الجمهورية الجزائرية الديمقراطية الشعبية
République Algérienne Démocratique et Populaire
وزارة التعليم العالي والبحث العلمي
Ministère de l'Enseignement Supérieur et de la Recherche Scientifique
جامعة قاصدي مرباح ورقلة
Université Kasdi Merbeh Ouargla
كلية التكنولوجيات الحديثة للمعلومات والاتصال
Faculté Des Nouvelles Technologies de l'Information Et de la Communication
قسم الاعلام الآلي وتكنولوجيا المعلومات
Département d'Informatique et des Technologies de l'Information

Mémoire de Master Académique

Pour l'obtention du diplôme de Master

Domaine : Mathématique et Informatique
Filière : Informatique
Spécialité : Informatique Industrielle

Mémoire présenté par :

Babkeur Firdaous
Chemlal Djana

Sujet :

Analyse des tendances dans les réseaux sociaux

Discussion publique : 23/06/2024
Membre du jury :

Dr. Bouhyaoui Nasria	Président(e)	UKM Ouargla
Pr. Abderrahim Mohammed El Amine	Superviseur	UKM Ouargla
Dr. Toumi Chahrazed	Examinateur(rice)	UKM Ouargla

Année universitaire : 2023 /2024

الملخص

لقد أدى الانتشار السريع لشبكات التواصل الاجتماعي إلى تأثير متزايد على مختلف جوانب الحياة، بما في ذلك التعليم الجامعي. ومع ازدياد استخدام الطلاب الجامعيين الجزائريين لهذه المنصات للحصول على الأخبار والمعلومات حول الدراسة والحياة الجامعية، أصبح من الضروري فهم اتجاهاتهم وآرائهم بطريقة علمية ومنهجية.

يأتي مشروع البحث هذا ضمن مجال المعالجة الآلية للغة الطبيعية والذكاء الاصطناعي، لقد استخدمنا خوارزميات التعلم الآلي المتقدم، وبشكل خاص التعلم الآلي العميق، لتحليل البيانات التي تم جمعها من شبكات التواصل الاجتماعي حول النصوص المتعلقة بالموضوعات الجارية التي تهم الطلاب الجامعيين. لأجل عملية التصنيف، لقد استخدمنا نموذج اللغة BERT.

من نتائج هذا المشروع تطوير برنامج يعتبر كأداة فعالة وقوية لتحليل الاتجاهات والآراء حول الموضوعات التي يناقشها الطلاب على وسائل التواصل الاجتماعي، مما سيساهم حتما في تعزيز التواصل بين الجامعات والطلاب.

الكلمات المفتاحية: تحليل التوجهات، الشبكات الاجتماعية، التعلم الآلي، نموذج اللغة (LLM)، معالجة اللغة الطبيعية، BERT.

Abstract

The rapid proliferation of social networks has led to a growing impact on various aspects of life, including university education. With the increasing use of these platforms by Algerian university students to obtain news and information about study and university life, it has become necessary to understand their trends and opinions in a scientific and methodological manner. This research project falls within the field of Natural Language Processing (NLP) and Artificial Intelligence (AI). We have used advanced machine learning approaches, and more particularly deep learning, for the analysis of data collected from social networks concerning texts related to current topics of interest to university students. For classification, we used the BERT language model.

The outcome of this project includes an effective and powerful tool for analyzing trends and opinions on the topics discussed by students on social networks, thus contributing to strengthening communication between universities and students.

Keywords: Trend Analysis, Social Networks, Machine Learning, LLM, Natural Language Processing (NLP), BERT.

Résumé

La propagation rapide des réseaux sociaux connaît une croissance croissante de son impact sur divers aspects de la vie, y compris l'enseignement universitaire. Avec l'utilisation croissante de ces plates-formes par les étudiants universitaires algériens pour obtenir des nouvelles et des informations sur l'étude et la vie universitaire, il est devenu nécessaire de comprendre leurs tendances et leurs opinions de manière scientifique et méthodologique. Le présent projet de recherche s'insère dans le domaine du traitement automatique de la langue (NLP) et de l'intelligence artificielle (AI). Nous avons utilisé des approches avancées de l'apprentissage automatique et plus particulièrement de l'apprentissage profond pour l'analyse des données recueillies à partir des réseaux sociaux concernant des textes relatives aux sujets courants qui intéressent les étudiants universitaires.
Pour la classification, nous avons utilisé le modèle de langue BERT.
Le résultat de ce projet comprend un outil efficace et puissant pour analyser les tendances et des opinions sur les sujets abordés par les étudiants sur les réseaux sociaux, contribuant ainsi à renforcer la communication entre les universités et les étudiants.

Mots clés : analyse des tendances, réseaux sociaux, Apprentissage automatique, Modèle de langue (LLM), Traitement automatique de la langue naturelle (NLP), BERT.

Dédicaces

“

A mos très chers parents

”

- Firdaous et Djana

REMERCIEMENTS

Tout d’abord, nous rendons grâce à DIEU Tout-Puissant, qui nous a donné la force, la patience et l’audace pour surmonter toutes les difficultés.

Nous remercions également nos parents, qui nous ont encouragé et motivés à atteindre ce niveau d’étude.

Nous remercions sincèrement notre encadreur Pr. Abderrahim Mohammed El Amine pour ses efforts continus pour nous soutenir, guider et motiver a résoudre les problèmes de recherche.

Nous adressons également nos sincères remerciements aux membres du jury....., ...qu’on participe à examiner notre travail.

Sans oublier nos professeurs pour leurs conseils au cours des cinq dernières années au Département de technologie de l’information à l’Université de Kasdi Merbeh.

Un grand merci aux membres de notre famille et à nos amis pour leur soutien.

Enfin, nous tenons à remercier tous ceux qui nous ont aidés de près ou de loin pendant nos années universitaires et à la préparation de cette thèse.

Tout le respect et la gratitude

Table des matières

Introduction Générale……2

Chapitre 01: Concepts généraux 4

1.1 Introduction 4

1.2 Réseaux sociaux 4

1.2.1 Types des réseaux sociaux 5

1.3 L'intelligence artificielle 6

1.4 Apprentissage Automatique (Machine Learning) 6

1.4.1 Types d'apprentissage automatique 6

1.4.2 Cas d'utilisation et applications de l'apprentissage automatique 9

1.5 Réseaux de neurones 9

1.6 L'apprentissage profond (Deep Learning) 9

1.7 Traitement Automatique de la Langue Naturelle 10

1.7.1 Utilisations de NLP 10

1.7.2 Avantages du NLP 11

1.7.3 Modèles de l'apprentissage automatique en NLP 11

1.8 Grand modèle de langage (Large Language Model) 12

1.8.1 Utilisation des LLM 12

1.8.2 Fonctionnement des LLM 12

1.8.3 Avantages des LLM 13

1.9 Topic modeling 13

1.10 BERT 13

1.11 Bertopic 14

1.12 Conclusion 14

Chapitre 02: Analyse des tendances : Etat de l'art 16

2.1 Introduction 16

2.2 L'analyse de tendance 16

2.2.1 Relation entre NLP et analyse de tendance 16

2.2.2 Relation entre LLM et Analyse de tendance 16

2.2.3 Méthodologie de l'analyse des tendances 17

2.2.4 Utilisation de l'analyse des tendances 17

2.3 Travaux reliés 17

2.4 Récapitulatif des travaux réalisés 19

2.5 Conclusion 20

Table des matières

Chapitre 03: Analyse des tendances : Expérimentation 21
3.1 Introduction 21
3.2 Banque de données (Dataset) 21
3.3 Les outils utilisés 21
➢ Visual Studio Code 21
➢ Python 21
3.4 Bibliothèques utilisées dans le code 22
➢ RegEx 22
➢ NLTK 22
➢ Word tokenize 22
➢ PyArabic 22
➢ Sentence Transformer 22
➢ Approximation et projection d'une variété uniforme (UMAP) 22
➢ Regroupement spatial hiérarchique basé sur la densité des applications avec bruit (HDBSCAN) 23
➢ Scikit-learn 23
➢ Arabic stop words 23
➢ Bertopic 23
➢ Pandas 23
➢ Matplotlib 23
➢ Seaborn 24
➢ Pathlib 24
3.5 Configuration expérimentale 24
3.6 Expérimentation (Implémentation) 24
3.6.1 Prétraitement 24
3.6.2 Intégration des données (Data Embedding) 25
3.6.3 Approximation et projection d'une variété uniforme (UMAP) 25
3.6.4 Regroupement spatial hiérarchique basé sur la densité des applications avec bruit (HDBSCAN) 25
3.6.5 Vectorisation (Vectorizer) 26
3.6.6 KeyBERT Inspired 26
3.6.7 BERTopic 26
3.7 Interface de l'application 27
3.7.1 Editeur utilisé pour développer l'interface 27
➢ Qt Creator 27
3.7.2 Description de l'interface 27

Table des matières

3.7.3 Exemples d'analyse des tendances 28
3.8 Conclusion 29
Conclusion Générale ……...……... 30
Références Bibliographique ……31

Liste des figures

Figure 1 : Exemples de réseaux sociaux 4
Figure 2 : Le cycle de vie d'un modèle d'apprentissage automatique 6
Figure 3: l'apprentissage supervisé 7
Figure 4: L'apprentissage non supervisé 8
Figure 5: L'apprentissage par renforcement 8
Figure 6 : Architecture d'un réseau neuronal multicouche 9
Figure 7: Relation entre l'IA, le ML, le DL et NLP 10
Figure 8: Les applications du traitement du langage naturel 11
Figure 9 : Modélisation par sujets pour l'analyse des données 13
Figure 10: Interface utilisateur 28
Figure 11 : Exemple d'analyse des tendances 29

Liste des tableaux

Tableau 1: Types des réseaux sociaux 5
Tableau 2: Travaux réalisés sur la détection des tendances 19

Liste des abréviations

IA : Intelligence Artificielle

ML: Machine Learning

DL: Deep Learning

NLP: Natural Language Processing

LLM: Large Language Model

UTF-8 : Format de transformation Unicode - 8 bits

API: Application Programming Interface

NBM: Naïve Bayes Multinomial

NLTK : Natural Language Toolkit

MBSCI : Modèle Basé sur un Système de Capture d'Influence

RE: RegEx

CSV: Comma-Separated Values

UMAP: Uniform Manifold Approximation and Projection

HDBSCAN: Hierarchical Density-Based Spatial Clustering of Applications with Noise

VS Code : Visual Studio Code

INTRODUCTION GENERALE

Dans le contexte du développement rapide de la technologie de l'automatisation et l'utilisation généralisée des réseaux sociaux, il est devenu important d'étudier l'impact de ces réseaux sur le comportement des utilisateurs et leurs directives. Nous constatons que parmi les catégories utilisées pour les sites de communication est celle des étudiants ayant l'intention de connaître les actualités et les nouvelles liées à l'université.

Il est de plus en plus important de comprendre les préférences et les intérêts des étudiants sur les plateformes de médias sociaux chez les responsables, car ces plateformes sont considérées comme une source majeure d'information et de communication entre les étudiants. En analysant les orientations des étudiants sur ces réseaux, leurs intérêts et leurs orientations peuvent être mieux compris.

Notre projet de fin d'études vise à développer un programme qui permet de détecter les tendances dans les réseaux sociaux concernant les étudiants universitaires. Pour se faire nous allons utiliser l'apprentissage automatique (ML) et le traitement de la langue naturelle (NLP) pour créer un modèle de langue capable d'analyser et de déduire ces tendances. Notre modèle doit répondre efficacement aux besoins des responsables (administration des instituts ou des universités).

Ce mémoire est structuré en trois chapitres,
Dans le premier chapitre nous abordons les concepts fondamentaux liés à notre sujet, y compris les réseaux sociaux, l'apprentissage (ML, DL), le traitement des langues naturelles (NLP) et tout ce qui y est lié.
Dans le deuxième chapitre, nous présentons la notion de l'analyse des tendances et ses travaux reliés.
Dans le troisième chapitre, nous décrivons la partie conception et réalisation de notre analyseur de tendance.

Chapitre I :
Concepts généraux

Chapitre 01: Concepts généraux

1.1 Introduction

Dans ce chapitre, nous discutons de nombreux concepts de base liés à l'analyse des tendances dans les médias sociaux comme : l'apprentissage artificielle, le traitement du langage naturel (NLP) et les grands modèles de langage (LLM).

1.2 Réseaux sociaux

L'expression « réseau social » est employée la première fois en 1954 par le sociologue australien John Arundel Barnes pour caractériser l'ensemble des relations entre les différentes personnes d'un groupe (familial, amical ou professionnel).

→ Les réseaux sociaux désignent des plateformes qui permettent aux masses de se connecter entre eux (voir Figure 1 pour des exemples de réseaux sociaux). Ils ont révolutionné la manière dont nous communiquons, nous informons et nous organisons [1].

Figure 1 : Exemples de réseaux sociaux [2]

1.2.1 Types des réseaux sociaux

Les réseaux sociaux peuvent être classés comme le montre le tableau ci-dessous (voir tableau 1) [3]:

Tableau 1: Types des réseaux sociaux

Type	Exemple	Description
Généralistes	Facebook	Permet irréguliers bonshommes d'échanges verso sa tribu d'amis.
	Twitter	Permets d`stipendier des messages derrière d'disparates internautes derrière une borne de caractères.
	Myspace	Site interactif qui offre à ses abonnés de multiples services combinant blog, espace personnel, espace communautaire.
Professionnels	LinkedIn	Un lacis accoucheur international qui permet de défi en fait à cause professionnelle.
	Viadeo	Permet d'édifier et de préconiser son réseau professionnel
Médias de partages	Instagram	Permet de dispatcher, d`affecter des photos et vidéos brève envers sa conversion d'amis uniquement item de dispatcher des stories d'une légitimité de 24 heures
	Dailymotion	Espace où on peut télécharger, partager et regarder des vidéos.
	YouTube	Grand espace de disposition de vidéos.

1.3 L'intelligence artificielle

L'intelligence artificielle est un processus d'imitation de l'intelligence humaine qui repose sur la création et l'application d'algorithmes exécutés dans un environnement informatique dynamique. Son but est de permettre à des ordinateurs de penser et d'agir comme des êtres humains.

Pour y parvenir, trois composants sont nécessaires :

- Des systèmes informatiques
- Des données avec des systèmes de gestion
- Des algorithmes d'IA avancés (code)

Pour se rapprocher le plus possible du comportement humain, l'intelligence artificielle a besoin d'une quantité de données et d'une capacité de traitement élevées [4].

1.4 Apprentissage Automatique (Machine Learning)

L'apprentissage automatique peut être défini comme étant une technologie d'intelligence artificielle permettant aux machines d'apprendre sans avoir été au préalablement programmées spécifiquement à cet effet (voir Figure 2). L'apprentissage automatique est explicitement lié au Big Data, étant donné que pour apprendre et se développer, les ordinateurs ont besoin de flux de données à analyser, sur lesquelles s'entraîner [5].

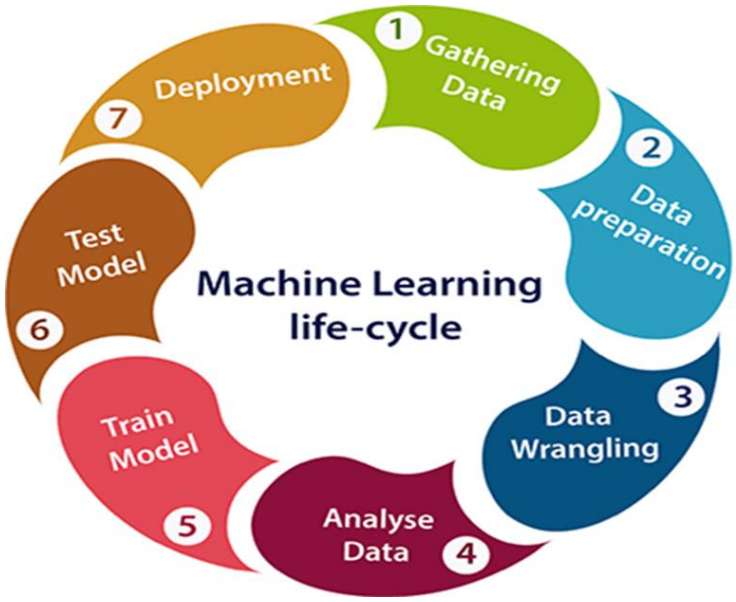

Figure 2 : Le cycle de vie d'un modèle d'apprentissage automatique [6]

1.4.1 Types d'apprentissage automatique

On peut citer trois types d'algorithme d'apprentissage automatique :

- Apprentissage supervisé.
- Apprentissage non supervisé.
- Apprentissage par renforcement.

1.4.1.1 Apprentissage supervisé

L'apprentissage supervisé est la tâche d'apprentissage automatique la plus simple et la plus connue, (fait en utilisant une vérité), Il est basé sur un certain nombre d'exemples pré classifiés, dans lesquels est connu à priori la catégorie à laquelle appartient chacune des entrées utilisées comme exemples (voir Figure 3).

Par conséquent, le but de ce type d'apprentissage est d'apprendre une fonction qui, compte tenu d'un échantillon de données et de résultats souhaités, se rapproche le mieux de la relation entre les entrées et les sorties observables dans les données.

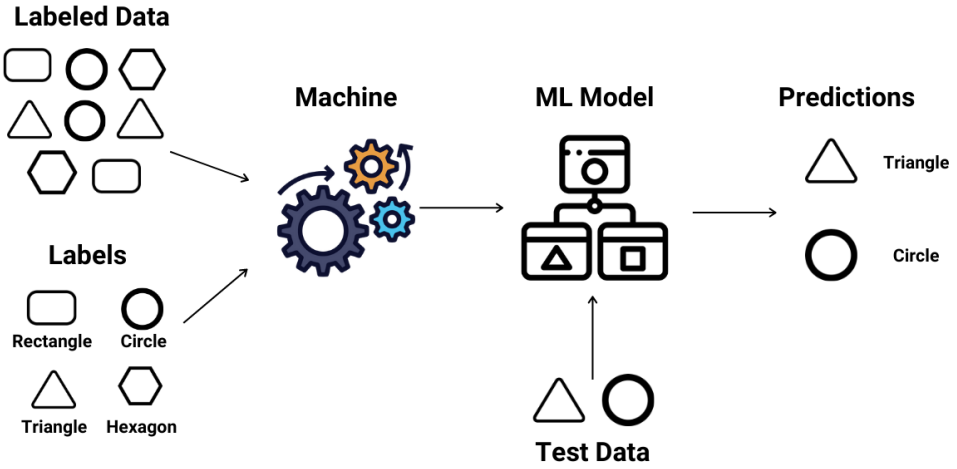

Figure 3: l'apprentissage supervisé *[7]*

Dans l'apprentissage supervisé, on a deux types d'algorithmes :

- Les algorithmes de **régression**, qui cherchent à prédire une valeur continue, une quantité.
- Les algorithmes de **classification**, qui cherchent à prédire une classe/catégorie [10].

Apprentissage non supervisé

La deuxième classe d'algorithmes d'apprentissage automatique est appelée apprentissage non supervisé, dans ce cas, nous n'étiquetons pas les données au préalable, nous laissons plutôt l'algorithme arriver à sa conclusion. Le modèle en question étudie ses données d'entrainement dans le but de déduire une fonction pour décrire une structure cachée à partir ces données (voir Figure 4). À aucun moment le système ne connaît la sortie correcte avec certitude. Au lieu de cela, il tire des inférences des ensembles de données quant à ce que la sortie devrait être. Une approche standard consiste à définir une mesure de similarité entre deux objets, puis à rechercher tout groupe d'objets plus similaires les uns aux autres, par rapport aux objets des autres clusters [8].

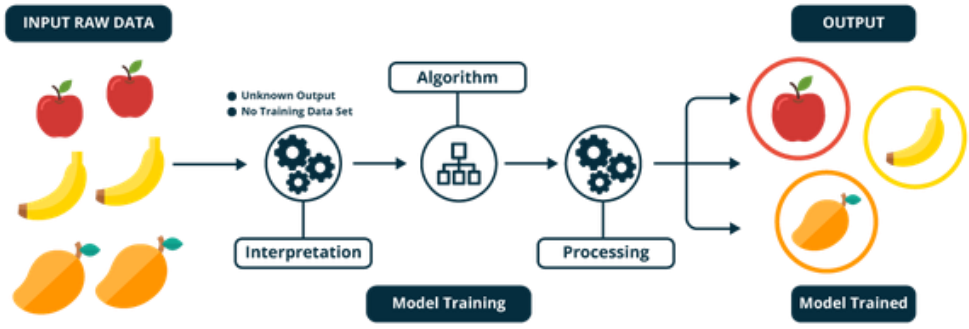

Figure 4: L'apprentissage non supervisé [7]

On peut utiliser les algorithmes de ce genre d'apprentissage pour résoudre trois types de problèmes :

- Association : Un souci dans lequel on souhaite trouver des règles qui définissent de vastes parties de ses données.
- Regroupement : Un souci où il est nécessaire de trouver les ensembles inhérents aux données.
- La réduction de dimension : L'objectif est de diminuer le nombre de variables à intégrer dans l'analyse.

1.4.1.2 Apprentissage par Renforcement

L'apprentissage par renforcement se concentre sur l'acquisition de connaissances par le système à travers ses interactions avec son environnement (voir Figure 5). En utilisant la méthode de renforcement de l'apprentissage, le système ajuste ses paramètres en fonction des réactions de l'environnement, ce qui donne ensuite un retour d'information sur les décisions prises. [8]

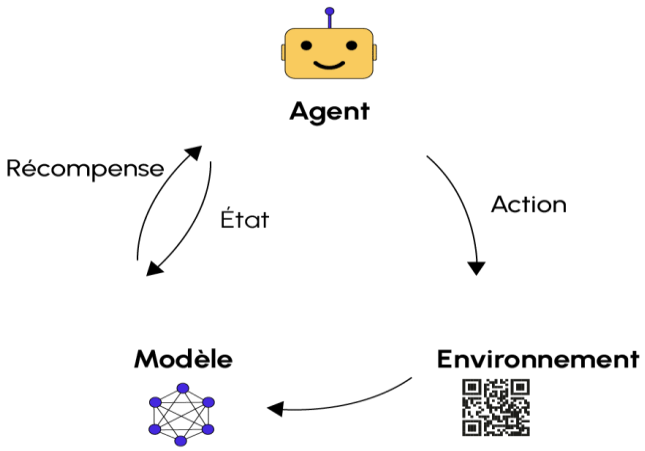

Figure 5: L'apprentissage par renforcement *[9]*

1.4.2 Cas d'utilisation et applications de l'apprentissage automatique

Nous trouvons plusieurs applications de l'apprentissage automatique :

- Systèmes de recommandation tels que Netflix, YouTube et Spotify.
- Moteurs de recherche : tels que Google et Baidu.
- Algorithmes d'information sur les réseaux sociaux tels que Facebook et Twitter.
- Applications pratiques : Aspirateurs robotisés, détection de spam (dans les services de messagerie), analyse d'images médicales (aider les médecins à détecter les tumeurs), voitures autonomes.
- Assistants numériques : tels que Siri, Alexa et Google Assistant.
- Systèmes de reconnaissance vocale et de synthèse vocale [12].

1.5 Réseaux de neurones

Les réseaux neuronaux se caractérisent par la capacité d'apprendre des données et de modifier les résultats en fonction de cet apprentissage (voir Figure 6) [13].

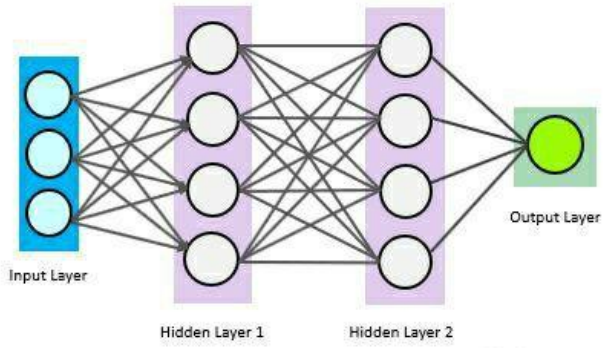

Figure 6 : Architecture d'un réseau neuronal multicouche [10]

1.6 L'apprentissage profond (Deep Learning)

L'apprentissage profond est un sous-domaine d'apprentissage automatique consistant à créer des systèmes capables d'apprendre, prévoir et décider en toute autonomie. Cette forme d'intelligence artificielle fonctionne avec des algorithmes capables d'imiter le cerveau humain grâce à un large réseau de neurones artificiels [15].

Exemples d'application de l'apprentissage profond

L'apprentissage profond est employé dans différents domaines, allant de la conduite automatisée aux dispositifs médicaux.

Grâce à l'apprentissage profond [16],

- Il est possible d'ajouter des sons à des films qui restent silencieux.
- Effectuer une traduction automatisée.
- Classer les objets en utilisant des photographies.
- Produire des écrits automatiques.
- Création d'une légende visuelle.
- Le jeu automatique.

1.7 Traitement Automatique de la Langue Naturelle

"Le Traitement Automatique des Langues ou Natural Language Processing (NLP)، est une branche de l'AI qui s'attache à donner la capacité aux machines de comprendre, générer ou traduire le langage humain tel qu'il est écrit et/ou parlé. Les chatbots figurent parmi les logiciels de NLP les plus populaires, par exemple, les assistants vocaux Alexa, Google Home ou encore Siri" [11].

La figure suivante (voir Figure 8) présente la relation entre NLP et IA.

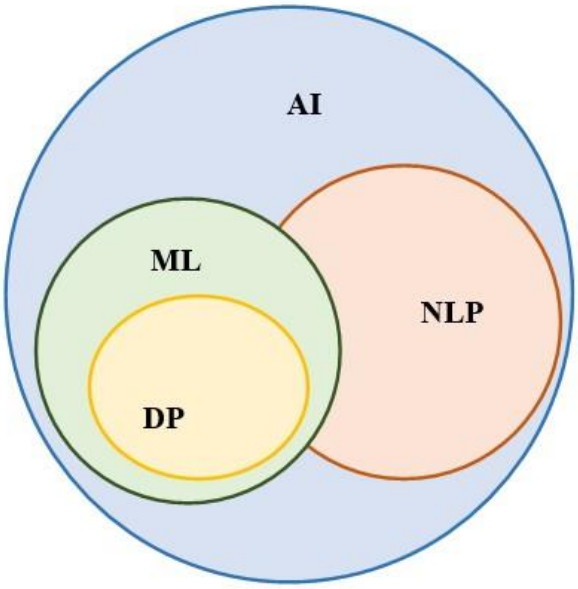

Figure 7: Relation entre l'IA, le ML, le DL et NLP *[12]*

1.7.1 Utilisations de NLP

" Le NLP a pour but de doter les logiciels de processus de traitement automatique du langage vocal ou textuel. Partant de là, il recouvre de nombreux cas d'usage plus ou moins élaboré : [11]

- La classification de texte
- La reconnaissance de texte
- Le résumé automatique
- La traduction automatique
- Les chatbot, voicebot ou callbot"

La figure suivante (voir Figure 7) résume l'utilisation du NLP.

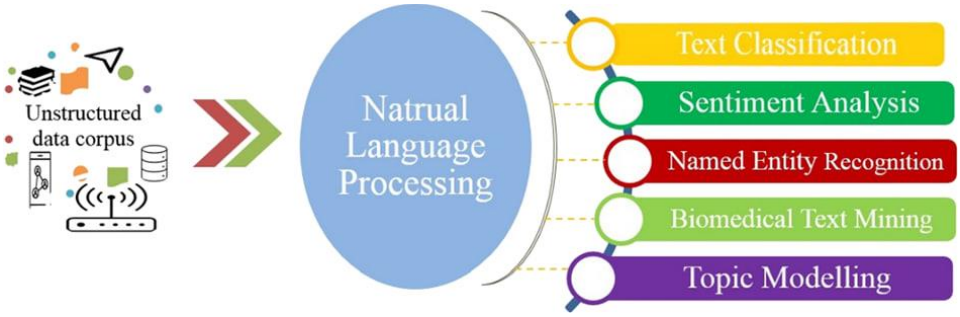

Figure 8: Les applications du traitement du langage naturel *[13]*

1.7.2 Avantages du NLP

Le NLP offre de nombreux bénéfices dans différents secteurs. Voici certains des principaux bénéfices :

- Automatisation des tâches
- Amélioration de l'expérience utilisateur
- Analyse de sentiment et compréhension des opinions
- Gestion de l'information et extraction de connaissances
- Amélioration de la recherche et du développement
- Support multilingue.

1.7.3 Modèles de l'apprentissage automatique en NLP

Globalement, le traitement du langage naturel (NLP) se décline en deux grandes catégories de modèles de l'apprentissage automatique :

- Les modèles de l'apprentissage automatique orientés NLU (natural language understanding) qui s'attachent à saisir le sens d'une langue et d'un discours dans son contexte.
- Les modèles de l'apprentissage automatique orientés NLG (natural language generation) qui ont pour but de générer un texte à la manière d'un humain [11].

1.8 Grand modèle de langage (Large Language Model)

Un modèle de langue (ou LLM), est une forme d'intelligence artificielle conçue pour comprendre et générer du langage humain. Ces modèles sont formés sur d'énormes ensembles de données textuelles (d'où "Large"), leur permettant de répondre aux requêtes, de rédiger des textes, et même d'interagir de manière conversationnelle. Grâce à des architectures comme les Transformers, les LLMs peuvent traiter des informations avec une précision et une nuance sans précédent [14].

1.8.1 Utilisation des LLM

Les LLM peuvent être utilisés pour une multitude de tâches. Par exemple : [15]

- Les questions-réponses
- L'analyse des sentiments
- L'extraction d'informations
- La capture d'images
- La reconnaissance d'objet
- La génération de texte
- Le résumé de texte
- La création de contenu
- Les chatbots, les assistants virtuels et les IA conversationnelles (c'est typiquement le cas du logiciel open source ChatGPT)
- La traduction
- La détection de fraude

Du fait de leurs multiples fonctionnalités, les LLM s'adaptent parfaitement à tous les secteurs d'activité (bancaire, logistique, santé, industrie…).

1.8.2 Fonctionnement des LLM

L'objectif des LLM étant d'apprendre la complexité du langage humain, ils sont pré-entraînés sur une grande quantité de données (comme du texte, des images, des vidéos, des discours, des données structurées…). Plus un LLM utilise de paramètres, meilleures sont ces performances. À ce titre, les grands modèles linguistiques nécessitent donc des ressources importantes en termes de données, de calcul et d'ingénierie.

En particulier, lors de la phase de pré-entraînement. À ce stade, les LLM doivent apprendre les tâches et fonctions linguistiques de base. Dès lors que le modèle d'apprentissage est pré-entraîné, il peut être entraîné avec de nouvelles données spécifiques. L'objectif est d'affiner ses capacités pour des cas d'utilisation particuliers. On parle alors de méthode fine tuning. Cette phase de l'apprentissage nécessite moins de données et d'énergie.

1.8.3 Avantages des LLM

Les avantages des LLM pour les organisations sont nombreux :

- **Automatisation des processus :** Les LLM peuvent automatiser divers processus tels que le service client, la génération de texte, les prédictions et la classification, ce qui réduit le temps de travail manuel et les coûts associés.
- **Favorisation de la personnalisation :** Les LLM permettent de fournir un service client disponible 24h/24 et 7j/7 grâce à des chatbots et des assistants virtuels. Ces systèmes peuvent traiter de grandes quantités de données pour personnaliser les interactions avec les clients, ce qui augmente la satisfaction client.
- **Augmentation de la précision des tâches :** En traitant de grandes quantités de données, les LLM améliorent la précision des tâches de prédiction et de classification, telles que l'analyse de sentiments à partir d'avis de clients.

1.9 Topic modeling

Est une technique en NLP utilisée pour identifier les thèmes cachés dans un grand corpus de textes (voir Figure 9). Elle permet d'extraire automatiquement des sujets potentiels à partir des mots utilisés et de leur distribution dans les documents [16].

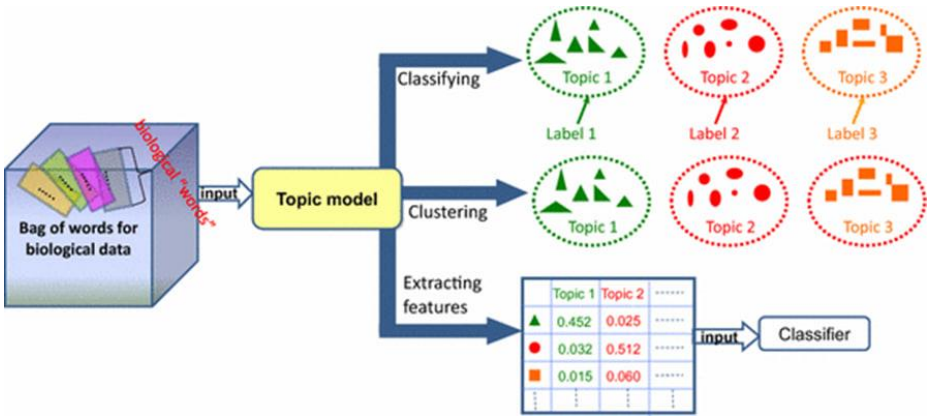

Figure 9 : Modélisation par sujets pour l'analyse des données [17]

1.10 BERT

BERT signifie « Bidirectionnel Transformer Encoded Representations » et fait référence à un algorithme basé sur NLP et les réseaux de neurones qui représente une avancée majeure dans le domaine ML. BERT se caractérise par la capacité de comprendre les textes de manière bidirectionnelle, ce qui lui permet de comprendre le contexte environnant de chaque mot du texte [18].

1.11 Bertopic

BERTopic est un modèle de modélisation de sujets qui exploite les intégrations BERT et c-TF-IDF pour créer des collections denses qui permettent d'analyser des textes et d'en extraire des sujets clés, ce qui aide à comprendre et à classer le contenu avec plus de précision et d'efficacité [19].

1.12 Conclusion

En conclusion, Les réseaux sociaux, l'analyse des tendances et le traitement du langage naturel sont des outils puissants qui peuvent être utilisés pour découvrir tout ce qui est nouveau dans de grandes quantités de données publiées par la société. En les utilisant ensemble, nous pouvons obtenir des informations précieuses sur le comportement humain, les tendances sociales et l'opinion publique.

Chapitre II
Analyse des tendances : Etat de l’art

Chapitre 02: Analyse des tendances : Etat de l'art

2.1 Introduction

Ce chapitre décrit quelques travaux effectués dans le domaine l'analyse des tendances. Il vise à résumer les principales conclusions et à identifier les lacunes de la recherche.

2.2 L'analyse de tendance

L'analyse des tendances est définie comme une technique statistique et analytique utilisée pour évaluer et identifier des modèles, des tendances ou des changements dans les données au fil du temps. Il s'agit d'examiner des données historiques afin d'obtenir des informations sur la direction ou les tendances d'un phénomène particulier. L'analyse des tendances est appliquée dans divers domaines, afin de prendre des décisions éclairées et de faire des prédictions sur la base de performances ou de comportements antérieurs. [20]

2.2.1 Relation entre NLP et analyse de tendance

Lorsqu'il s'agit d'analyser les tendances sur les réseaux sociaux, le NLP joue un rôle crucial en comprenant et en analysant les données linguistiques échangées sur ces plateformes. Par exemple, les techniques de NLP peuvent être utilisées pour extraire les sujets principaux dont les utilisateurs discutent, pour identifier l'évolution de l'intérêt pour ces sujets au fil du temps, et pour détecter les sentiments, les tendances et les opinions exprimées dans ces discussions.

En utilisant les techniques de NLP, les données textuelles présentes sur les réseaux sociaux peuvent être transformées en données analytiques, permettant ainsi aux décideurs et aux analystes de comprendre les tendances actuelles, de prédire les changements futurs et de prendre des décisions éclairées en conséquence. Ainsi, la relation entre le NLP et l'analyse des tendances réside dans l'utilisation des techniques de NLP pour extraire des informations précieuses des données linguistiques sur les réseaux sociaux, afin de comprendre les comportements, les tendances et de guider les décisions en conséquence.

2.2.2 Relation entre LLM et Analyse de tendance

La relation entre les LLM et l'analyse des tendances réside dans l'utilisation des LLM pour analyser de grands ensembles de données linguistiques afin de comprendre les tendances et les informations qui en découlent. Les LLM peuvent être utiles pour extraire des données à partir de grands volumes de texte, tels que les médias sociaux, les articles de presse, les blogs, etc., puis analyser ces données pour comprendre les tendances du marché, les opinions des clients, les sujets populaires, les événements en cours, etc. En utilisant les LLM, il est possible de générer des rapports d'analyse approfondis qui aident les organisations et les entreprises à prendre des décisions stratégiques basées sur une compréhension précise des données et des tendances actuelles.

2.2.3 Méthodologie de l'analyse des tendances

La méthodologie de l'analyse des tendances englobe trois points importants : [21]

- **Définition des objectifs :** La méthodologie commence par la définition claire des objectifs afin d'orienter la collecte de données.
- **Outils d'analyse :** Des techniques avancées telles que l'analyse sectorielle, l'analyse comparative et les modèles prédictifs sont utilisées dans la méthodologie.
- **Validation des données :** Un processus rigoureux de validation et de vérification des données est suivi pour assurer la fiabilité des résultats.

2.2.4 Utilisation de l'analyse des tendances

L'utilisation de l'analyse des tendances concerne : [21]

- **Orientation stratégique :** L'analyse des tendances éclaire les choix stratégiques d'une organisation en identifiant les opportunités et les menaces.
- **Innovation compétitive :** Elle favorise l'innovation en anticipant les besoins changeants des clients et en suivant les évolutions du marché.
- **Gestion des risques :** Elle contribue à une gestion proactive des risques en identifiant les potentiels impacts des évolutions futures sur l'organisation.

2.3 Travaux reliés

Le travail décrit dans [22], consiste à développer un système pour détecter les tendances des réseaux sociaux en utilisant des techniques de NLP. Le système collecte des données à partir de Twitter en utilisant une Interface de Programmation d'Applications (API) et utilise des bibliothèques NLP telles que NLTK, Gensim, TextBlob et SpaCy pour les traiter. Ensuite, le système classe les sujets en utilisant des algorithmes d'apprentissage automatique tels que le Naïve Bayes Multinomial (NBM). Enfin, le système analyse les résultats pour déterminer les tendances.

L'article [23] aborde la manière de découvrir les sujets tendances sur les plateformes de médias sociaux, en détaillant les méthodes et les techniques utilisées à cet effet. Les tweets en langue cinghalaise ont été collectés à partir de Twitter sur une période donnée, et des techniques de NLP ont été utilisées pour analyser et extraire les fonctionnalités importantes de ces tweets. Le regroupement hiérarchique séquentiel a été utilisé pour regrouper les tweets pertinents en groupes, et la classification SVM a été utilisée pour identifier les sujets tendances et les classer en fonction des fonctionnalités extraites. L'objectif de ces méthodes est d'identifier les sujets tendances dans les tweets en cinghalais et de les classer afin d'aider les entrepreneurs et les chercheurs à comprendre les préoccupations du public et à y répondre efficacement en temps opportun.

L'article [24] aborde l'importance du domaine du NLP et des techniques d'analyse du Big Data pour reconnaître les sujets d'actualité dans les textes des médias sociaux au manière de découvrir les sujets tendances sur les plateformes de médias sociaux à partir de données textuelles, en détaillant les méthodes et les techniques utilisées à cet effet. Cet article explique plusieurs méthodes utilisées pour l'analyse de données textuelles, le système collecte des données en utilisant une API et utilise des bibliothèques NLP et les algorithmes de clustering K-Means et Topic model. Enfin, cet article discute différentes méthodes largement utilisées pour identifier des thèmes communs.

L'étude présentée dans le mémoire [25] cherche à développer trois systèmes pour la détection des tendances afin de prédire la demande d'articles de mode. Dans le premier système "Impact de la Covid-19" essaye d'expliquer les ventes en ligne pendant la période de confinement, il se base donc sur le comportement des clients et incorpore l'impact de la pandémie. Pour ce là il a utilisé plusieurs algorithmes qui sont K-moyennes pour le clustering, Random Forest (RF) pour la classification, et régression polynomiale pour la correction des prévisions. Le deuxième système est "Modèle Basé sur un système de capture d'influence" ce système utilise trois modèles de prévision RF, LSTM et MBSCI. Le troisième système "Tendances de la mode" ce système se concentre sur les prévisions à court terme par famille de produits, il montre des résultats mitigés. Enfin, ce système permettra d'intégrer les recherches futures dans les systèmes de prévision des ventes.

Le travail proposé dans [26]abord le problème de la prévision des débits fluviaux, pour ce faire il utilise des ensembles et des réseaux de neurones profonds pour une grande précision, et il utilise des arbres de décision pour expliquer les prédictions et des outils comme CatBoost et t-SNE. Les prévisions ont une précision de 60% pour un horizon de trois ans.

Au fur et à mesure que les clients discutent de leurs produits sur Internet, l'article proposé dans [27] filtre les données collectées sur les réseaux sociaux pour recueillir un texte spécifique pour une balise PR. L'utilisation de ngrams pour extraire un texte spécifique des commentaires des clients en temps réel est examinée dans cette recherche. Cette étude a démontré que l'analyse des sentiments par n-gramme est efficace pour identifier les tendances dans les réseaux sociaux.

L'article [28] propose la détection des événements populaires en utilisant la méthode des dérivées sur Twitter. Le cadre proposé repose sur l'analyse des sentiments, la technique K-means et l'algorithme des dérivées de second ordre pour trouver les sujets d'événements populaires. L'objectif de la recherche est de découvrir les événements populaires, quel que soit le domaine, sur la plateforme de médias sociaux. Le cadre proposé a été testé sur quatre ensembles de données différents, tels que les événements scientifiques, politiques, sociaux et sportifs, et il a été constaté que la précision du cadre proposé varie de 88 % à 95 %, ce qui montre des résultats positifs. Le cadre proposé peut être étendu pour analyser les données de flux en continu afin de prédire les événements populaires en temps réel.

2.4 Récapitulatif des travaux réalisés

Le tableau 2 présente un récapitulatif des travaux examinés sur la détection des tendances

Tableau 2: Travaux réalisés sur la détection des tendances

Papiers	Approches	Base de données (Dataset)	Précision
[22]	TextBlob,	Ensemble de données collectées sur twitter	75%
[23]	Hierarchical Clustering, LDA, SVM	Ensemble de données collectées sur twitter	69.29%
[24]	K-Means Clustering, Approche bayésienne,	Pages Web, Hackforums, Microblogging Networks, Conférences, Bases de données académiques	/
[25]	LSTM	Ensemble des données sur réseaux sociaux	/
[26]	CatBoost	Conférences	60%
[27]	N-grams	Ensemble des données sur réseaux sociaux (Twitter)	/
[28]	K-Means	Ensemble de données collectées sur twitter	95%

2.5 Conclusion

Dans ce chapitre nous avons décrit quelques travaux en relation avec le domaine de l'analyse des tendances. Plusieurs approches ont été proposées et évaluées dans le cadre de ces travaux, certaines ont donné des résultats intéressants allant jusqu'à 95% de précision dans le cas des travaux de [28].

Chapitre III
Analyse des tendances : Expérimentation

Chapitre 03: Analyse des tendances : Expérimentation

3.1 Introduction

Dans le domaine du NLP, l'analyse de texte et la classification thématique sont des tâches fondamentales pour extraire des informations significatives à partir de grandes quantités de données textuelles. Ce processus comprend la préparation des données et leur traitement avec diverses techniques d'apprentissage automatique et d'apprentissage profond, et enfin les textes sont classés en groupes pour en extraire des mots-clés. Dans ce contexte, l'utilisation de modèles pré-entraînés tels que BERT pour classer les sujets joue un rôle crucial dans l'amélioration des performances du modèle.

3.2 Banque de données (Dataset)

L'ensemble de données que nous avons utilisé pour former le modèle provient du bigIR Research Group [29], qui est une vaste collection d'articles spécifiques des médias arabes sur le thème de la politique. Elle comprend 6499 fichiers texte. La taille totale de cet ensemble de données est 32 Ko, chaque fichier contient un texte arabe groupé.
L'ensemble utilisé fait partie d'un ensemble de données beaucoup plus vaste sur des dossiers pour d'autres catégories (par exemple culture, finance, sports, etc.). Cet ensemble de données le plus important s'appelle SANAD.

Comme exemple: de texte :

"ونقل المرصد عن مصادر متقاطعة أن "التنظيم المتطرف أرسل إلى جبهات القتال في عين العرب (كوباني) كتيبة مؤلفة من نحو 140 عنصراً غالبيتهم دون سن الـ 18، من المنضمين حديثاً إلى معسكرات التدريب التابعة للتنظيم"، وتمكن المرصد السوري لحقوق الإنسان من توثيق مصرع 6 عناصر منهم، قضوا في اشتباكات مع وحدات حماية الشعب الكردي ..."

3.3 Les outils utilisés

Pour réaliser notre projet, nous avons employé divers outils :

- **Visual Studio Code**

(VS Code) est un éditeur de code source et un environnement de développement intégré (IDE) de Microsoft. Il est open-source et cross-platform, c'est-à-dire qu'il fonctionne sur Windows, Linux et Mac. Il a été conçu pour les développeurs web, mais il prend en charge de nombreux autres langages de programmation tels que C++, C#, Python, Java, etc [30].

- **Python**

Est un langage de programmation puissant et facile à apprendre. Il dispose de structures de données de haut niveau et permet une approche simple mais efficace de la programmation orientée objet [31].

3.4 Bibliothèques utilisées dans le code

Nous avons utilisé un ensemble de bibliothèques Python à savoir :

➢ RegEx

La bibliothèque re est un outil puissant pour la manipulation de texte et l'identification de modèles. Les expressions régulières ou RegEx sont utilisées pour savoir si le texte correspond à un certain modèle ou pour en extraire des informations spécifiques.

La bibliothèque fournit un large éventail de fonctions pour travailler avec des expressions régulières, elle joue un rôle crucial dans les étapes de préparation et de nettoyage des données [32].

➢ NLTK

La bibliothèque NLTK (Natural Language Toolkit) est une bibliothèque spécialisée dans le NLP. Sa fonction principale est de fournir des outils et des ressources pour analyser et traiter des textes en langage naturel. Cette bibliothèque comprend un large éventail de fonctions et d'outils tels que :

- Divisez les textes en mots et en phrases.
- Identifier des parties du discours dans des textes.
- Identifier les structures grammaticales dans les textes.

Analyser les significations et les concepts des textes.

La bibliothèque NLTK contient un ensemble de fonctions que nous avons utilisé comme :

➢ Word tokenize

Cette bibliothèque divise le texte en mots individuels (jetons ou tokens), elle prend en compte la ponctuation et l'espacement des mots [33].

➢ PyArabic

C'est une bibliothèque de logiciels àpen source dédiée au traitement des textes en langue arabe. Cette bibliothèque fournit une gamme d'outils et de fonctions pour analyser les textes et diviser les mots… Nous l'avons utilisé par exemple :

→ Strip Taksheel

Cette bibliothèque permet de faire la suppression des accents et des signes diacritiques (des voyelles et des signes diacritiques) du texte arabe [34].

➢ Sentence Transformer

Cette bibliothèque peut convertir des textes en représentations numériques (embeddings) à l'aide de modèles pré-entraînés. La fonction de cette bibliothèque est de convertir des phrases et des textes en représentations numériques vectorielles. Elle utilise des techniques d'apprentissage profond (DL) et d'apprentissage automatique (ML) [35].

➢ Approximation et projection d'une variété uniforme (UMAP)

Cette bibliothèque est un outil puissant pour l'exploration et l'analyse de données complexes en Python. Elle permet de faire la réduction de dimensionnalité pour une meilleure compréhension et visualisation des données.

➢ Regroupement spatial hiérarchique basé sur la densité des applications avec bruit (HDBSCAN)

La bibliothèque HDBSCAN est un outil qui collecte et organise les données en fonction de leur densité, créant une structure arborescente hiérarchique d'éléments. Cela aide à analyser différents ensembles de données et à gérer efficacement le bruit.

HDBSCAN est une bibliothèque de clustering utilisée dans le cadre de l'apprentissage automatique (ML) [36].

➢ Scikit-learn

C'est bibliothèque de logiciels open source est utilisée pour exécuter de nombreux algorithmes d'apprentissage automatique et d'analyse de données.

Elle contient de nombreuses instructions et fonctions pour traiter les textes et les convertir en représentations numériques, parmi lesquelles nous avons utilisé :

- **Count Vectorizer**

C'est un outil qui convertit un ensemble de textes en un raccourci contenant le nombre de répétitions par mot dans le texte [37].

➢ Arabic stop words

La bibliothèque Arabic Stop words comprend une liste des mots vides. Elle est utilisée pour supprimer les mots vides des textes arabes. Les mots vides sont des mots qui apparaissent fréquemment dans les textes et qui n'ont pas de signification spécifique, tels que « et », « dans », « sur » et autres [38].

➢ Bertopic

La bibliothèque BERTOPIC est un outil puissant utilisé pour analyser les sujets tirés des textes. La Bibliothèque s'appuie sur des modèles d'apprentissage automatique pour analyser les textes et découvrir les sujets clés.

Pour cette bibliothèque, nous avons utilisé un ensemble de formats et de modèles affiliés :

→ KeyBERT Inspired

Permet l'extraction de mots-clés à partir de textes à l'aide de BERT.

Le fondement de KeyBERT réside dans l'exploitation de la puissance des modèles d'apprentissage automatique (ML) et des techniques d'apprentissage profond (DL) [39].

➢ Pandas

Permet de manipuler et organiser les données. Pandas offre des structures de données faciles à utiliser pour représenter les données de manière homogène et fournit des fonctions de lecture et d'écriture à partir et vers différentes sources de données telles que les fichiers CSV et Excel… etc. [40]

➢ Matplotlib

C'est une bibliothèque open source de Bayton utilisée pour créer des graphiques et des diagrammes. Fournit des outils flexibles pour concevoir et personnaliser plusieurs types de graphiques. Nous avons utilisé :

→ Matplotlib.pyplot

Une sous-unité dans Matplotlib qui facilite la création de graphiques à l'aide d'une interface similaire à MATLAB. [41]

➢ Seaborn

Une bibliothèque conçue pour dessiner des données statistiques facilement utilisables. Il est basé sur Matplotlib. [42]

➢ Pathlib

C'est une bibliothèque est utilisée pour traiter les fichiers. Nous pouvons par exemple :

- Créer un fichier.
- Extraire le nom du fichier.
- Lire et écrire des fichiers.
- Modifier les extensions de fichiers.
- ...etc. [43]

3.5 Configuration expérimentale

Nous avons utilisé :

- Un ordinateur portable HP.
- Mémoire vive : 8 Go.
- Système d'exploitation : Windows 10 Professionnel 64 bits.
- Processeur : Intel ® Core(™) i5-6200U CPU @ 2.30 GHz, ~ 2.4 GHz .
- Carte graphique : Intel ® HD Graphics 520.

3.6 Expérimentation (Implémentation)

Notre expérimentation regroupe un ensemble d'étape :

3.6.1 Prétraitement

Le prétraitement est une technique courante dans le domaine du NLP. À cette étape, les données sont nettoyées et préparées pour être utilisées par le modèle. Cette préparation comprend les étapes suivantes :

- Décode et encode le texte au format UTF-8 pour une gestion cohérente.
- Suppression des liens : remplacer tout lien par le mot « lien ».
- Remplacement des chiffres par le mot nombre : pour améliorer le regroupement.
- Remplacement des noms d'utilisateurs par le mot utilisateur : pour nettoyer les textes des informations non nécessaires.
- Suppression des caractères répétitifs pour rendre les textes plus homogènes.
- Suppression des diacritiques : pour simplifier les mots.
- Conversion du texte en minuscules et suppression des espaces superflus.

Par exemple : nous avons le texte suivant.

"ونقل المرصد عن مصادر متقاطعة أن "التنظيم المتطرف أرسل إلى جبهات القتال في عين العرب (كوباني) كتيبة مؤلفة من نحو 140 عنصراً غالبيتهم دون سن الـ 18، من المنضمين حديثاً إلى معسكرات التدريب التابعة للتنظيم"، وتمكن المرصد السوري لحقوق الإنسان من توثيق مصرع 6 عناصر منهم، قضوا في اشتباكات مع وحدات حماية الشعب الكردي "

Après avoir appliqué la fonction de prétraitement, le texte devient comme ceci :

"ونقل المرصد عن مصادر متقاطعة ان التنظيم المتطرف أرسل الى جبهات القتال في عين العرب كوباني كتيبة مؤلفة من نحو رقم عنصرا غالبيتهم دون سن ال رقم من المنضمين حديثا الى معسكرات التدريب التابعة للتنظيم وتمكن المرصد السوري لحقوق الانسان من توثيق مصرع رقم عناصر منهم قضوا في اشتباكات مع وحدات حماية الشعب الكردي "

3.6.2 Intégration des données (Data Embedding)

La fonction Embedding est une technique d'apprentissage profond (DL). Dans le domaine du NLP, elle prend en entrée les textes traités par la fonction de prétraitement ; elle est utilisée à ce stade pour convertir des textes en représentations numériques comme suit:

- Charger le modèle LaBSE pré-entraîné à l'aide de la bibliothèque Sentence-transformers pour convertir les textes en représentations numériques.
- La fonction encode du modèle LaBSE reçoit en entrée une liste de textes, divise le texte en mots ou phrases, et chaque mot et phrase est converti en une représentation numérique.

Par exemple : nous avons le texte :

"ونقل المرصد عن مصادر متقاطعة ان التنظيم المتطرف أرسل الى جبهات القتال في عين العرب كوباني كتيبة مؤلفة من نحو رقم عنصرا غالبيتهم دون سن ال رقم من المنضمين حديثا الى معسكرات التدريب التابعة للتنظيم وتمكن المرصد السوري لحقوق الانسان من توثيق مصرع رقم عناصر منهم قضوا في اشتباكات مع وحدات حماية الشعب الكردي"

Après avoir appliqué la fonction embedding ce texte devient (ces chiffres ne sont que des exemples illustratifs) : " [0.123, -0.456, 0.789, ……., 0.321]"

3.6.3 Approximation et projection d'une variété uniforme (UMAP)

UMAP prend en entrée des données transformées en représentations numériques par une fonction embedding. Elle fonctionne de la manière suivante :

- Convertir un ensemble de données de grande dimension en un ensemble de données de plus petite dimension tout en préservant la structure et les relations de base entre les données afin de faciliter la visualisation. Ce qui conduit à améliorer les performances du modèle et à réduire le temps d'exécution.

3.6.4 Regroupement spatial hiérarchique basé sur la densité des applications avec bruit (HDBSCAN)

HDBSCAN est un algorithme de clustering d'apprentissage automatique basé sur la densité qui prend en entrée des données de plus petite dimension traitées par une fonction UMAP. Il est principalement utilisé pour identifier des groupes (clusters) de points de données dans un ensemble de données capables de gérer des ensembles de données qui ont :

- **Densités variées** : les clusters peuvent avoir différentes densités (nombre de points par unité de surface).
- **BRUIT** : il peut identifier et exclure les valeurs aberrantes (points de bruit) des clusters.
- **Formes inégales** : les grappes peuvent avoir une forme irrégulière.

Les étapes les plus importantes qu'une fonction HDBSCAN applique aux données à ce stade sont :

- **Identification des points essentiels** : L'algorithme commence par identifier les points essentiels, qui sont des points à haute densité dans les données.

- **Construction du graphe de densité** : Les clusters se développent autour de chaque point en incorporant les points voisins à haute densité, ce qui crée un graphe reflétant les connexions entre les points en fonction de leur densité.
- Ce processus d'expansion se poursuit jusqu'à ce que la densité diminue considérablement.
- **Analyse hiérarchique** : Analyser la hiérarchie de densité pour extraire les clusters.
- **Extraction des clusters finaux** : déterminer les clusters finaux en regroupant les points qui n'appartiennent à aucun groupe dans la catégorie bruit.

3.6.5 Vectorisation (Vectorizer)

Vectorizer est une techninque utilisée par l'apprentissage automatique pour transformer des textes en une matrice de termes basée sur la fréquence des mots et prenant des ensembles de données spécifiés par la fonction hdbscan comme entrée pour extraire les modèles les plus courants dans les textes de chaque ensemble. Son travail à ce stade est le suivant :

- Supprimez les mots vides à l'aide de stop words qui contient la liste des mots vides tels que : qui, dans, ça...
- Diviser le texte en unités telles que des mots individuels ou des groupes de mots (n-grammes). « Ngram range » est défini sur (1, 3), ce qui signifie que des mots impairs, des bigrammes et des trigrammes sont utilisés.

Par exemple : on a cette phrase

" العراق يخسر أمام كوريا الجنوبية في نصف نهائي كأس آسيا , تأهل منتخب كوريا الجنوبية إلى نهائي كأس أمم آسيا لكرة القدم"

Après avoir appliqué la technique Vectorizer

	يخسر	أمام	أمم	منتخب	القدم	كرة	كأس اسيا	تأهل	نهائي	نصف	الجنوبية	كوريا	العراق
0	1	1	1	1	1	1	2	2	1	2	1	2	2	1

3.6.6 KeyBERT Inspired

La fonction KeyBERT Inspired est une version modifiée ou inspirée de la bibliothèque KeyBERT et est utilisée pour extraire des mots-clés de textes à l'aide d'un modèle d'apprentissage automatique préentraîné BERT. Cette fonction analyse le contexte et la signification des textes pour générer une liste des mots-clés les plus importants dans chaque texte. Son travail à ce stade est le suivant :

- **Extraire les mots-clés** : À l'aide d'un modèle linguistique préentraîné BERT, les phrases clés sont extraites en appliquant le modèle à chaque phrase du texte et en identifiant les mots les plus pertinents et les plus utiles pour représenter le sens de la phrase.

3.6.7 BERTopic

Dans cette partie, nous créons et entraînons un modèle de modélisation thématique utilisant la techninque BERTopic, qui utilise les techniques et algorithmes que nous avons mentionnés précédemment, qui sont les suivants :

- embedding_model : pour obtenir une représentation du texte.
- umap_model : pour réduire la dimensionnalité et représenter les données dans un espace de faible dimension.

- hdbscan_model : pour collecter les données représentées après réduction de dimensionnalité.
- vectorizer_model : pour la conversion de texte en une matrice de nombres basée sur la fréquence des mots.
- representation_model: pour créer une liste des mots-clés les plus importants

Ensuite, nous utilisons la fonction fit_transform pour entraîner et appliquer le modèle BERTopic à l'ensemble de données que nous connaissions auparavant comme entrée. Le modèle analyse les données textuelles après avoir exécuté toutes les étapes de traitement mentionnées précédemment, afin d'en extraire finalement les sujets et les actualités tendances.

3.7 Interface de l'application

3.7.1 Editeur utilisé pour développer l'interface

- **Qt Creator**

Qt Creator est un environnement de développement intégré (IDE) spécialement conçu pour créer des applications pour diverses plates-formes, notamment les ordinateurs de bureau, les systèmes embarqués et les appareils mobiles (Android et iOS). Il fait partie du framework Qt populaire et est utilisé pour développer des applications logicielles en utilisant le langage C++. Qt Creator fournit un éditeur de texte puissant avec des fonctionnalités telles que la coloration syntaxique et la complétion automatique du code, ainsi que des outils de conception d'interface utilisateur graphique (GUI) pour créer et modifier facilement des interfaces. Il fournit également des outils de test et de débogage d'applications, ce qui en fait un outil utile et puissant pour développer des applications à l'aide du framework Qt [44].

3.7.2 Description de l'interface

Cette interface utilisateur (voir Figure 10) a été conçue pour permettre aux utilisateurs d'accéder facilement aux tendances en fonction du texte spécifique qu'ils entrent d'une manière facile et organisée.

Lors de l'accès à l'interface utilisateur, l'utilisateur se trouve devant un champ de saisie de texte qui permet à l'utilisateur d'entrer le texte spécifique qu'il souhaite rechercher pour les nouvelles associées. À côté du champ d'entrée, il y a le bouton Rechercher, qui permet à l'utilisateur d'effectuer la recherche. Après avoir cliqué sur le bouton Rechercher, les résultats sont affichés dans une liste triée par importance, où les actualités les plus courantes (tendances) sont présentées. Chaque titre de la liste est affiché de manière ciblée, ce qui permet à l'utilisateur de comprendre le contenu de base de la nouvelle.

Grâce à cette interface, les utilisateurs peuvent facilement et rapidement rechercher des nouvelles courantes et connaître les derniers développements dans différents domaines. La figure suivante donne une vue d'ensemble de l'interface :

Figure 10: Interface utilisateur

3.7.3 Exemples d'analyse des tendances

L'analyse des tendances de notre Datasets produit le résultat montré dans la figure 11. Il faut noter que l'utilisateur peut choisir en fonction de ses besoins le nombre des tendances à afficher, par exemple dans la figure 11 l'utilisateur a demandé un nombre de 12 de tendances.

Figure 11 : Exemple d'analyse des tendances

3.8 Conclusion

Dans ce chapitre, nous avons décrit brièvement le processus de mise en œuvre de notre projet en décrivant l'environnement de développement, la banque de données et l'approche adoptée pour l'extraction des tendances. Nous avons utilisé Python, un langage de programmation qui offre de nombreuses fonctionnalités. L'expérimentation réalisée consiste à appliquer un certain nombre d'opérations de prétraitement pour nettoyer et préparer ces données pour fine tunner un modèle préentrainé spécifique pour extraction de tendances. Le résultat est un modèle LLM qui analyse les textes pour en extraire les sujets et les actualités tendances.

CONCLUSION GENERALE

Notre projet de fin d'étude a commencé par un examen complet de l'importance des réseaux sociaux comme moyen de communication et d'information entre les étudiants, les enseignants et l'administration des établissements universitaires. Il met en lumière la façon dont ces plates-formes sont devenues une source majeure de nouvelles et d'informations, ce qui nous incite à comprendre et à analyser attentivement les tendances.

Ce projet a adopté une approche composée de plusieurs étapes principales. Tout d'abord, des données ont été recueillies à partir des différentes plateformes de médias sociaux utilisées par les étudiants algériens. Nous avons ensuite appliqué des techniques d'analyse des tendances pour comprendre les tendances dominantes. En effet, nous avons utilisé des algorithmes d'apprentissage automatique et en particulier de l'apprentissage en profondeur (Bertopic) pour construire un modèle de langue capable de classer efficacement les textes arabes. Pour faciliter l'interaction des utilisateurs avec ce modèle nous avons développé une interface utilisateur qui demande en entrée des textes et renvoi en sortie les tendances.

L'outil réalisé peut être utilisé dans un large éventail de domaines, y compris les médias, le marketing et même les institutions académiques, pour mieux comprendre les besoins et les tendances.

Sur la base de ces résultats, nous pouvons dire que notre projet est encore à son début et par conséquent, il est recommandé de mener d'avantage de recherches pour développer des modèles plus précis pour comprendre les textes arabes, ainsi que d'améliorer l'interface utilisateur pour inclure des fonctionnalités supplémentaires améliorant sa compétence et son efficacité dans la présentation d'informations.

Références

[1]	Les-r-seaux-sociaux, : https://www.slideserve.com/ghalib/les-r-seaux-sociaux. [Consulté : Mars 2024].
[2]	vente en ligne mckinsey dessine la strategie de l'ugap, : https://www.consultor.fr/articles/vente-en-ligne-mckinsey-dessine-la-strategie-de-l-ugap. [Consulté : Avril 2024].
[3]	M. Thelwall, «Social network sites: Users and uses. Advances in computers,,» 2009.
[4]	What is artificial intelligence : https://www.netapp.com/fr/artificial-intelligence/what-is-artificial-intelligence/. [Consulté : janvier 2024].
[5]	«ia-data-analytics.fr,» 2022. : https://ia-data-analytics.fr/machine-learning/. [Accès le janvier 2024].
[6]	Le cycle de vie du développement de l'apprentissage automatique, : https://medium.com/@dancerworld60/the-machine-learning-development-life-cycle-mldc-a-comprehensive-guide-8f6ff35541f5. [Consulté : Avril 2024].
[7]	P. É. Blent, Apprentissage supervise, : https://blent.ai/blog/a/apprentissage-supervise-definition. [Consulté : Mai 2024].
[8]	S. Y. e. T. d. Salim, «Extraction de motifs basée sur word2vec,» Université Saad Dahleb Blida - 1, 2020.
[9]	J. Robert, Reinforcement learning, : https://datascientest.com/reinforcement-learning. [Consulté : Mai 2024].
[10]	L. Galiana, Modelisation : https://pythonds.linogaliana.fr/content/modelisation/. [Consulté : Avril 2024].
[11]	A. Crochet-Damais, Natural language processing ,27 Juillet 2022. : https://www.journaldunet.fr/intelligence-artificielle/guide-de-l-intelligence-artificielle/1501887-natural-language-processing-nlp/. [Consulté : Février 2024].
[12]	M. Hasanuzzaman, Relationship-between AI ML DL and NLP, : https://www.researchgate.net/figure/Relationship-between-AI-ML-DL-and-NLP-7_fig8_343079524. [Consulté : Avril 2024].
[13]	N. K. N. e. P. Singh, Impact of word embedding models on text analytics in deep learning environment , : https://link.springer.com/article/10.1007/s10462-023-10419-1. [Consulté : Mai 2024].
[14]	P. JEAN-BAPTISTE, Les llm large language model, 28 Septembre 2023. : https://fr.linkedin.com/pulse/ia-les-llm-large-language-model-philippe. [Consulté : Février 2024].
[15]	Large Language Models (LLM),18 Mai 2023. : https://datascientest.com/large-language-models-tout-savoir. [Consulté : Février 2024].
[16]	Topic modeling, : https://www.qualtrics.com/experience-management/research/topic-modeling/. [Consulté : Mars 2024].
[17]	What is topic modelling, : https://anivation.wordpress.com/2018/05/02/tuberculosis-topic-modeling/. [Consulté : Avril 2024].
[18]	D. Groult,Definition google bert,26 November 2023. : https://www.noiise.com/definition/google-bert/. [Consulté : Mars 2024].
[19]	BERTopic, : https://maartengr.github.io/BERTopic/api/bertopic.html. [Consulté : Mars 2024].
[20]	N. Jain, Quest ce que lanalyse des tendances, 30 Novembre 2023. : https://ideascale.com/fr/blogues/quest-ce-que-lanalyse-des-tendances/. [Consulté : Janvier 2024].
[21]	Quest ce que lanalyse de donnees, : https://www.questionpro.com/blog/fr/quest-ce-que-lanalyse-de-donnees. [Consulté : janvier 2024].
[22]	B. a. Djebrani.A, «Détection de tendances et prévision de la demande d'articles de mode par les données massives et l'intelligence artificielle,» universite Saad Dahlab, Alger -Blida-, 2020.

[23]	L. J. e. S. Ahangama, «Trend Detection in Sinhala Tweets Using Clustering and Ranking Algorithms,» 2020.
[24]	R. K. K. e. G. A. A. Al-Talib, «A Survey Study on Extracting Trending Topics from Textual Data,» 2022.
[25]	R. Sleiman, «Détection de tendances et prévision de la demande d'articles de mode par les données massives et l'intelligence artificielle,» DOCTORAT DELIVRE PAR CENTRALE LILLE (Doctorat), 2022.
[26]	M. C. a. all, «Visualization of Research Trending Topic Prediction: Intelligent Method for Data Analysis,» 2021.
[27]	C. M. N. a. all, «An Ngram-Based Approach to Determine Trends and Patterns in the Social Networks,» 2023.
[28]	P. V. T. a. J. Rani, «DETECTING TRENDING EVENT TOPICS USING SENTIMENT DRIVEN DERIVATIVES METHOD ON TWITTER,» 2021.
[29]	«bigIR Research Group,» 2016 mai 6. : https://www.dropbox.com/scl/fo/tvw89pzdplgn890k6ypew/AADIS10uGQ31WjPoN17uAuk?rlkey=lfcx66yz6y3cvkuyhl49nfu5h&e=1&st=5nmwbxzz&dl=0. [Accès le mai 2024].
[30]	Visual studio code, : https://bility.fr/definition-visual-studio-code/. [Consulté : Mars 2024].
[31]	Python, : https://docs.python.org/3/library/csv.html. [Consulté : Février 2024].
[32]	Python RegEx , : https://www.w3schools.com/python/python_regex.asp. [Consulté : Féverier 2024].
[33]	Nltk, : https://www.nltk.org. [Consulté : Janvier 2024].
[34]	PyArabic, : https://snyk.io/advisor/python/PyArabic/functions/pyarabic.araby.strip_tashkeel. [Consulté : Janvier 2024].
[35]	Sentence Transformer, : https://www.sbert.netindex.html. [Consulté Janvier 2024].
[36]	HDBSCAN, : https://itsudit.medium.com/discovering-the-power-of-hdbscan-clustering-for-unsupervised-learning-d67273e28c5b. [Consulté : Février 2024].
[37]	Using CountVectorizer,07 Juillet 2022. : https://www.geeksforgeeks.org/using-countvectorizer-to-extracting-features-from-text. [Consulté : Février 2024].
[38]	Arabic stop words, : https://www.kaggle.com/code/mpwolke/arabic-stop-words-w2v. [Consulté : Janvier 2024].
[39]	BERTopic, : https://github.com/MaartenGr/BERTopic. [Consulté : Janvier 2024].
[40]	Python pandas, : https://www.w3schools.com/python/pandas/pandas_intro.asp. [Accès le Avril 2024].
[41]	Python matplotlib, : https://realpython.com/python-matplotlib-guide/. [Consulté Avril 2024].
[42]	Python seaborn, : https://www.geeksforgeeks.org/python-seaborn-tutorial/. [Consulté Avril 2024].
[43]	Python Pathlib, : https://towardsdatascience.com/10-examples-to-master-python-pathlib-1249cc77de0b. [Accès le Avril 2024].
[44]	Qt Creator, : https://en.wikipedia.org/wiki/Qt_Creator. [Consulté : Mai 2024].