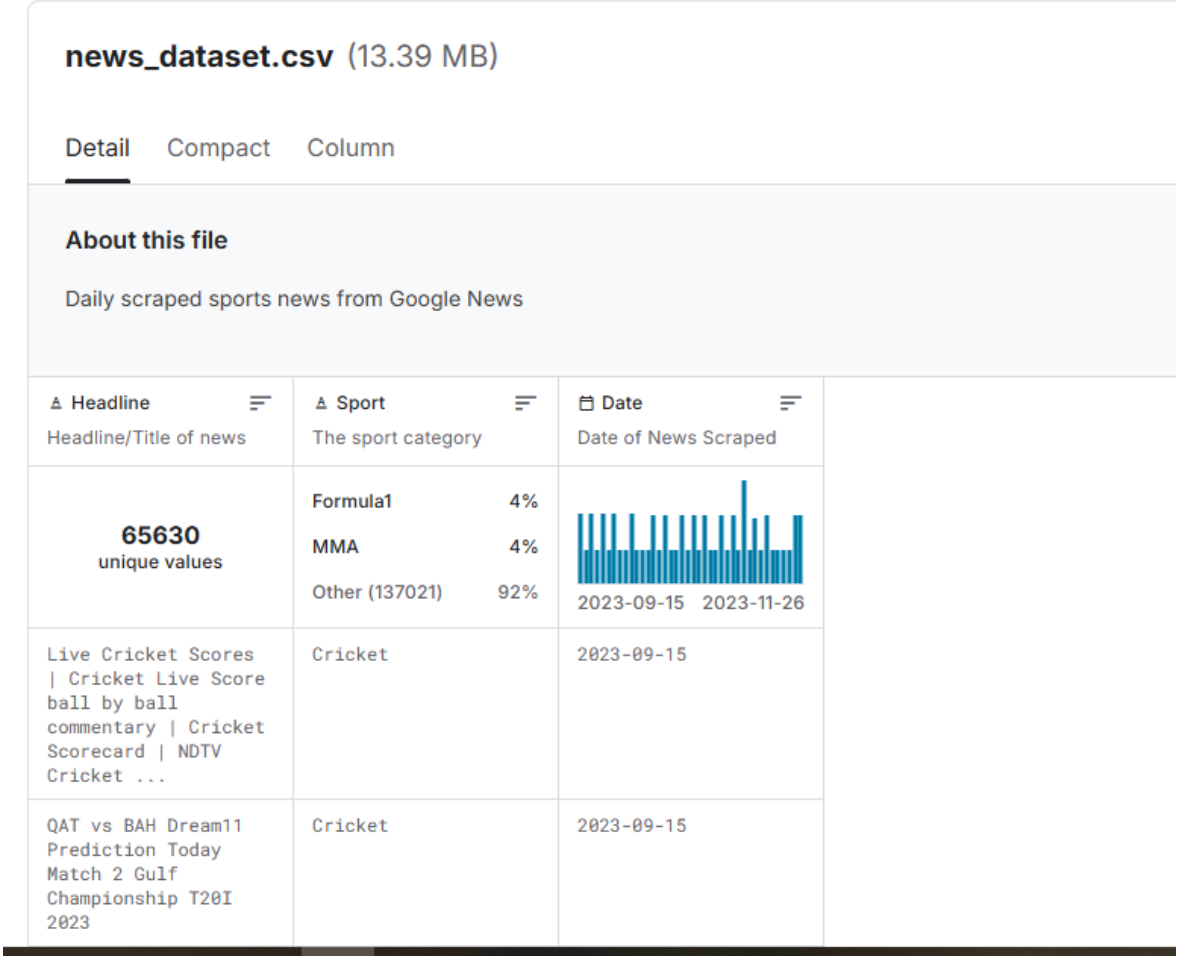

Figure 2.1: Fichier news dataset.csv

2.4 Prétraitement de données

Il est primordial de prétraiter les données textuelles pour assurer la qualité et l'uniformité des textes avant leur emploi dans les modèles de classification. Il facilite la purification des données brutes en retirant les éléments superflus (caractères spéciaux, répétitions, mots vides...), la standardisation des textes (minuscules, lemmatisation), ce qui améliore l'interprétation par les algorithmes. Cette étape participe directement à l'amélioration de la précision et de l'efficacité du système de classification thématique. Cela inclut [16] :

- **Élimination des caractères spéciaux, des chiffres et de la ponctuation:** Cette phase implique le retrait de tous les caractères non alphabétiques comme les marques de ponctuation (`.`, `!`, `?`, `etc.`), les symboles spéciaux (`@`,`#`,`%`,`$`,`etc.`) ainsi que les nombres. Ces composants n'offrent généralement aucune valeur sémantique pertinente dans le contexte d'une catégorisation thé-

matique.

- **Tokenisation des phrases en mots :** La tokenisation est l'opération qui vise à segmenter une chaîne de texte en éléments linguistiques identifiés comme tokens, habituellement des mots. Cette méthode facilite la conversion d'une phrase en une série de mots utilisables par des algorithmes. On utilise des bibliothèques telles que **NLTK (Natural Language Toolkit)** pour effectuer cette tâche [17].

- **Suppression des mots vides (stopwords):** Les stopwords sont des mots couramment utilisés dans une langue (tels que « **The** », « **and** », « **for** » en anglais) qui n'apportent que peu ou pas d'informations pour la classification. L'élimination de ceux-ci aide à diminuer le bruit présent dans les données et à se focaliser sur les termes pertinents.

- **Élimination des doublons et traitement des valeurs manquantes :** Dans le but d'éviter les préjugés durant l'apprentissage, les entrées répétées présentes dans le corpus sont éliminées. Ainsi, les lignes comportant des données manquantes **(comme un titre ou une catégorie absente)** sont éliminées pour éviter toute perturbation lors de l'entraînement des modèles.

2.5 Représentation vectorielle des textes

La transformation des textes en vecteurs est une phase cruciale dans tout système de traitement automatique du langage naturel **(TAL)**, y compris pour la segmentation thématique. Elle vise à convertir les textes (présentés sous forme de chaînes de caractères) en **représentations numériques** utilisables par des algorithmes d'apprentissage automatique. [18]

2.5.1 Le modèle TF-IDF

La technique de **TF-IDF** figure parmi les méthodes les plus couramment employées pour vectoriser les documents. Elle offre la possibilité d'estimer la valeur d'un mot dans un document, en considérant à la fois sa fréquence au sein de ce document et sa rareté dans l'ensemble du corpus.La formule de TF-IDF est donnée par [19] [20]:

$$\textbf{TF-IDF} = \textbf{TF} * \textbf{IDF}$$

TF (Term Frequency) : dénombre la fréquence d'apparition d'un terme dans un document spécifique. Plus un mot est couramment utilisé, plus il est jugé essentiel pour ce document.

IDF (Inverse Document Frequency) : évaluation du nombre de documents présents dans le corpus où un mot est mentionné. Si un mot figure de manière récurrente dans presque tous les documents **(par exemple : "le", "est","dans")**, son importance diminue puisqu'il n'est pas distinctif.

2.5.2 Application dans la segmentation thématique

Dans le cadre de ce projet, nous avons fait appel à l'approche **TF-IDF (Fréquence du Terme - Fréquence Inverse du Document)** afin de convertir les titres d'articles sportifs en vecteurs numériques que les modèles de classification peuvent manipuler. Cette méthode permet d'apprécier la pertinence d'un mot dans un document par rapport à l'ensemble du corpus, en mettant l'accent sur les termes propres à un sujet tout en minimisant l'importance des mots répétitifs et peu significatifs. En utilisant **TF-IDF**, nous avons réussi à dégager des caractéristiques significatives pour diverses disciplines sportives **(comme le football et le tennis)**, ce qui a favorisé l'entraînement des modèles et perfectionné la précision de la classification thématique.

2.6 Méthodes classiques de classification

2.6.1 Le classifieur Naive Bayes

2.6.1.1 Définition

L'algorithme Naïve Bayes, basé sur la règle de Bayes, est une méthode d'apprentissage simple qui postule que les attributs présentent une indépendance conditionnelle en considérant la classe. Même si cette supposition d'indépendance est fréquemment transgressée en situation réelle, Naïve Bayes propose tout de même une exactitude de classification compétitive. Sa large utilisation en pratique peut être attribuée à son efficacité de calcul et à diverses autres caractéristiques attrayantes [21] [22].

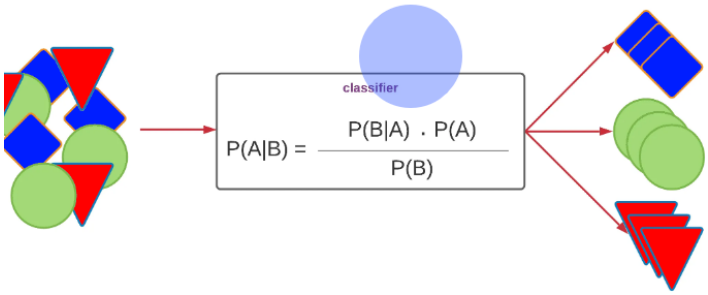

Figure 2.2: Illustration du fonctionnement du classifieur Naive Bayes

2.6.1.2 Les avantages

- Très rapide à entraîner et à exécuter.
- Performant même avec peu de données.
- Facile à interpréter.
- Particulièrement efficace avec des textes courts ou bien séparés thématiquement.

2.6.1.3 Utilisation de l'algorithme Naïve Bayes

Pour notre projet visant à classifier thématiquement les articles sportifs, nous avons fait appel à l'algorithme Naïve Bayes, un outil réputé pour son efficacité en matière de classification textuelle. Cet algorithme est basé sur le théorème de Bayes, en supposant de manière « **naïve** » que les caractéristiques (dans ce cas, les mots) sont indépendantes entre elles.

2.6.2 Les machines à vecteurs de support (SVM)

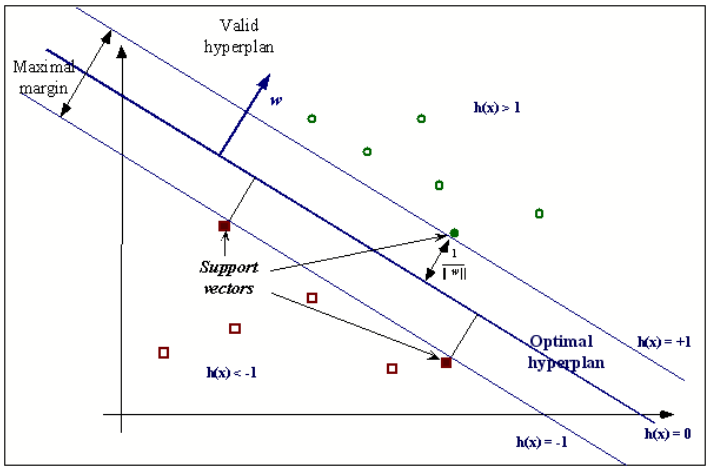

Figure 2.3: Le principe de SVM

[23]

Un SVM (Support Vector Machine) est un algorithme d'apprentissage supervisé employé pour classer les données. L'objectif est de déterminer un hyperplan qui divise deux classes de données en garantissant la plus grande marge possible. On désigne par vecteurs de support les points qui se situent à proximité de cet hyperplan. [24]

Lorsqu'il n'est pas possible de séparer les données de manière linéaire, on fait appel à une fonction noyau (kernel) pour projeter les données dans un espace où cette séparation peut être réalisée. [24]

Les SVM trouvent leur application dans divers domaines tels que la classification de textes, la détection des courriers indésirables ou le diagnostic médical, et ils produisent fréquemment des résultats satisfaisants. [24]

2.6.2.1 Algorithme 1 : Classification de textes par SVM

1. Charger les données textuelles (titres) et leurs étiquettes.
2. Nettoyer les textes (suppression de ponctuation, stopwords, lemmatisation...).
3. Appliquer la vectorisation TF-IDF sur les titres nettoyés.
4. Diviser les données en ensemble d'entraînement et de test.
5. Initialiser le modèle SVM avec des paramètres optimaux.
6. Entraîner le modèle SVM sur les données vectorisées.
7. Évaluer la performance du modèle (accuracy, F1-score...).

8. Utiliser le modèle pour prédire la classe de nouveaux articles.

2.6.2.2 utilisé SVM dans notre projet

Dans le contexte de notre initiative visant à regrouper les articles sportifs selon des critères thématiques, nous avons eu recours à l'algorithme SVM en raison de sa compétence à traiter efficacement les données textuelles sous forme de vecteurs et son aptitude à fournir une classification précise, même si les catégories se ressemblent ou utilisent un lexique commun.

Nous avons utilisé SVM sur les titres d'articles une fois qu'ils ont été prétraités et transformés en vecteurs à l'aide de la méthode TF-IDF. L'algorithme a été formé pour identifier le type de sport associé à chaque article en se basant sur ses attributs lexicaux. SVM, grâce à sa solidité et à son aptitude à représenter des limites complexes, a produit d'excellents résultats dans la catégorisation de sports tels que le football, le cricket ou le tennis.

2.7 Les Réseaux de neurones

2.7.1 MLP (Multilayer Perceptron)

Le Perceptron **Multicouche (MLP)** est une forme de réseau de neurones artificiels qui se classe parmi les algorithmes d'apprentissage supervisé. Il est composé de plusieurs strates : une couche d'entrée, une ou plusieurs couches cachées, et enfin, une couche de sortie. Chaque neurone utilise une fonction d'activation **(généralement ReLU ou sigmoïde)** pour acquérir des représentations non linéaires des données. [25]

Dans le contexte de ce projet, le Perceptron Multicouche a été employé en tant que troisième technique de classification, en plus de **Naive Bayes** et **SVM**, afin d'évaluer les performances dans la tâche de segmentation thématique des textes sportifs multilingues.

2.7.1.1 Étapes d'implémentation du MLP dans le projet

1. **Prétraitement et vectorisation** : Initialement, les textes ont été transformés en vecteurs numériques à travers le processus de TF-IDF, permettant ainsi leur

utilisation par des modèles d'apprentissage automatique.

2. **Division des données** : Le corpus a été fractionné en un lot de formation représentant 80 % des données et un lot de test qui constitue les 20 % restants.

3. **Entraînement du MLP** : Le modèle a été formé en se servant d'un classificateur MLP doté d'une architecture simple, comprenant une couche cachée de 100 neurones et un nombre maximal de 300 itérations.

4. **Prédiction et Évaluation** : L'évaluation du modèle a été réalisée en utilisant la précision, le score F1, le rappel et la matrice de confusion.

2.8 Traduction automatique multilingue des textes

2.8.1 mBERT (Multilingual Bidirectional Encoder Representations from Transformers)

Le modèle **Multilingual BERT (mBERT)** est une déclinaison multilingue du modèle BERT créé par **Google**. À l'opposé du modèle BERT initial, qui a été formé uniquement sur des documents en anglais, **mBERT** a bénéficié d'une formation basée sur **Wikipédia** dans **104 langues**, y compris l'arabe, le français et diverses langues moins représentées. [26]

mBERT, en raison de son aptitude à assimiler des représentations linguistiques communes à plusieurs langues, est couramment employé pour diverses tâches de traitement automatique des langues multilingues comme la classification de texte, l'identification d'entités nommées **(NER)** et la traduction assistée. [27]

2.8.2 NLLB (No Language Left Behind)

C'est un projet lancé par Meta AI (anciennement Facebook AI) pour améliorer la traduction automatique dans plus de 200 langues, y compris des langues peu représentées ou à faibles ressources, qui sont souvent négligées par les systèmes de traduction traditionnels.

Son objectif principal est de concevoir des systèmes de traduction de haute qualité couvrant plus de **200 langues**, avec une attention particulière portée aux **langues à faibles ressources**. [28]

2.9 Conclusion

Ce chapitre détaille la réalisation technique du système de segmentation thématique, couvrant toutes les étapes, de la préparation des données à l'évaluation des modèles. Nous avons testé et comparé trois classificateurs : **Naive Bayes**, **SVM** et **MLP**. Les résultats indiquent que **MLP** présente de bonnes performances, cependant cela implique un temps de calcul plus important. Cette réalisation confirme l'approche suggérée et pave la route pour des améliorations à venir.

Chapter 3

Implémentation

3.1 Introduction

Ce chapitre se focalise sur l'application pratique du système suggéré. Une fois que nous avons défini les concepts théoriques et les modèles utilisés, nous en venons ici à leur mise en œuvre pratique. Nous décrivons l'environnement de développement, les outils sélectionnés et les bibliothèques Python qui ont contribué à la réalisation de notre système de segmentation thématique. Chaque phase technique, depuis la préparation des données jusqu'à l'évaluation des résultats, y sera décrite en détail.

3.2 Environnement et outils d'implémentation

3.2.1 Matérial

Caractéristiques	**Poste de travail N°01**	**Poste de travail N°02**
PC	Hp	DELL
Système d'exploitation	Windows 11 Professionnel	Windows 10 Professionnel
Processeur	Intel(R) Core(TM) i5-8365U CPU @ 1.60GHz	Intel(R) Core(TM) i5-6200U CPU @ 2.30GHz , 2.40 GHz
RAM	16,00 Go	8,00 Go
Type de système	SE 64 bits	SE 64 bits

Table 3.1: Caractéristiques des matériels utilisés

3.2.2 Langage de programmation :

Python est un langage de programmation interprété, polyvalent et libre, apprécié pour sa syntaxe simple et sa clarté. On le retrouve fréquemment dans les secteurs de la science des données, du développement web, de l'automatisation et principalement dans le domaine du **traitement automatique des langues (TAL)**. Pour ce projet, j'ai surtout employé les versions **Python 3.13.3** et **Python 3.9.0**, qui présentent une large compatibilité avec les bibliothèques actuelles indispensables pour le prétraitement et la classification de textes en plusieurs langues. [29] [30]

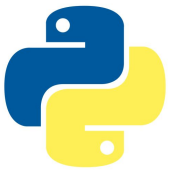

Figure 3.1: Logo de Python

3.2.3 Environnement de programmation :

3.2.3.1 Visual Studio Code

Un éditeur de code à la fois léger et performant, qui peut être étendu grâce à une vaste gamme d'extensions. Ce logiciel a été principalement utilisé pour rédiger des scripts Python, structurer le projet et s'intégrer avec Git. [31]

Figure 3.2: Logo de VS Code

3.2.3.2 Jupyter Notebook

Un environnement interactif largement employé en science des données, qui offre la possibilité d'intégrer du code, des représentations graphiques et des annotations au sein d'un unique document. Il s'est révélé particulièrement utile lors des étapes de prototypage, d'exploration des données et de représentation des résultats. [32]

Figure 3.3: Logo de Jupyter

3.2.4 Les packages Python utilisés

Le développement a essentiellement été effectué en Python, en faisant appel à diverses bibliothèques spécialisées, y compris :

1. **Bibliothèque NLTK:** Le Natural Language Toolkit **(NLTK)** est une bibliothèque Python open source destinée au traitement automatique du langage naturel **(TALN)**. Elle propose divers outils pour analyser et manipuler des textes, comme la tokenisation, la lemmatisation, l'analyse grammaticale ou la classification. Utilisée en recherche et en enseignement, **NLTK** permet aussi l'accès à plusieurs corpus linguistiques, ce qui en fait une ressource précieuse pour les projets en linguistique computationnelle et en intelligence artificielle. [25]

2. **Pandas:** est une bibliothèque Python très répandue en matière de science des données et qui est de plus open source. Elle offre la possibilité de charger, épurer et structurer des données en DataFrames, ce qui rend le prétraitement des textes avant analyse plus aisé. Pandas, en s'intégrant avec d'autres outils tels que NumPy et Scikit-learn, représente une phase essentielle dans le prétraitement des données pour les projets de traitement du langage naturel. [33]

3. **Numpy:** est une bibliothèque Python open source, développée par Travis Oliphant en 2005, dédiée à la manipulation performante de tableaux multidimensionnels. Elle offre aussi des capacités avancées pour réaliser des opérations mathématiques,

en particulier en algèbre linéaire comme le traitement matriciel. Cela la rend indispensable dans plusieurs domaines tels que l'analyse de données, l'apprentissage automatique et le traitement du langage naturel. [34]

4. **Matplotlib:** est une bibliothèque Python à code source ouvert consacrée à la représentation graphique des données. Elle offre la possibilité de générer une vaste gamme de graphiques, comme des courbes, des histogrammes, des diagrammes de dispersion ou des heatmaps, rendant ainsi l'analyse et la visualisation des données plus aisées. On utilise fréquemment cet outil en complément de Pandas et NumPy, principalement pour l'exploration des données textuelles et la présentation des résultats issus des projets de science des données et d'analyse du langage naturel. [35]

5. **Wordcloud:** Le nuage de mots, aussi connu sous le nom de Word Cloud, est une représentation visuelle des termes les plus fréquemment utilisés dans un ensemble de textes. Les termes les plus couramment utilisés sont généralement représentés en taille de police plus grande et en position dominante, tandis que les termes moins courants sont présentés en taille de police réduite et en position subalterne. [36]

6. **Seaborn:** Seaborn est une librairie Python conçue pour la représentation graphique de données. Elle offre la possibilité de concevoir aisément des graphiques statistiques à la fois clairs et esthétiques. [37]
   Dans le cadre de ce projet, elle a été employée pour présenter les matrices de confusion, dans le but d'évaluer visuellement l'efficacité des modèles de classification. Grâce à son intégration avec pandas et matplotlib, c'est un outil utile et performant pour l'analyse des résultats.

7. **Joblib:** est une librairie Python conçue pour l'enregistrement et le chargement efficace de modèles et de données de grande taille, notamment celles comprenant des matrices NumPy. [38]

3.3 Génération du nuage de mots

Pour une analyse exploratoire, nous avons eu recours à la bibliothèque **WordCloud** de Python afin de représenter graphiquement les 100 mots les plus courants dans les titres des articles. Cette représentation permet d'identifier de manière visuelle les termes prédominants par catégorie **(Tennis, Cricket, Football)**, ce qui facilite l'appréhension des particularités lexicales propres à chaque discipline avant la phase de classement.

Après l'application de cette bibliothèque sur les deux bases de données, nous avons obtenu les nuages suivants :

Figure 3.4: Nuage de mots — Football

Dans cette image, nous présentons un nuage de mots élaboré à partir de titres en lien avec le football. Les termes affichés en grands caractères, tels que **"football"**, **"team"** et **"fantasy"**, correspondent aux mots les plus fréquemment rencontrés, justifiant ainsi leur mise en valeur. À l'inverse, les mots apparaissant en plus petits caractères, tels que **"play"**, **"night"**, **"game"** sont ceux dont la fréquence est moindre. À travers la création de ce nuage de mots, nous visons à offrir une visualisation rapide des termes dominants et à dégager les thématiques principales abordées dans les titres analysés.

Figure 3.5: Nuage de mots — Cricket

Dans cette image, nous présentons en avant un nuage de mots créé à partir de titres associés au cricket. Les mots qui apparaissent le plus souvent, comme " **Cricket** ", " **World** ", " **Cup** ", sont présentés en lettres majuscules, illustrant leur fréquence dans les titres examinés. Au contraire, des termes tels que " **run** ", " **score** ", " **live** " ou encore " **pitch** " sont affichés en caractères plus petits du fait de leur occurrence moins importante. Ce nuage de mots offre une représentation visuelle rapide des idées dominantes associées aux tournois internationaux, aux équipes nationales et aux performances individuelles, mettant en lumière les principaux sujets du cricket dans le corpus examiné.

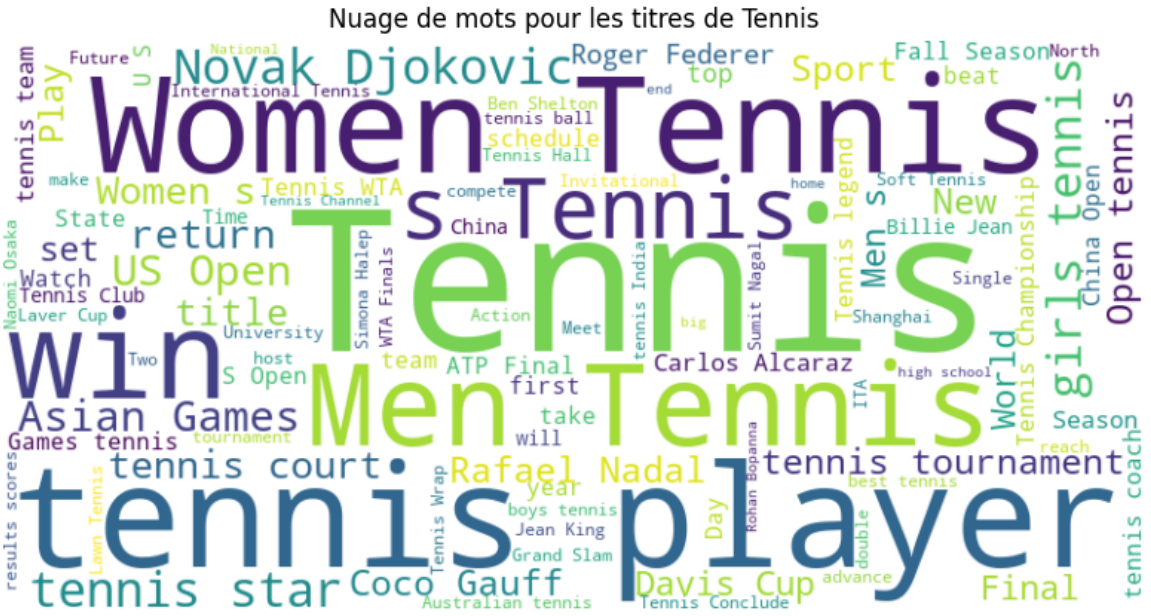

Figure 3.6: Nuage de mots — Cricket

Dans cette image, nous présentons un nuage de mots élaboré à partir de titres en lien avec le tennis. Les mots les plus mis en avant, comme **"tennis"**, **"player"**, **"win"**, **"women"**, et **"men"**, apparaissent en grands caractères, indiquant leur forte présence dans les titres collectés. D'autres termes tels que **"coach"**, **"court"**, **"title"** ou **"return"** apparaissent en plus petit, car ils sont moins fréquents. Ce nuage de mots permet d'identifier rapidement les thématiques majeures abordées, telles que les performances des joueurs, les compétitions internationales, et la distinction entre les tournois masculins et féminins, constituant ainsi une base pertinente pour l'analyse thématique des textes liés au tennis.

3.4 Diagramme des catégories de sport

Dans notre travail sur la classification thématique des écrits sportifs, il est crucial de saisir la structure de la base de données examinée. Le diagramme présenté a été réalisé dans le but de représenter la distribution des titres selon la discipline sportive. Cette représentation offre une vision précise de la fréquence de chaque discipline sportive dans l'ensemble des données, crucial pour juger de la pertinence des sujets. Ce diagramme, en pinpointant les sports qui sont le plus et le moins présents, permet de prévoir d'éventuels déséquilibres dans l'analyse thématique et d'ajuster les

modèles de traitement du langage naturel en conséquence.

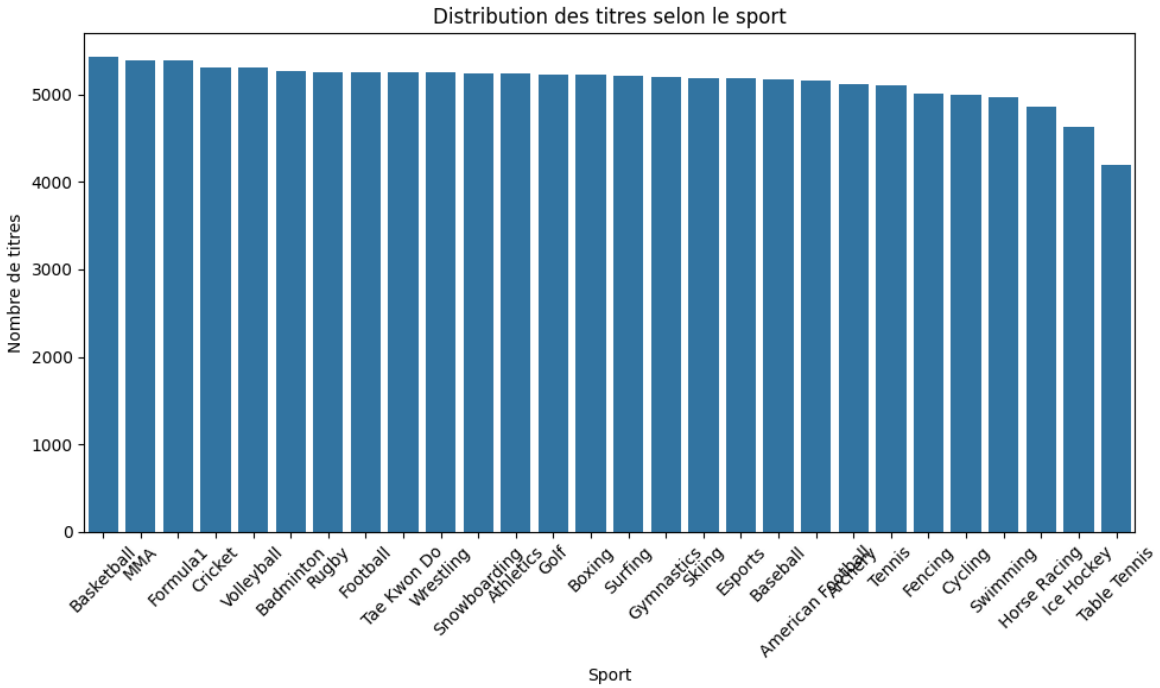

Figure 3.7: Distribution des titres selon le sport

Nous notons que des sports comme **le basketball**, **le MMA** et **la Formule 1** sont fortement représentés, ce qui reflète leur importante visibilité médiatique dans les sources rassemblées. En revanche, des sports tels que le tennis de table ou le hockey sur glace semblent beaucoup moins courants. Cette représentation inégale met en relief la nécessité d'intégrer la diversité des sujets dans le traitement automatique des textes pour prévenir une préférence biaisée envers les sports les plus médiatisés.

3.5 Application des algorithmes de classification de données

Dans cette partie, nous expliquons comment les algorithmes de classification sont mis en œuvre sur notre collection de textes liés au sport. Suite au prétraitement et à la vectorisation des données, nous avons réparti l'ensemble en deux segments : 80 % pour l'entraînement et 20 % pour le test. Cette distribution, démontrée à la figure 3.9, facilite l'entraînement des modèles sur une part importante des données tout en gardant un sous-ensemble pour tester leur efficacité.

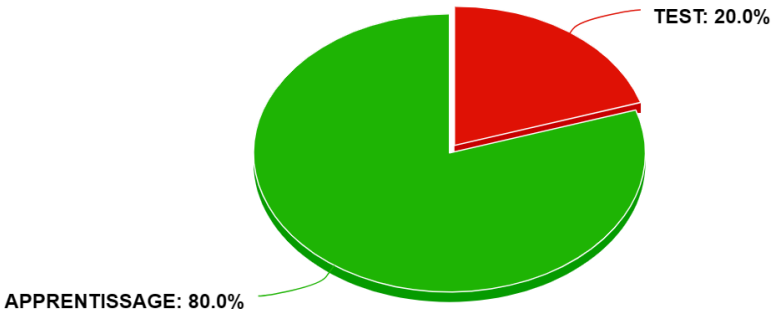

Figure 3.8: Découpage d'un ensemble de données en 80% pour l'entrainement et 20 % pour le test

- Les fonctions utilisées pour préparer et nettoyer les données brutes ont été détaillées précédemment dans le chapitre 2 (prétraitement) ainsi que dans la section 3.2. Ces opérations ont joué un rôle essentiel dans l'amélioration de la qualité des textes, en les rendant exploitables pour les étapes de vectorisation et de classification.
- Afin d'illustrer concrètement l'impact de ce traitement, le tableau ci-dessous présente quelques exemples de titres avant et après nettoyage. On y observe la suppression des éléments non pertinents et la simplification du contenu textuel, ce qui permet d'obtenir une représentation plus cohérente et normalisée pour l'analyse automatique.

Titre original	**Titre Nottoyer**
Live Cricket Scores | Cricket Live Score ball by ball commentary | Cricket Scorecard | NDTV Cricket ..	live cricket score cricket live score ball ball commentary cricket scorecard ndtv cricket
QAT vs BAH Dream11 Prediction Today Match 2 Gulf Championship T20I 2023	qat v bah dream prediction today match gulf championship ti
IND: 259-9 (49.5) | IND VS BAN Live Cricket Score and Updates, Asia Cup 2023 Super 4: Bangladesh Win	Iind ind v ban live cricket score update asia cup super bangladesh win
RGD vs CAS Dream11 Prediction, Fantasy Cricket Tips, Playing XI, Pitch Report, & Injury Updates for ECS Rome T10 2023, Match 36	rgd v ca dream prediction fantasy cricket tip playing xi pitch report injury update ec rome match

Table 3.2: une comparaison entre le titre original et le titre nettoyé

3.6 Interface Web de classification multilingue avec traduction

Nous avons développé cette interface utilisateur en utilisant HTML. Elle autorise l'utilisateur à entrer un texte dans l'une des trois langues disponibles : l'arabe, le français ou l'anglais. Après avoir inséré le texte, l'utilisateur a la possibilité de cliquer sur le bouton « **Classer le texte** ». Nous engageons ensuite un processus de détection automatique de la langue, traduisons le texte en anglais si besoin, avant de le diriger vers le modèle de classification pour déterminer le sujet principal du contenu (comme le football, la natation, le basketball, etc.).

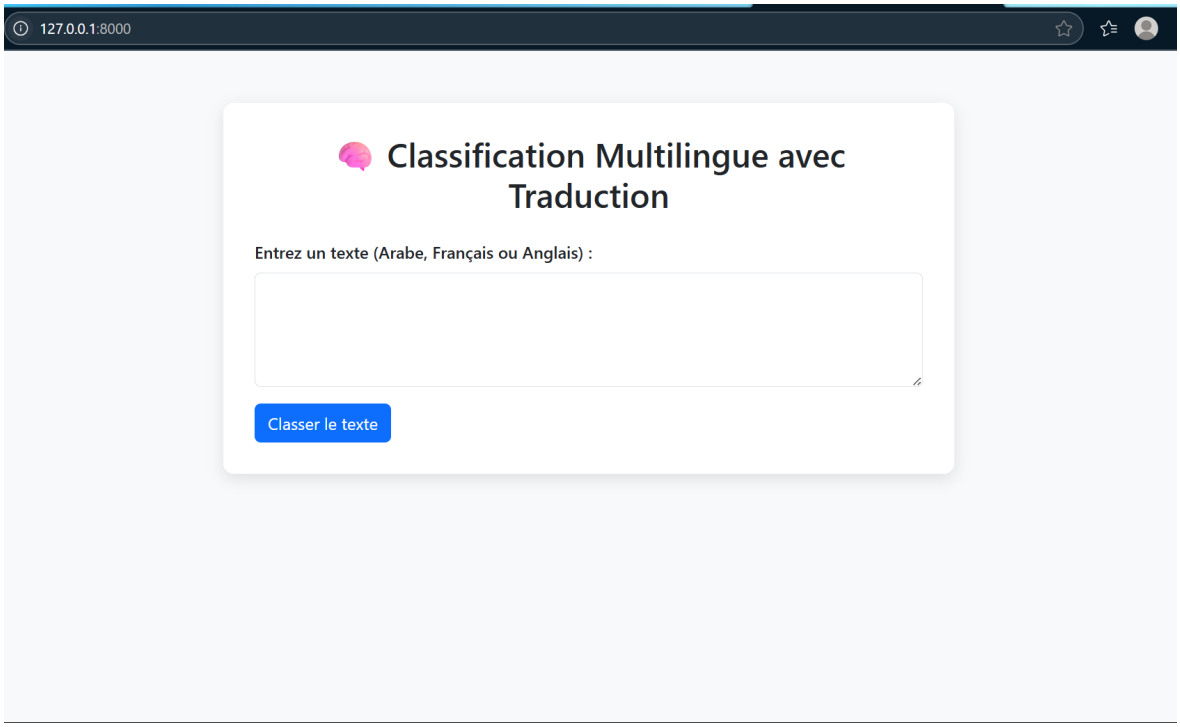

Figure 3.9: Interface web de classification multilingue avec traduction

3.7 Conclusion

Ce chapitre a synthétisé la réalisation technique de notre projet, allant du traitement des données à l'application des algorithmes. Les résultats obtenus corroborent l'efficacité de notre méthode pour classer thématiquement les textes sportifs en plusieurs langues.

Chapter 4

Évaluation des résultats et discution

4.1 Introduction

Dans ce chapitre, nous allons présenter et analyser les résultats issus de nos expérimentations.

4.2 Métriques d'évaluation

Pour évaluer l'efficacité des modèles de classification, nous avons recouru à diverses métriques fréquemment utilisées dans le domaine de l'apprentissage automatique :

- **Accuracy :** ratio des prédictions justes.
- **Précision (Precision) :** proportion de prédictions positives vraies par rapport à l'ensemble des prédictions positives réalisées.
- **Rappel (Recall)**: ratio entre les vraies positives et l'ensemble des instances positives.
- **F1-score :** une valeur moyenne harmonique entre la précision et le rappel, particulièrement pertinente pour les classes déséquilibrées.
- **Matrice de confusion :** tableau indiquant les prédictions justes et fausses pour chaque catégorie.

4.3 Résultats des modèles de classification

4.3.1 Modèle Naive Bayes

On a utilisé l'algorithme Naive Bayes sur les données qui ont été vectorisées à l'aide de la méthode **TF-IDF**.Il s'est montré rapide à entraîner, mais avec des performances modérées.

Classe	**Précision**	**Rappel**	**F1-score**	**Support**
Cricket	0.92	0.93	0.92	1069
Football	0.84	0.73	0.78	1059
Tennis	0.91	0.82	0.86	1002
Moyenne (macro)	0.88	0.88	0.88	-
Moyenne (pondérée)	0.88	0.88	0.88	28804

Table 4.1: Rapport de classification du modèle - Naive Bayes

• Le modèle Naive Bayes a montré des performances satisfaisantes, notamment grâce à sa simplicité et sa rapidité. Il a atteint une accuracy globale de **87.87%**.

4.3.2 Modèle SVM (Support Vector Machine)

Le modèle SVM linéaire a produit de meilleures performances, présentant une séparation plus distincte entre les classes.

Classe	**Précision**	**Rappel**	**F1-score**	**Support**
Cricket	0.97	0.97	0.97	1069
Football	0.90	0.90	0.90	1059
Tennis	0.96	0.95	0.95	1002
Moyenne (macro)	0.94	0.93	0.94	-
Moyenne (pondérée)	0.94	0.93	0.94	28804

Table 4.2: Rapport de classification du modèle SVM (extrait)

• Le modèle SVM a affiché les meilleures performances, avec une précision générale de **93,50%**. Il a réussi à distinguer clairement les divers thèmes sportifs.

4.3.3 Modèle MLP (Multilayer Perceptron)

Le réseau de neurones MLP a été le plus performant en termes de précision globale.

Classe	**Précision**	**Rappel**	**F1-score**	**Support**
Cricket	0.97	0.97	0.97	1069
Football	0.87	0.87	0.87	1059
Tennis	0.93	0.95	0.94	1002
Moyenne (macro)	0.93	0.93	0.93	-
Moyenne (pondérée)	0.94	0.93	0.93	28804

Table 4.3: Rapport de classification du modèle réseau de neurones (extrait)

• Le réseau de neurones MLP a également démontré une précision générale élevée de **93,35%**, comparable à celle obtenue par SVM.

4.3.4 Analyse des matrices de confusion

L'analyse des matrices de confusion nous permet une évaluation précise des performances des trois modèles de classification explorés dans cette étude : SVM, Naive Bayes et MLP (réseau de neurones multicouche). Ces tableaux illustrent la compétence de chaque modèle à catégoriser correctement les articles en fonction de leurs thèmes sportifs associés.

4.3.4.1 Matrice de confusion – Naive Bayes

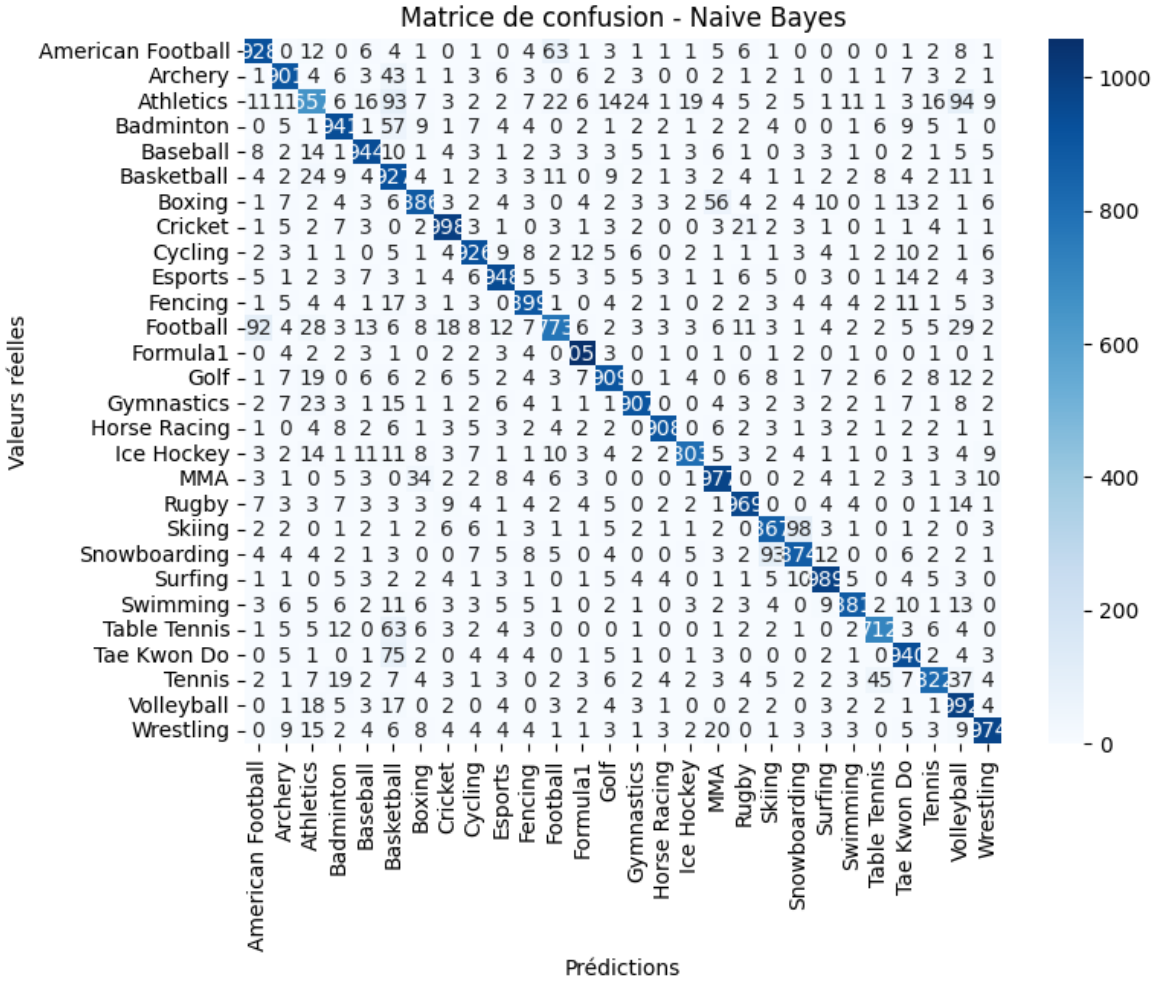

Figure 4.1: Matrice de confusion du modèle Naive Bayes

Le modèle **Naive Bayes**, qui repose sur des principes de probabilité, présente une performance inférieure :

- De nombreuses classes présentent des confusions significatives. Par exemple, **l'Athlétisme**, **le Football** et **le Basketball** sont fréquemment confondus avec d'autres disciplines sportives.
- Néanmoins, des classes comme **le Wrestling (974)**, **le Golf (900)** et **le Snowboarding (972)** demeurent clairement définies.
- Ce modèle est plus réceptif aux déséquilibres de classes et aux superpositions lexicales, ce qui entrave son efficacité dans ce contexte multilingue et thématique.

4.3.4.2 Matrice de confusion – SVM

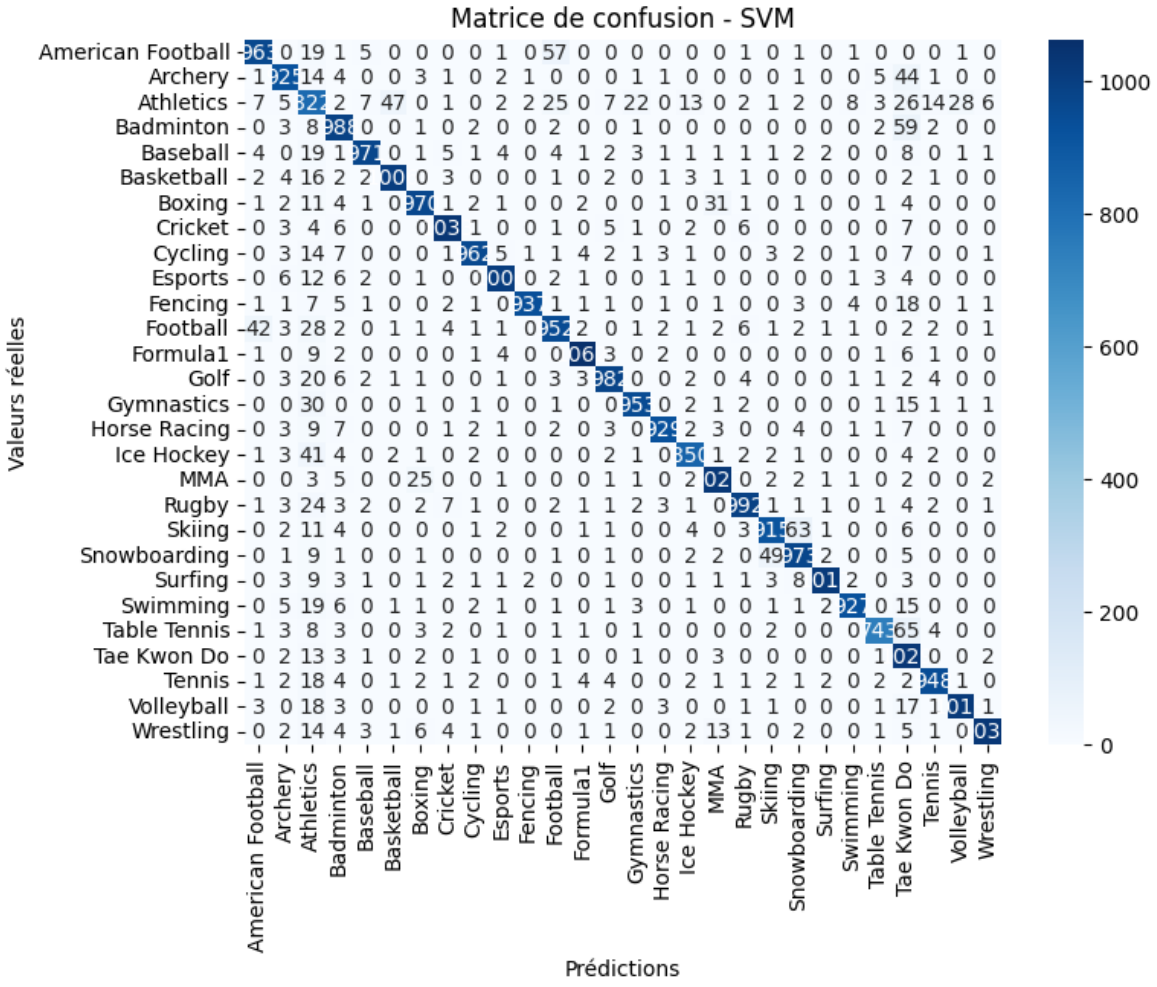

Figure 4.2: Matrice de confusion du modèle SVM

Le **Support Vector Machine (SVM)** a démontré une performance globale satisfaisante. On note :

- Une reconnaissance remarquable pour des catégories telles que **la lutte (1036)**, **le snowboard (972)**, **la natation (822)** et **le tennis de table (765)**.
- Il peut y avoir des confusions légères entre certains sports qui sont thématiquement similaires **(par exemple : Athlétisme, Tennis, Gymnastique).**
- Le modèle **SVM** parvient efficacement à distinguer la plupart des classes, ce qui le rend un choix solide pour la segmentation thématique.

4.3.4.3 Matrice de confusion – MLP (Réseau de neurones)

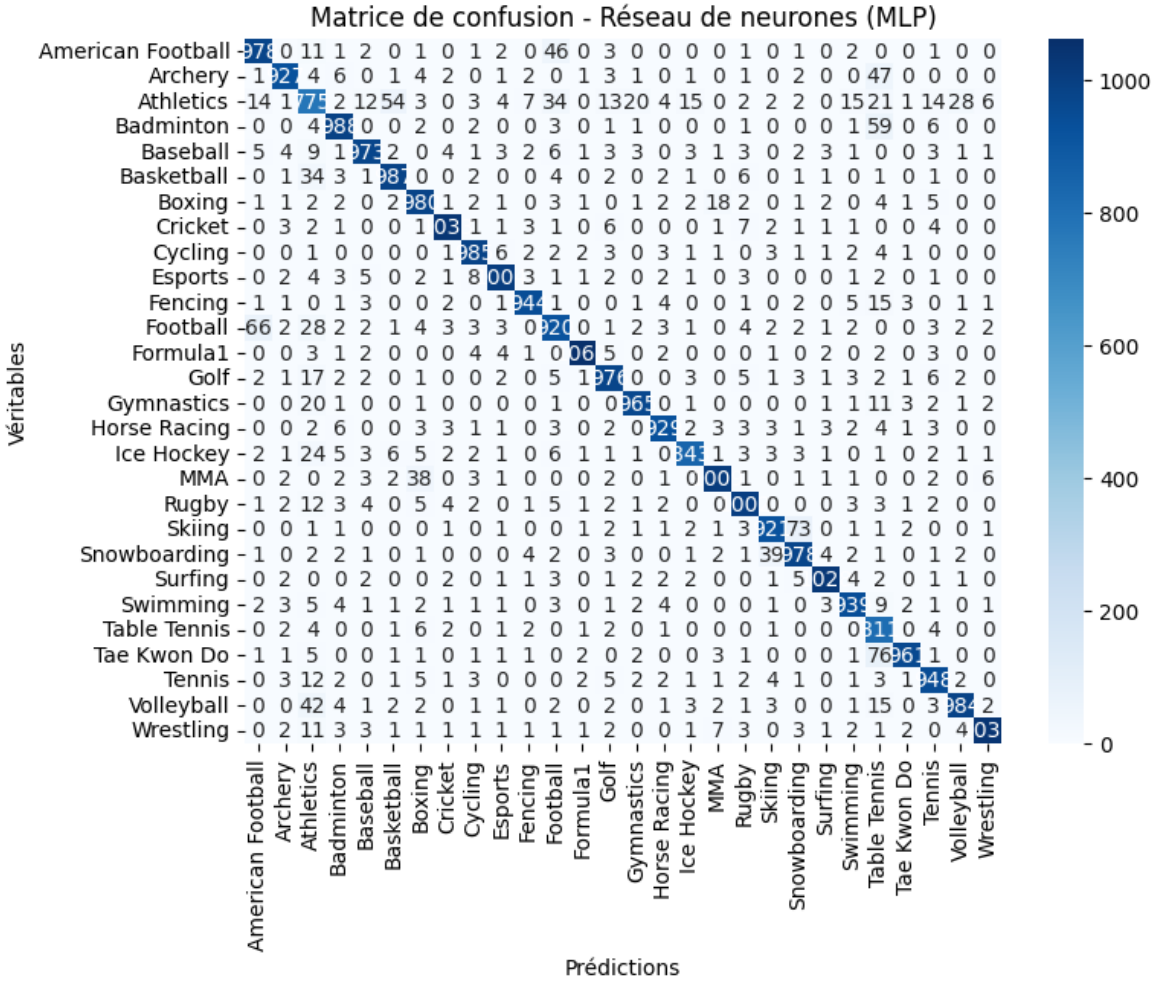

Figure 4.3: Matrice de confusion du modèle MLP

Le **Multi-Layer Perceptron (MLP)** a montré les meilleurs résultats dans la majorité des cas :

- Très bonne classification pour **Swimming (766)**, **Snowboarding (974)**, **Table Tennis (811)** et **Wrestling (1033)**.
- Moins de confusion que **Naive Bayes**, et parfois plus performant que **SVM** pour certaines classes.
- Le modèle semble mieux capter la complexité sémantique des textes, notamment grâce à sa capacité d'apprentissage non linéaire.

4.3.5 Comparaison des modèles

Modèle	**Accuracy**	**F1-score (macro)**	**Remarques**
Naive Bayes	87.87 %	0.88	Simple, rapide
SVM	**93.50 %**	**0.94**	Très précis, stable
MLP	93.35 %	0.93	Performant, mais coûteux

Table 4.4: Comparaison des performances des modèles de classification

• Sur les trois modèles examinés, nous avons observé que le **SVM (Support Vector Machine)** affiche les meilleures performances, avec un taux de précision général de **93,50 %** et un score **F1** de **0,94**. Il a démontré une stabilité et une efficacité notables pour différencier les différentes catégories, même face à des similarités lexicales. Le modèle **MLP (Perceptron Multicouche)** se classe en deuxième place avec une précision légèrement plus basse **(93,35 %)**, néanmoins, il démontre une aptitude notable à saisir des relations complexes entre les données, bien qu'il exige un temps de formation plus important. Pour conclure, le modèle **Naive Bayes**, malgré sa rapidité et sa simplicité de mise en œuvre, est moins précis avec un rendement global de **87,87 %**.Cependant, il peut être bénéfique dans des situations où la vitesse prime sur l'exactitude. Nous en déduisons donc que le **SVM** est le modèle le mieux approprié pour notre mission de classification thématique des textes sportifs dans plusieurs langues.

4.3.6 Optimisation du modèle SVM

Pour améliorer les performances de classification, nous avons procédé à une phase d'optimisation sur le modèle SVM. Cette phase a impliqué la diminution du nombre initial de **28 catégories sportives** à **18 catégories choisies**, en supprimant celles qui étaient soit faiblement représentées dans les résultats.

Cette sélection nous a permis de diminuer la complexité du jeu de données et d'obtenir une séparation plus nette entre les classes restantes. Le modèle **SVM** a ensuite été réentraîné exclusivement sur ce sous-ensemble filtré.

L'efficacité a considérablement progressé, affichant **une précision globale** de **96,61 %**, comme le démontrent la mise à jour de la matrice de confusion et le rapport de classification joints ci-dessous :

4.3.6.1 La matrice de confusion

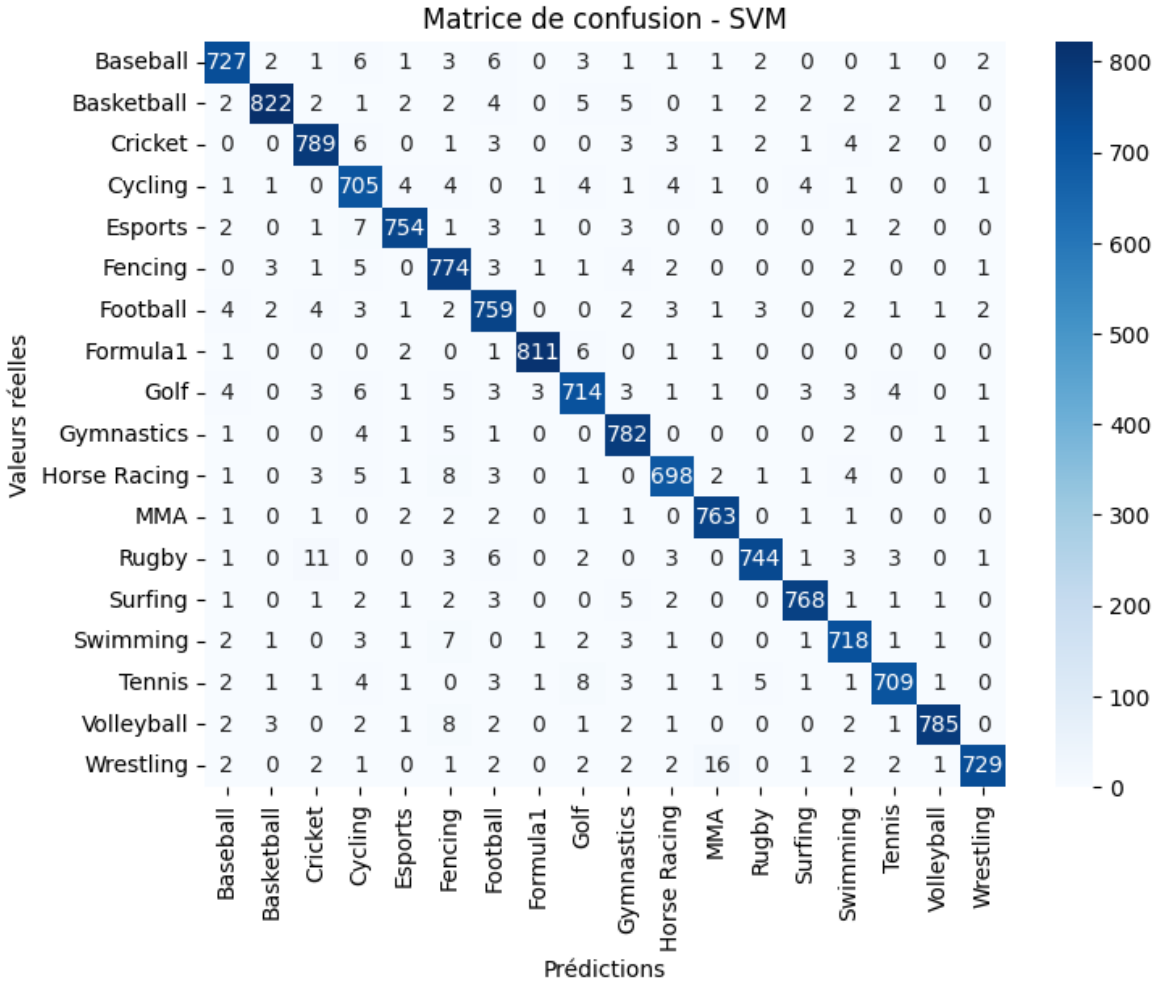

Figure 4.4: Matrice de confusion du modèle SVM

La matrice de confusion générée démontre une performance remarquable du modèle à effectuer des prédictions correctes concernant les catégories sportives choisies. Une diagonale prononcée est notée, ce qui signifie que la plupart des articles ont été rangés correctement.

Les fautes de classification sont minimes et touchent essentiellement des disciplines qui partagent des ressemblances lexicales ou contextuelles. Par exemple, des confusions mineures pourraient surgir entre des disciplines comme le **Basketball** et le **Football**, ou entre le **Cycling** et la **Formula1**, à cause de l'emploi de termes techniques similaires.

4.3.6.2 Matrice de confusion binaire globale

Pour apprécier les performances globales du modèle de classification binaire mis en œuvre pour tous les sports, nous avons produit une matrice de confusion que vous trouverez ci-dessous.

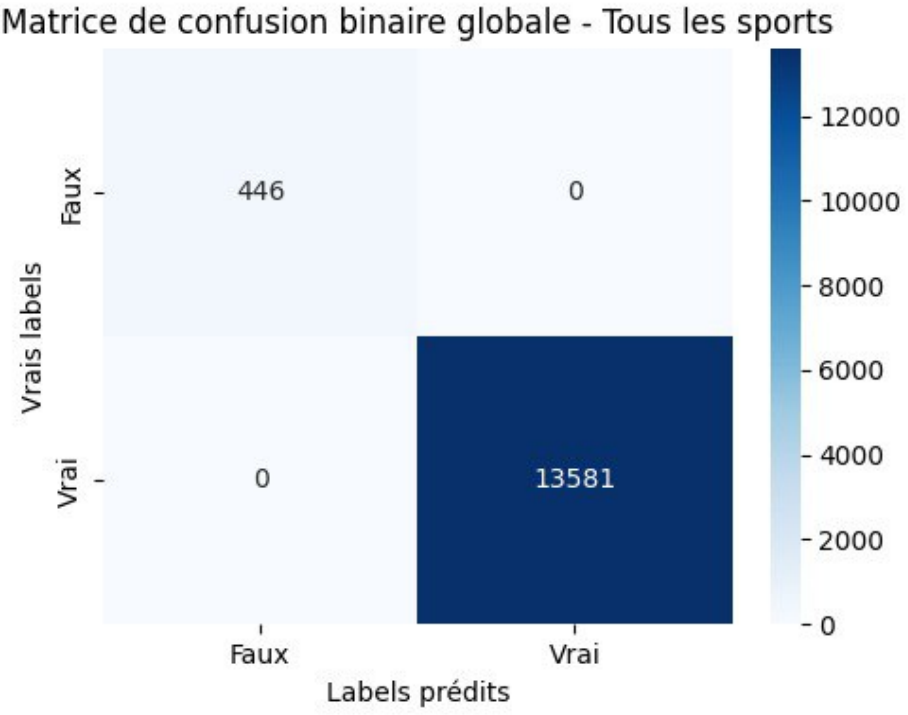

Figure 4.5: Matrice de confusion binaire globale

La matrice indique que le modèle réussit à séparer parfaitement les deux catégories **(Vrai et Faux)**. Effectivement, il n'y a pas eu de mauvaise classification : tous les cas de la classe **«Faux»** ont été justes **(446)**, tout comme ceux de la classe **«Vrai» (13 581)**. Cela indique que le modèle obtient une précision et un rappel de **100 %**, illustrant ainsi une excellente aptitude à la généralisation dans ce cas particulier.

4.3.6.3 Le rapport de classification

Catégorie	**Précision**	**Rappel**	**F1-score**	**Support**
Cricket	0.96	0.97	0.97	815
Football	0.94	0.96	0.95	790
Tennis	0.97	0.95	0.96	743
Moyenne	**0.96**	**0.96**	**0.96**	—

Table 4.5: Rapport de classification du modèle SVM optimisé (extrait)

4.3.7 Comparaison entre le modèle SVM initial et le modèle SVM optimisé

Dans le contexte de notre recherche, nous avons mis en parallèle les performances du modèle **SVM** formé sur les **28 catégories intégrales** et celles du **modèle SVM affiné**, formé uniquement sur **18 catégories choisies**. Cette comparaison souligne l'importance de sélectionner les classes en amont.

Modèle	**Nb de catégories**	**Précision globale**	**Commentaires**
SVM initial	28	93.50 %	Bonne performance générale, mais impactée par la présence de classes peu représentées et des confusions fréquentes entre certaines disciplines.
SVM optimisé	18	96.61 %	Précision améliorée grâce à l'élimination des classes ambiguës ou sous-représentées, ce qui a renforcé la stabilité du modèle.

Table 4.6: Comparaison entre le modèle SVM initial et le modèle SVM optimisé

Après cette analyse comparative, nous avons noté une amélioration significative des performances du modèle suite à son optimisation. En limitant le nombre de catégories aux plus représentatives et moins susceptibles de prêter à confusion, nous avons facilité l'apprentissage par le classifieur **SVM** des traits distinctifs de chaque classe. Cette décision stratégique a conduit à une amélioration de plus de 3 points en termes de précision globale, illustrant ainsi l'effet bénéfique du filtrage des classes sur la solidité et la fiabilité du modèle.

4.4 Intégration de la traduction automatique dans le processus de segmentation thématique

4.4.1 Contexte et motivation

Dans le contexte de notre projet de segmentation thématique portant sur un ensemble multilingue d'articles sportifs, nous avons dû faire face à une grande diversité

linguistique. Le corpus comprenait des écrits en diverses langues, y compris **l'anglais, le français, Arabe** et d'autres. Cette variété a engendré des défis pour standardiser le traitement du texte et mettre en œuvre des méthodes de classification. Pour harmoniser la représentation linguistique des données, nous avons inclus un processus de traduction automatique, dont le but est de convertir tous les textes en anglais, langue pivot sélectionnée pour son abondance en ressources **NLP**.

4.4.2 Techniques utilisées

Dans un premier temps, nous avons utilisé le modèle **mBERT (Multilingual BERT)** proposé par **Google**. Ce modèle multilingue pré-entraîné sur plus de 100 langues permet une représentation sémantique partagée entre différentes langues. Il s'avère particulièrement utile pour des tâches de traitement automatique des langues (TAL), notamment dans un contexte multilingue, en facilitant l'extraction de caractéristiques communes aux textes traduits.

Toutefois, les performances de **mBERT** n'ont pas été entièrement satisfaisantes, en raison notamment de la traduction parfois inexacte des termes spécifiques au milieu sportif.

Pour perfectionner la qualité de notre traduction, nous avons aussi testé le modèle **NLLB (No Language Left Behind)**, également conçu par **Meta AI**. Ce modèle, expert en traduction multilingue, s'est montré plus efficace que **mBERT**, offrant des traductions plus précises et mieux ajustées à l'univers sportif.

4.5 Résultat d'exécution

Après l'intégration de la phase de traduction, nous avons rendu le système capable de traiter des textes rédigés en plusieurs langues **(arabe, français, anglais)**. Cette étape nous a permis de traduire automatiquement tout texte saisi vers l'anglais avant de le soumettre à la phase de classification thématique. L'objectif de cette traduction est de normaliser les entrées multilingues afin d'assurer une meilleure cohérence dans le traitement et d'améliorer la précision des prédictions. Grâce à cette amélioration, nous avons pu classifier correctement le thème principal d'un texte, même lorsqu'il comporte des segments dans différentes langues, comme le montre la figure ci-dessous.

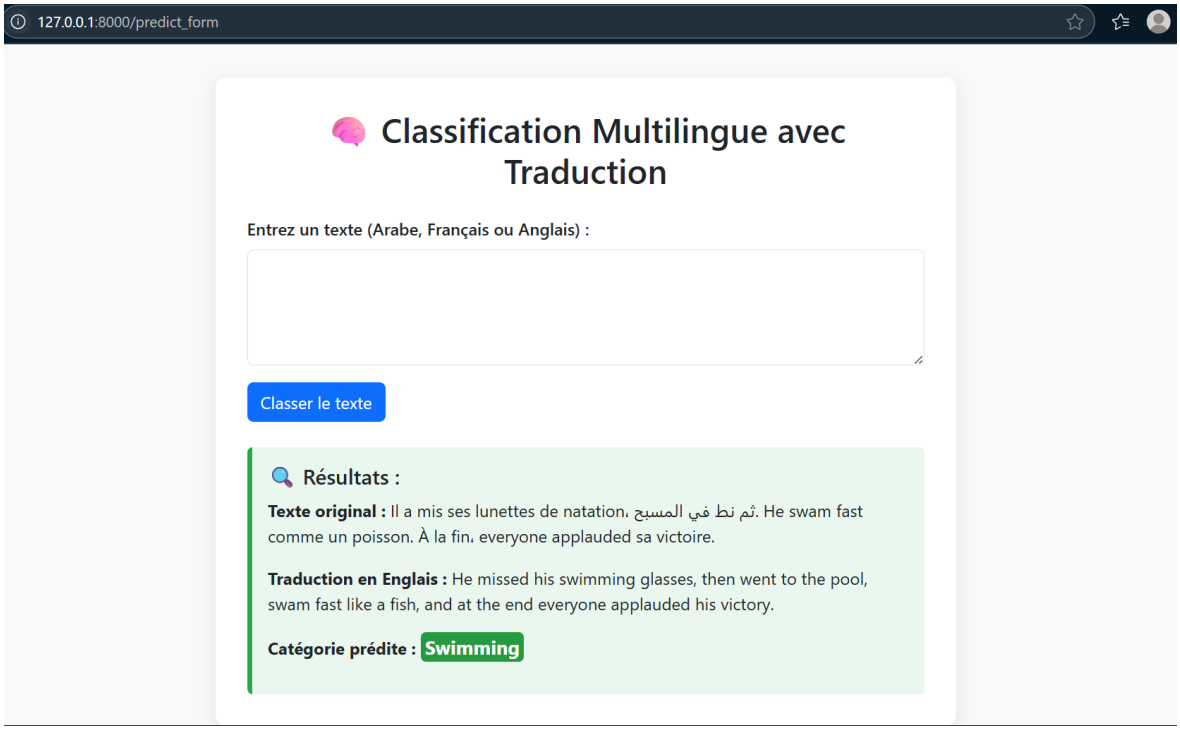

Figure 4.6: interface des résultats

Ci-dessous, nous présentons une capture d'écran illustrant un exemple d'utilisation de l'interface. Dans cet exemple, nous avons saisi un texte multilingue contenant des parties en arabe, en français et en anglais. Le système a effectué la traduction et a prédit la catégorie **"Swimming"**, ce qui démontre sa capacité à traiter des textes multilingues et à identifier le thème principal.

4.6 Conclusion

Ce chapitre a fourni l'occasion d'examiner et de comparer différents modèles de classification dans le cadre de la segmentation thématique des textes sportifs. Les données expérimentales indiquent que les méthodes fondées sur les réseaux neuronaux **(MLP)** dépassent légèrement les modèles traditionnels en matière de performance. Cependant, chaque algorithme a ses propres avantages en fonction du contexte et des ressources à disposition. Ces observations constitueront la fondation pour suggérer des axes d'amélioration dans la conclusion finale.

Conclusion Générale

À la fin de ce projet, nous avons réalisé une analyse détaillée sur la segmentation thématique des documents multilingues liés au sport, en associant des procédures de traitement automatique du langage naturel **(TAL)** à des techniques de classification supervisée. Le projet a été structuré autour de plusieurs piliers importants, de l'étude théorique des méthodes de segmentation à la réalisation technique d'un système de classification efficace.

Travaux réalisés

Les travaux réalisés dans le cadre de ce mémoire ont permis l'acquisition et le nettoyage d'un corpus multilingue d'articles sportifs provenant de la plateforme Kaggle. Nous avons mis en place un prétraitement linguistique rigoureux, incluant la tokenisation, la lemmatisation, la suppression des mots vides ainsi que la vectorisation des textes à l'aide de la méthode TF-IDF. Ensuite, trois modèles de classification supervisée ont été évalués de manière comparative : **Naive Bayes**, **Support Vector Machine (SVM)** et **Multilayer Perceptron (MLP)**. Afin d'uniformiser les contenus linguistiques, un modèle de traduction automatique **(NLLB)** a été intégré pour traduire les textes dans une langue unique. Une interface web interactive a été développée, permettant de classifier automatiquement un texte sportif en arabe, en français ou en anglais. Enfin, une phase d'optimisation du modèle **SVM** a été effectuée, notamment en filtrant certaines catégories moins représentées pour améliorer la précision du système.

Les résultats obtenus ont montré que les modèles **SVM** et **MLP** offraient une meilleure précision en matière de classification thématique, tandis que le modèle **Naive Bayes** s'est distingué par sa rapidité et sa simplicité d'exécution.

Cette complémentarité entre précision et performance computationnelle a permis d'aboutir à un système robuste et efficace, capable de traiter des textes sportifs multilingues avec fiabilité.

Perspectives

Une évolution intéressante consisterait à associer un système de résumé automatique à la segmentation thématique. Cela permettrait de classer les textes tout en en offrant une synthèse claire, facilitant ainsi la lecture rapide et l'accès à l'essentiel.

Il serait également pertinent de concevoir un système capable de gérer des textes multilingues contenant plusieurs thématiques. Cela améliorerait la capacité du modèle à analyser des articles et des textes complexes avec plus de précision et de pertinence.

En définitive, ce travail de recherche représente une contribution initiale dans le secteur de la segmentation thématique pour le sport en plusieurs langues. Il pave également la voie pour des travaux futurs d'envergure qui allient intelligence artificielle, linguistique informatisée et accessibilité linguistique à grande échelle.

Bibliography

[1] A. Widlocher, F. Bilhaut, N. Hernandez, F. Rioult, T. Charnois, S. Ferrari, and P. Enjalbert, "Une approche hybride de la segmentation thematique : collaboration du traitement automatique des langues et de la fouille de texte," *GREYC, Université de Caen, CNRS UMR 6072*, 2006.

[2] A. Benamar, "Segmentation de texte non-supervisée pour la détection de thématiques à l'aide de plongements lexicaux," in *6e conférence conjointe Journées d'Études sur la Parole (JEP, 33e édition), Traitement Automatique des Langues Naturelles (TALN, 27e édition), Rencontre des Étudiants Chercheurs en Informatique pour le Traitement Automatique des Langues (RÉCITAL, 22e édition)*, pp. 1–14, 2020.

[3] H. Neggaz, "Segmentation thématique des textes anglais," mémoire de master, Université Ain Temouchent Belhadj Bouchaib, Algérie, 2021 2021.

[4] L. Liu, "Segmentation thématique de transcriptions automatiques de données audiovisuelles," mémoire de master, Institut National des Langues et Civilisations Orientales (INALCO), France, 2022 2022. Encadrante : Camille Guinaudeau.

[5] M. Laignelet and C. Pimm, "La segmentation thématique texttiling comme indice pour le repérage de segments d'information évolutive dans un corpus de textes encyclopédiques," in *RÉCITAL 2007*, (Toulouse, France), pp. 387–396, Université Toulouse 2 – Le Mirail, juin 2007.

[6] M. A. Hearst, "Texttiling: Segmenting text into multi-paragraph subtopic passages," *Computational Linguistics*, vol. 23, no. 1, pp. 33–64, 1997.

[7] C. zhi Liu, Y. xiu Sheng, Z. qiang Wei, and Y.-Q. Yang, "Research of text classification based on improved tf-idf algorithm," in *Proceedings of the International*

Conference of Intelligent Robotic and Control Engineering (IRCE), pp. 218–222, IEEE, 2018.

[8] U. C. Calistus, M. O. Onyesolu, A. C. Doris, and C. V. Egwu, "Exploring latent dirichlet allocation (lda) in topic modeling: Theory, applications, and future directions," *Newport International Journal of Engineering and Physical Sciences*, vol. 4, pp. 9–16, March 2024.

[9] L. Sitbon and P. Bellot, "Évaluation de méthodes de segmentation thématique linéaire non supervisées après adaptation au français," *TALN 2004*, avril 2004.

[10] Wikipédia contributors, "Variation linguistique." *Wikipédia, l'encyclopédie libre*, février 2024.

[11] Olliewood, "Variantes linguistiques : Qu'est-ce que c'est ? types, caractéristiques, exemples." Olliewood.fr, février 2024.

[12] Inforsid, "Stratégies optimales pour l'analyse multidimensionnelle de contenus." Consulté sur Inforsid.fr, 2024.

[13] Nugg.ad, "Corpus en nlp et ia." Nugg.ad - Glossaire IA, février 2024.

[14] Ichi.pro, "Xlm-roberta : Apprentissage de la représentation multilingue non supervisé à grande échelle." Ichi.pro, février 2024.

[15] L. D. du Code, "Bert : Le transformer model qui s'entraîne et qui représente." Les Dieux du Code - Blog, avril 2019.

[16] S. Bird, E. Klein, and E. Loper, *Natural Language Processing with Python: Analyzing Text with the Natural Language Toolkit*. Sebastopol, CA: O'Reilly Media, Inc., 2009.

[17] D. Jurafsky and J. H. Martin, *Speech and Language Processing*. Pearson Education, draft of january 12, 2025 ed., 2025. Accessed via Google Books.

[18] B. Trouvilliez, "Représentation vectorielle de textes courts d'opinions – analyse de traitements sémantiques pour la fouille d'opinions par clustering," in *Actes de la conférence RECITAL*, (Montréal, Canada), pp. 19–23, juillet 2010.

[19] P. Bafna, D. Pramod, and A. Vaidya, "Document clustering: Tf-idf approach," in *International Conference on Electrical, Electronics, and Optimization Techniques (ICEEOT)*, pp. 61–66, IEEE, 2016.

[20] V. Claveau, "Vectorisation, okapi et calcul de similarité pour le tal: pour oublier enfin le tf-idf," in *Actes de la conférence conjointe JEP-TALN-RECITAL 2012, volume 2: TALN*, (Grenoble, France), pp. 85–98, ATALA & AFCP, 2012.

[21] Unknown, "Lecture 7: Naïve bayes classifier." `Lect-7-DM.pdf`. Accessed April 2025.

[22] G. I. Webb, "Naïve bayes," in *Encyclopedia of Machine Learning and Data Mining*, Springer, 2016.

[23] M. Hasan and F. Boris, "Svm : Machines à vecteurs de support ou séparateurs à vastes marges," Jan. 2006. Rapport de projet, BD Web.

[24] M. Hasan and F. Boris, "Svm : Machines à support de vecteurs," tech. rep., BD Web, ISTY3, Versailles St Quentin, janvier 2006.

[25] G. S. Lumacad and R. A. Namoco, "Multilayer perceptron neural network approach to classifying learning modalities under the new normal," *IEEE Transactions on Computational Social Systems*, 2023. Accepted version.

[26] S. Wu and M. Dredze, "Are all languages created equal in multilingual bert?," in *Proceedings of the 5th Workshop on Representation Learning for NLP (RepL4NLP)*, 2020. arXiv preprint arXiv:2005.09093.

[27] S. Wu and M. Dredze, "Are all languages created equal in multilingual bert?," in *Proceedings of the 5th Workshop on Representation Learning for NLP (RepL4NLP-2020)*, pp. 120–130, Association for Computational Linguistics, 2020.

[28] NLLB Team, M. R. Costa-jussà, J. Cross, O. Çelebi, M. Elbayad, K. Heafield, K. Heffernan, E. Kalbassi, J. Lam, D. Licht, J. Maillard, A. Sun, S. Wang, G. Wenzek, A. Youngblood, B. Akula, L. Barrault, G. M. Gonzalez, P. Hansanti, J. Hoffman, S. Jarrett, K. R. Sadagopan, D. Rowe, S. Spruit, C. Tran, P. Andrews, N. F. Ayan, S. Bhosale, S. Edunov, A. Fan, C. Gao, V. Goswami, F. Guzmán, P. Koehn,

A. Mourachko, C. Ropers, S. Saleem, H. Schwenk, and J. Wang, "No language left behind: Scaling human-centered machine translation," tech. rep., Meta AI, 2022. Available at `https://github.com/facebookresearch/fairseq/tree/nllb`.

[29] G. Van Rossum and F. L. Drake, *Python 3 Reference Manual*. CreateSpace, 2009.

[30] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer, 2nd ed., 2009.

[31] Microsoft, "Visual studio code." `https://code.visualstudio.com/`, 2024. Accessed: 2025-04-15.

[32] P. Jupyter, "Jupyter documentation." `https://jupyter.org/`, 2025. Consulté en 2025.

[33] DataScientest, "Pandas : la bibliothèque Python dédiée à la Data Science," Apr. 2024. Consulté le 22 avril 2024.

[34] NumPy Community, "NumPy –," Apr. 2024. Consulté le 8 avril 2024.

[35] "Matplotlib — visualization with python." `https://matplotlib.org/`, 2024. Consulté le: 22 avril 2024.

[36] Zygomatic, "Générateur de nuage de mots clés gratuit en ligne et Générateur de nuage de tags," Apr. 2024. Consulté le 8 avril 2024.

[37] M. L. Waskom, *Seaborn: statistical data visualization*, 2025. `https://seaborn.pydata.org/`.

[38] P. Sharma, "How to save and load machine learning models in python using joblib library," *Analytics Vidhya*, 2025.

Annexe

Dans cette annexe, nous fournissons des extraits de code pertinents utilisés dans le cadre de notre étude.

Classification

Prétraitement des données

Cette partie du code supprime les lignes contenant des valeurs manquantes ou des doublons dans le dataframe. Ensuite, elle conserve uniquement les colonnes importantes « Headline » et « Sport » et s'assure que ces colonnes ne contiennent pas de valeurs vides.

```
# Supprimer les lignes contenant des valeurs manquantes
df = df.dropna()

# Supprimer les lignes dupliqu es
df = df.drop_duplicates()
# Conserver uniquement les colonnes importantes : 'Headline' et 'Sport
    '
df = df[['Headline', 'Sport']]
# V rifier que les titres ne sont pas vides
df = df[df['Headline'].str.strip() != '']
# V rifier que les cat gories ne sont pas vides
df = df[df['Sport'].str.strip() != '']
```

Listing 4.1: Nettoyage du dataset

Cette fonction nettoie un texte en supprimant les liens, les caractères non alphabétiques et les mots vides.

```
def nettoyer_texte(text):
    # Convertir le texte en minuscules
    text = text.lower()

    # Supprimer les liens web
    text = re.sub(r"http\S+|www\S+|https\S+", '', text, flags=re.
        MULTILINE)

    # Supprimer les caract res non alphab tiques
    text = re.sub(r"[^a-z\s]", '', text)

    # Supprimer les mots vides (stopwords)
    words = text.split()
    words = [word for word in words if word not in stop_words]

    # Appliquer la lemmatisation
    words = [lemmatizer.lemmatize(word) for word in words]

    # Reconstituer le texte nettoy
    return " ".join(words)
```

Listing 4.2: Fonction de nettoyage de texte avec NLTK

Entrainement avec SVM

Ce code entraîne un modèle SVM linéaire pour la classification, puis évalue ses performances à l'aide d'indicateurs comme la précision, le rapport de classification et la matrice de confusion.

```
# Cr ation et entra nement du mod le SVM lin aire
model_svm = SVC(kernel='linear', random_state=42)
model_svm.fit(X_train, y_train)

\sharp Pr diction sur les donn es de test
y_pred_svm = model_svm.predict(X_test)

# Affichage de la pr cision et du rapport de classification
print(f"Accuracy: {accuracy_score(y_test, y_pred_svm) * 100:.2f}%")
print(classification_report(y_test, y_pred_svm))

# Affichage de la matrice de confusion
plt.figure(figsize=(10,8))
sns.heatmap(confusion_matrix(y_test, y_pred_svm), annot=True, fmt='d',
            xticklabels=model_svm.classes_, yticklabels=model_svm.
                classes_, cmap='Blues')
plt.xlabel("Pr dictions")
plt.ylabel("R el")
plt.title("Matrice de confusion - SVM")
plt.show()
```

Listing 4.3: Entrainement avec SVM et affichage de la matrice de confusion

Le serveur et le modèle de traduction NLLB

Ce code charge le modèle de traduction NLLB et configure les langues supportées. Nous avons mis en place un serveur FastAPI pour combiner ce système de traduction avec notre modèle SVM de classification.

```
# --- Chargement du mod le de traduction NLLB ---
tokenizer_translation = AutoTokenizer.from_pretrained("facebook/nllb
    -200-distilled-600M")
model_translation = AutoModelForSeq2SeqLM.from_pretrained("facebook/
    nllb-200-distilled-600M")

# --- Mappage des codes linguistiques NLLB ---
lang_code_to_id = {
    "ar_AR": "arb_Arab",
    "fr_XX": "fra_Latn",
    "en_XX": "eng_Latn"
}

# --- Configuration de FastAPI ---
app = FastAPI()
templates = Jinja2Templates(directory="templates")

# --- Fonction de traduction utilisant NLLB ---
def translate_text(text: str, source_lang: str, target_lang: str) ->
    str:
    src = lang_code_to_id[source_lang]
    tgt = lang_code_to_id[target_lang]
```

Listing 4.4: Chargement du modèle NLLB et configuration du serveur FastAPI