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

Rapport du projet text mining

Analyse, Résumé et Extraction de Données depuis des Articles Scientifiques

Réalisé par :
SOUSSOU Youness
EL HARNAF Nihad

Encadré par :
Mr. TABII Youness

Année universitaire : 2024 / 2025

Contents

1 Contexte général de projet : 2
1.1 Introduction : . . . 2
1.2 Enoncé de la problématique : . . . 2
1.3 Objectifs : . . . 2
1.4 Conclusion : . . . 3

2 Cadre théorique et Revue de Littérature : 4
2.1 Introduction . . . 4
2.2 Fondements théoriques du NLP : . . . 4
2.2.1 Définition et Objectifs du NLP : . . . 4
2.2.2 Techniques et Approches du NLP . . . 5
2.3 BERT :ModèlesBaséssurlesTransformers pour le Question Answering . . . 6
2.3.1 Aperçu de BERT : . . . 6
2.3.1.1 Conception des Transformers : . . . 6
2.3.1.2 Architecture et Mécanisme de BERT : . . . 7
2.3.1.3 Pré-Entraînement et Fine-Tuning : . . . 9
2.3.2 Application de BERT au Question Answering et à la Génération de Résumés 9
2.3.2.1 Fondements du Question Answering . . . 9
2.3.2.2 Utilisation de BERT pour le Question Answering . . . 10
2.4 Résumé Automatique de Texte . . . 10
2.4.1 Approches Extractives et Abstractives . . . 10
2.4.1.1 Approches Extractives . . . 11
2.4.1.2 Approches Abstractives . . . 11
2.4.2 Utilisation de Modèles de Langage pour le Résumé . . . 11
2.4.2.1 Avantages des Modèles de Langage . . . 11
2.4.2.2 BERT Extractive Summarizer . . . 11
2.5 État de l'Art dans le Question Answering et le Résumé Automatique . . . 12
2.5.1 Évolution du Question Answering . . . 12
2.5.2 Avancées dans le Résumé Automatique . . . 12
2.6 Conclusion . . . 13

3 Analyse et conception du projet 14
3.1 Introduction . . . 14
3.2 Analyse de l'existant . . . 14
3.3 Extraction de Texte depuis des Sources Variées et traitement de texte . . . 15
3.4 Question Answering par BERT . . . 16
3.5 Génération de Résumés par BERT Summarizer . . . 17
3.6 Interface Utilisateur avec Streamlit . . . 18
3.7 Conception du fonctionnement Global du ChatBot . . . 19
3.8 Conclusion . . . 19

4 La Réalisation de la solution **21**
4.1 Introduction 21
4.2 Outils de Réalisation 21
4.2.1 Python : Langage de Programmation Principal 21
4.2.2 Librairies et Frameworks Utilisés 22
4.3 Mise en Œuvre 23
4.3.1 Sélection des Sources d'Articles pour les Tests 23
4.3.2 Résultats de l'Extraction de Texte et de la génération de Résumés 25
4.3.2.1 Pour l'article du site HTML 25
4.3.2.2 Pour l'article du site PDF 27
4.3.3 Résultats des Réponses aux Questions 28
4.3.3.1 Pour l'article du site HTML 28
4.3.3.2 Pour l'article du site PDF 29

List of Figures

2.1 L'architecture des transformers 8

3.1 Extraction de Texte depuis des Sources Variées et traitement de texte 15
3.2 Question Answering par BERT 16
3.3 Génération de Résumés par BERT Summarizer 17
3.4 Interface Utilisateur avec Streamlit 18
3.5 Conception Global Du projet 19

4.1 Logo- Python 22
4.2 La page d'accueil de l'interface utilisateur 23
4.3 Capture d'écran du site de test HTML 24
4.4 Capture d'écran du site de test PDF 24
4.5 Résultats de l'Extraction de Texte et de la génération de Résumés pour le site HTML 26
4.6 Résultats de l'Extraction de Texte et de la génération de Résumés pour un article pdf 27
4.7 question 1 - page HTML 28
4.8 question 2 - page HTML 28
4.9 question 3 - page HTML 29
4.10 question 1 - article PDF 29
4.11 question 2 - article PDF 30

Introduction

L'avènement du traitement automatique du langage naturel (NLP) a ouvert de nouvelles perspectives passionnantes dans la façon dont nous interagissons avec l'information textuelle et la façon dont nous extrayons des connaissances à partir de vastes corpus de données. Le NLP, en combinant l'intelligence artificielle et la linguistique, permet aux machines de comprendre, d'analyser et de générer du langage humain de manière similaire à celle des humains.

Dans ce contexte, l'analyse d'articles scientifiques occupe une place centrale pour la diffusion et la découverte de nouvelles connaissances. Cependant, avec la croissance exponentielle de la littérature scientifique, il devient de plus en plus complexe et chronophage pour les chercheurs de filtrer et de synthétiser les informations pertinentes. C'est dans ce cadre que notre projet trouve sa pertinence.

Le présent rapport détaille la conception et la mise en œuvre d'un système innovant de traitement automatique du langage naturel (NLP) dans le domaine de l'analyse d'articles scientifiques. Ce système, incarné par un ChatBot sophistiqué, vise à simplifier l'extraction de données significatives à partir d'articles scientifiques divers, à générer des résumés synthétiques et à répondre aux questions spécifiques des utilisateurs concernant le contenu de ces articles.. Ce projet se fonde sur l'utilisation de modèles de deep learning, en particulier ceux basés sur BERT, pour améliorer la compréhension du langage et la génération de réponses cohérentes.

Au cours de ce rapport, nous détaillerons les étapes clés de ce projet ambitieux.Nous aborderons les fondements théoriques du NLP, l'examen de la littérature existante, l'analyse des besoins, la conception du système et son implémentation.

Ce projet témoigne de l'impact croissant du NLP dans le domaine de la recherche et de la diffusion scientifiques. Il démontre la façon dont la convergence de la technologie et du language humain peut transformer la manière dont nous traitons l'information et interagissons avec elle.

Chapitre 1

Contexte général de projet :

1.1 Introduction :

Dans ce chapitre, Nous exposons ensuite le contexte général, la problématique et les objectifs du projet.

1.2 Enoncé de la problématique :

Dans un monde régi par l'internet, la fondation nationale pour la science (National Science Foundation- NSF) a observé une prolifération sans précédent d'articles scientifiques. Plus de deux millions d'articles de sciences et d'ingénierie (SE) ont été publiés dans le monde en 2022, avec une croissance annuelle moyenne de près de 9,25% au cours des dix dernières années. Les statistiques du centre national pour la science et l'ingénierie indiquent que le domaine de la santé domine le paysage de la recherche mondiale avec 23,64% des publications, suivi de près par l'ingénierie (16,94%), l'informatique (11,97%), la physique (9,12%), la chimie (6,20%) et les mathématiques (2,39%).

Cependant, cette abondance d'information présente un double défi. D'une part, les chercheurs doivent s'efforcer de filtrer les articles pertinents pour leurs besoins spécifiques parmi cette profusion. D'autre part, une fois identifiés, la compréhension en profondeur de ces articles peut s'avérer chronophage. Pour répondre à ces problèmes cruciaux, comment pouvons-nous exploiter les avancées du traitement automatique du langage naturel pour développer un ChatBot capable d'extraire des données à partir d'articles scientifiques, de générer des résumés pertinents et de répondre aux questions des utilisateurs, simplifiant ainsi l'accès à la connaissance scientifique et optimisant la recherche d'informations ?

1.3 Objectifs :

Les objectifs de ce projet se concentrent sur l'élaboration d'un système de traitement automatique du langage naturel (NLP) visant à faciliter l'analyse et l'accès à la vaste quantité de données contenues dans les articles scientifiques. Ces objectifs se déclinent comme suit :

- **Extraction de Données Précises :** Concevoir un ChatBot capable d'extraire de ma-

nière précise les informations essentielles contenues dans les articles scientifiques, en détectant et en sélectionnant les passages pertinents.

- **Génération de Résumés Pertinents :** Mettre en place un mécanisme de génération de résumés automatiques pour les articles extraits, fournissant aux utilisateurs une vue synthétique et concise du contenu.
- **Réponse aux Questions Utilisateurs :** Développer une fonctionnalité qui permet au ChatBot de répondre aux questions spécifiques des utilisateurs concernant le contenu des articles, en fournissant des réponses cohérentes et contextuellement appropriées.
- **Interactivité Utilisateur :** Créer une interface conviviale qui facilite l'interaction entre les utilisateurs et le ChatBot, offrant une expérience intuitive et accessible pour l'extraction d'informations, la génération de résumés et l'obtention de réponses.
- **Optimisation des Performances :** Évaluer et optimiser les performances du système,tant du point de vue de l'extraction de données que de la génération de résumés, afin de garantir des résultats de haute qualité.

Ces objectifs s'alignent sur la nécessité croissante de simplifier l'accès à l'information scientifique et de faciliter le processus de recherche et d'assimilation des connaissances. En combinant les avancées du NLP avec une conception centrée sur l'utilisateur, ce projet vise à repousser les limites actuelles de l'analyse d'articles scientifiques.

1.4 Conclusion :

À travers la formulation de notre problématique, nous avons souligné le besoin pressant de développer un ChatBot exploitant les avancées du traitement automatique du langage naturel pour simplifier l'accès et la compréhension des articles scientifiques. En alignant nos objectifs sur cette problématique, nous avons établi la base solide sur laquelle nous bâtirons les solutions dans les chapitres à venir.

Chapitre 2

Cadre théorique et Revue de Littérature :

2.1 Introduction

Ce chapitre plonge dans le cadre conceptuel de notre projet en explorant le Traitement Automatique du Langage Naturel (NLP) et en effectuant une revue approfondie de la littérature pertinente. Le NLP est essentiel à notre ChatBot pour l'analyse d'articles scientifiques et la génération de résumés. Nous examinerons les bases du NLP, discuterons du modèle BERT pour le Question Answering, explorerons le résumé automatique de texte et passerons en revue l'état actuel du Question Answering et du Résumé Automatique. Cette exploration théorique jettera les bases nécessaires pour la mise en œuvre de notre ChatBot.

2.2 Fondements théoriques du NLP :

Cette section explore les bases essentielles du Traitement Automatique du Langage Naturel (NLP) pour éclairer notre approche dans la création du ChatBot. Nous définirons les objectifs centraux du NLP et passerons en revue les techniques et approches clés. Cette compréhension sous-jacente est cruciale pour saisir le rôle du NLP dans notre projet.

2.2.1 Définition et Objectifs du NLP :

Le Traitement Automatique du Langage Naturel (NLP) constitue un domaine majeur de l'intelligence artificielle qui se consacre à permettre aux ordinateurs de comprendre, d'interpréter et de générer le langage humain de manière naturelle. Au cœur du NLP se trouve la tâche complexe de faire en sorte que les machines puissent traiter le langage humain, qui est riche en nuances, ambiguïtés et variations.

L'objectif fondamental du NLP est de créer des modèles et des algorithmes qui permettent aux machines d'analyser et de comprendre le langage humain dans toutes ses formes : texte, audio, et même langage corporel. Cela englobe plusieurs sous-objectifs interconnectés :

- **Analyse Syntaxique et Sémantique :**Le NLP cherche à comprendre la structure grammaticale des phrases ainsi que leur signification profonde. Cela implique la détection des

parties du discours, la construction de relations entre les mots, et la saisie du contexte pour interpréter le sens.

- **Reconnaissance d'Entités Nommées :**Les systèmes NLP identifient et catégorisent les éléments spécifiques tels que les noms de personnes, les lieux, les dates, etc., dans le texte, ce qui est crucial pour extraire des informations pertinentes.
- **Question Answering :** Le NLP vise à permettre aux machines de répondre aux questions posées en langage naturel, en comprenant la question et en localisant la réponse pertinente dans les données textuelles.
- **Traduction Automatique :** Les systèmes NLP cherchent à traduire automatiquement le texte d'une langue à une autre, en capturant les nuances et les subtilités du langage.
- **Résumé Automatique :** Le NLP peut générer automatiquement des résumés succincts d'articles ou de textes longs en identifiant les informations essentielles.
- **Analyse des sentiments :** Les systèmes NLP évaluent les opinions et les émotions exprimées dans le texte, ce qui trouve des applications dans l'analyse des réseaux sociaux, les critiques de produits, etc.

En résumé, le NLP vise à créer des modèles de langage intelligents et adaptatifs qui permettent aux machines de traiter le langage humain de manière fluide et cohérente, ouvrant ainsi la voie à une communication plus naturelle entre l'homme et la machine. Cette compréhension du NLP sert de fondement essentiel à la mise en œuvre réussie de notre ChatBot pour l'analyse d'articles scientifiques et la génération de résumés.

2.2.2 Techniques et Approches du NLP

Les techniques et approches du Traitement Automatique du Langage Naturel (NLP) couvrent un éventail diversifié de méthodes visant à accomplir les objectifs complexes de compréhension et de génération de langage naturel. Ces techniques sont le fruit d'une convergence entre les avancées en intelligence artificielle, en linguistique et en traitement de données massives. Voici un aperçu des principales techniques et approches du NLP :

- **Tokenization :** Cette étape consiste à diviser le texte en unités plus petites appelées "tokens", qui peuvent être des mots ou des caractères. Cela permet de traiter le texte de manière granulaire.
- **Analyse Syntaxique :** L'analyse syntaxique vise à comprendre la structure grammaticale d'une phrase en identifiant les rôles des mots (sujet, verbe, objet, etc.) et les relations entre eux.
- **Traitement des Stop Words :** Les stop words sont des mots courants tels que "et", "le", "de" qui n'apportent pas de sens significatif à l'analyse et sont souvent retirés pour réduire le bruit..

- **Modèles de Langage :** Ces modèles statistiques apprennent les probabilités d'apparition des mots dans un contexte donné, ce qui est essentiel pour comprendre la signification des phrases.
- **Réseaux de Neurones :**Lesréseaux de neurones, notamment les réseaux récurrents (RNN) et les transformers, ont révolutionné le NLP en permettant aux modèles d'apprendre des relations complexes entre les mots.
- **Apprentissage Supervisé et Non Supervisé :**L'apprentissage supervisé implique l'entraînement de modèles sur des données étiquetées, tandis que l'apprentissage non supervisé découvre des modèles à partir de données non étiquetées.
- **Méthodes d'Entraînement :**L'entraînement des modèles NLP peut utiliser des techniques telles que le Fine-Tuning pour adapter des modèles pré-entraînés à des tâches spécifiques.
- **Approches de Représentation de Mot :** Des approches telles que Word Embeddings et Word2Vec capturent les représentations sémantiques des mots pour mieux comprendre leur signification.
- **Traitement de Langage Naturel Conversationnel :** L'approche Conversationnelle NLP vise à rendre les interactions homme-machine plus naturelles en comprenant les dialogues complexes et les nuances du langage parlé.

Ces techniques et approches, parmi d'autres, constituent les briques essentielles qui permettent aux systèmes NLP de comprendre, de traiter et de générer du langage naturel avec une précision croissante. Leur compréhension est cruciale pour l'implémentation réussie de notre ChatBot pour l'analyse d'articles scientifiques et la génération de résumés.

2.3 BERT :ModèlesBaséssurlesTransformers pour le Question Answering

Au cœur de l'avancée révolutionnaire du Traitement Automatique du Langage Naturel (NLP), se trouvent les modèles basés sur les transformers, avec BERT (Bidirectional Encoder Representations from Transformers) en tête. Cette section se penche sur la conception innovante de BERT et explore comment ce modèle a transformé le paysage du Question Answering. Nous commencerons par un aperçu de BERT, avant de plonger dans son application captivante dans la tâche de Question Answering. La compréhension de BERT est cruciale pour saisir la puissance de ce modèle dans notre projet de ChatBot.

2.3.1 Aperçu de BERT :

2.3.1.1 Conception des Transformers :

Les transformers ont révolutionné le domaine du Traitement Automatique du Langage Naturel (NLP) en introduisant une architecture novatrice qui a permis des avancées significatives

dans la compréhension contextuelle des mots. La conception des transformers repose sur une structure à multiples couches qui permet une analyse approfondie des relations entre les mots d'un texte. Contrairement aux architectures précédentes, les transformers ne suivent pas une séquence linéaire, ce qui leur permet de capturer les interactions contextuelles à la fois vers l'avant et vers l'arrière.

Cette conception innovante repose sur deux composants principaux : les mécanismes d'atten tion et les couches de feedforward. Les mécanismes d'attention permettent au modèle de pondérer l'importance de chaque mot en fonction de son contexte, créant ainsi des représentations plus riches et nuancées. Les couches de feedforward suivent les mécanismes d'attention pour effectuer des transformations non linéaires, ajoutant de la complexité à la compréhension du modèle.

La flexibilité des transformers a permis de résoudre plusieurs problèmes de NLP, notamment la traduction automatique, la génération de texte et bien sûr, le Question Answering. BERT (Bidirectional Encoder Representations from Transformers), dont nous allons discuter en détail, est un modèle emblématique basé sur les transformers qui a propulsé les performances du NLP à de nouveaux sommets.

2.3.1.2 Architecture et Mécanisme de BERT :

L'architecture des transformers, sur laquelle repose BERT (Bidirectional Encoder Representations from Transformers), a radicalement redéfini la manière dont les modèles NLP captent les dépendances contextuelles dans le langage. Cette architecture révolutionnaire est basée sur une série d'étapes fondamentales qui permettent la compréhension bidirectionnelle du texte.

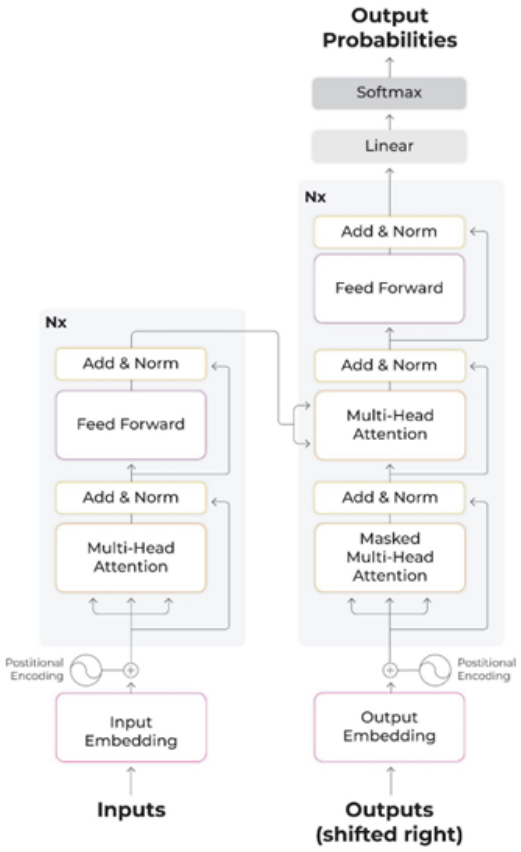

FIGURE 2.1 : L'architecture des transformers

Le processus commence par l'introduction des Embeddings d'Entrée pour les tokens d'entrée. Ces embeddings sont ensuite combinés avec l'Encodage Positionnel pour fournir au modèle une compréhension de la position relative des tokens dans la séquence.

Les étapes clés qui suivent sont les blocs Multi-Head Attention, qui capturent les relations contextuelles en évaluant les relations entre les tokens. Les étapes Add & Norm sont utilisées pour ajouter et normaliser les sorties des différentes parties du modèle, garantissant la stabilité du flux d'information.

Le bloc Feed Forward est responsable de l'ajout d'une couche de traitement non linéaire pour chaque token. Les Add & Norm qui suivent s'assurent que les sorties sont à nouveau normalisées. Cette structure en cascade de transformations Multi-Head Attention, Feed Forward, et Add & Normestrépétée plusieurs fois dans l'architecture pour capter des relations complexes.

Enfin, l'Encodage Positionnel est à nouveau appliqué pour les sorties, et les embeddings sont utilisés pour prédire les Probabilités d'apparition des tokens à l'aide d'une fonction Softmax. Les Outputs obtenus sont alors utilisés pour les prédictions et les tâches spécifiques.

En résumé, la conception des transformers présente une structure en couches qui permet une analyse contextuelle approfondie du langage grâce à l'attention multi-tête et à l'encodage positionnel. Cette architecture sous-tend la puissance de BERT dans la compréhension de phrases complexes et son application au Question Answering.

2.3.1.3 Pré-Entraînement et Fine-Tuning :

L'une des caractéristiques clés de BERT réside dans sa méthodologie de pré-entraînement et de fine-tuning, qui lui permet de capturer une compréhension profonde et contextuelle du langage. Cette approche novatrice repose sur l'utilisation de vastes corpus de texte non étiqueté pour pré-entraîner le modèle, suivi d'un fine-tuning sur des tâches spécifiques.

Pré-Entraînement : Lors de la phase de pré-entraînement, BERT est exposé à d'énormes quantités de texte non étiqueté. Il apprend à prédire des mots masqués dans une phrase en se basant sur le contexte environnant. Cette tâche, appelée "Masked Language Model", pousse BERT à saisir les relations sémantiques et syntaxiques entre les mots. Le modèle intègre ainsi une compréhension profonde du langage, ce qui en fait une base solide pour des tâches spécifiques ultérieures.

Fine-Tuning : Après la pré-entraînement, BERT est adapté à des tâches spécifiques via le fine-tuning. Cette étape consiste à entraîner le modèle sur des données étiquetées pour la tâche souhaitée, comme le Question Answering dans notre contexte. Le fine-tuning affine les représentations apprises lors du pré-entraînement pour qu'elles soient spécifiques à la tâche cible. Grâce à cette méthode, BERT peut être appliqué à différentes tâches en utilisant des ensembles de données relativement petits pour le fine-tuning.

En combinant pré-entraînement et fine-tuning, BERT parvient à capturer des informations de haut niveau sur la syntaxe, la sémantique et le contexte dans le langage, ce qui en fait un choix puissant pour les tâches de NLP, y compris le Question Answering. Dans notre projet, BERT est un élément clé qui permet à notre ChatBot d'extraire efficacement des réponses précises à partir d'articles scientifiques complexes

2.3.2 Application de BERT au Question Answering et à la Génération de Résumés

2.3.2.1 Fondements du Question Answering

Le Question Answering (QA) est une tâche majeure du Traitement Automatique du Langage Naturel (NLP) qui vise à développer des systèmes capables de comprendre et de répondre aux questions posées en langage naturel. La complexité du QA réside dans la nécessité de comprendre le contexte, de saisir les relations entre les phrases et les entités mentionnées, et de fournir une réponse précise et cohérente.

L'une des approches traditionnelles pour le QA consiste à utiliser des modèles basés sur des règles ou des modèles d'apprentissage automatique pour extraire des réponses à partir du texte. Cependant, ces approches sont souvent limitées en termes de compréhension contextuelle et de capacité à gérer des questions complexes.

C'est là que BERT entre en jeu, en apportant une nouvelle dimension au QA. En capturant des dépendances contextuelles complexes et en comprenant les nuances du langage, BERT a le potentiel de fournir des réponses plus précises et pertinentes. Son architecture de transformer bidirectionnelle permet de contextualiser chaque mot en fonction du texte environnant, ce qui améliore considérablement sa capacité à extraire des réponses avec un haut degré de pertinence.

Dans notre projet, nous exploitons la puissance de BERT pour résoudre la tâche exigeante du Question Answering dans le contexte de l'analyse d'articles scientifiques. En tirant parti de la compréhension contextuelle avancée de BERT, notre ChatBot peut extraire des réponses pertinentes et informatives à partir d'articles scientifiques, contribuant ainsi à une meilleure compréhension et à l'exploration des connaissances scientifiques.

2.3.2.2 Utilisation de BERT pour le Question Answering

L'utilisation de BERT pour le Question Answering (QA) marque une avancée significative dans la capacité des modèles de NLP à comprendre et à répondre aux questions en langage naturel. BERT excelle dans cette tâche en raison de sa conception innovante, qui lui permet de capturer des relations contextuelles entre les mots et de saisir les nuances du langage.

Lorsqu'il s'agit de QA, BERT prend en compte le contexte global du texte pour interpréter la signification des questions et extraire des réponses pertinentes. Contrairement aux méthodes traditionnelles qui ne considèrent que des mots individuels, BERT analyse les interactions complexes entre les mots, ce qui permet de fournir des réponses plus précises et cohérentes.

Le processus de Question Answering avec BERT se déroule en deux phases majeures. Tout d'abord, BERT est pré-entraîné sur de vastes corpus de texte non étiqueté, ce qui lui permet de développer une compréhension profonde du langage. Ensuite, le modèle est affiné pour la tâche spécifique de QA à l'aide de données étiquetées.

Lorsqu'une question est posée, BERT analyse la question et le contexte pour identifier les parties pertinentes du texte susceptibles de contenir la réponse. En utilisant ses représentations de mots riches en informations, BERT évalue les correspondances et extrait la réponse la plus appropriée.

Dans notre projet, BERT est déployé pour extraire des réponses précises à partir d'articles scientifiques dans le domaine de l'Intelligence Artificielle. Cette utilisation de BERT pour le QA renforce notre ChatBot en lui permettant d'interagir avec les utilisateurs de manière intelligente et informative, en fournissant des réponses basées sur une compréhension approfondie du contenu scientifique.

2.4 Résumé Automatique de Texte

Le résumé automatique de texte est une composante cruciale du traitement automatique du langage naturel (NLP) qui vise à extraire les informations essentielles d'un texte long et complexe tout en maintenant la cohérence et la signification. Dans cette section, nous explorerons les approches extractives et abstractives du résumé automatique, ainsi que l'utilisation de modèles de langage, en mettant l'accent sur notre utilisation spécifique du modèle "BERT Extractive Summarizer" pour générer des résumés de haute qualité à partir d'articles scientifiques.

2.4.1 Approches Extractives et Abstractives

Les approches du résumé automatique se déclinent en deux principales catégories : les approches extractives et les approches abstractives. Chacune de ces approches présente des avan-

tages et des défis distincts dans la création de résumés automatiques.

2.4.1.1 Approches Extractives

Les approches extractives consistent à sélectionner directement des phrases, des paragraphes ou des mots clés du texte source pour former le résumé. Ces approches tirent parti de l'information déjà présente dans le texte et garantissent une fidélité au contenu original. En identifiant les parties les plus informatives du texte, les approches extractives produisent souvent des résumés cohérents et pertinents. Cependant, elles peuvent parfois souffrir d'une certaine rigidité, car elles dépendent entièrement du contenu existant sans possibilité de reformulation ou de réorganisation

2.4.1.2 Approches Abstractives

Les approches abstractives, en revanche, visent à générer un résumé en reformulant et en ré organisant les informations du texte source. Ces approches ont l'avantage de la flexibilité, car elles peuvent produire des résumés plus fluides et naturels. En permettant la reformulation, elles peuvent également résoudre le problème de la redondance et présenter des informations dans un ordre logique. Cependant, les approches abstractives peuvent également être confrontées à des défis majeurs, tels que la préservation de la cohérence et la garantie de la fidélité au contenu d'origine.

2.4.2 Utilisation de Modèles de Langage pour le Résumé

L'utilisation de modèles de langage dans le processus de résumé automatique a considérablement évolué grâce à des avancées telles que BERT (Bidirectional Encoder Representations from Transformers). Les modèles de langage, en particulier ceux basés sur l'architecture des transformers, ont apporté une nouvelle perspective en permettant de capturer des dépendances contextuelles complexes et de saisir les nuances du langage naturel.

2.4.2.1 Avantages des Modèles de Langage

Les modèles de langage, par leur nature, peuvent analyser le contexte global d'un texte et identifier les relations sémantiques entre les mots et les phrases. Cette caractéristique les rend extrêmement précieux pour la génération de résumés. Ils peuvent extraire les parties les plus informatives et significatives d'un texte, tout en assurant la cohérence et la pertinence dans le résumé final. Les modèles de langage peuvent également résoudre des problèmes tels que la redondance et la répétition dans le résumé, en optimisant la sélection des informations.

2.4.2.2 BERT Extractive Summarizer

Dans notre projet, nous avons choisi d'utiliser le modèle "BERT Extractive Summarizer" pour générer des résumés automatiques à partir d'articles scientifiques en IA. Ce choix est motivé parla puissance de BERT à capturer des relations contextuelles et à extraire des informations pertinentes. Le modèle analyse le texte source pour identifier les parties les plus

informatives, puis les organise de manière à créer un résumé cohérent et significatif. Le modèle "BERT Extractive Summarizer" se révèle particulièrement efficace pour le résumé extractif, où l'objectif est de conserver fidèlement les informations du texte source. En tirant parti de la représentation riche et nuancée de BERT, notre ChatBot peut offrir aux utilisateurs des résumés concis et informatifs qui facilitent la compréhension et l'exploration des articles scientifiques en IA.

2.5 État de l'Art dans le Question Answering et le Résumé Automatique

L'état de l'art dans les domaines du Question Answering (QA) et du Résumé Automatique a connu des avancées significatives grâce aux progrès constants en matière de modèles de traitement automatique du langage naturel (NLP). Ces domaines étant étroitement liés, leur évolution a été alimentée par des modèles de langage avancés et des techniques novatrices.

2.5.1 Évolution du Question Answering

L'évolution du QA a été marquée par des approches de plus en plus sophistiquées, allant des méthodes traditionnelles aux modèles basés sur l'apprentissage profond. Les premières méthodes se concentraient sur la correspondance de motifs et la recherche de réponses dans le texte source.Cependant, avec l'avènement des modèles de langage pré-entraînés comme BERT, le QA a connu une révolution. Les modèles de type transformer ont permis de saisir les relations complexes entre les mots et ont introduit des capacités de compréhension contextuelle et d'inférence, améliorant considérablement les performances du QA.

2.5.2 Avancées dans le Résumé Automatique

De manière similaire, les avancées dans le domaine du Résumé Automatique ont été stimulées par l'émergence de modèles de langage pré-entraînés et de réseaux de neurones profonds. Les méthodes extractives traditionnelles ont été améliorées avec l'intégration de techniques de sélection et de pondération de phrases. En parallèle, les approches abstractives ont bénéficié de l'utilisation de modèles de génération de langage, permettant de créer des résumés plus fluides et naturels. Dans l'ensemble, l'utilisation de modèles pré-entraînés comme BERT a été un catalyseur majeur dans l'amélioration des performances du QA et du Résumé Automatique. Ces modèles ont permis une meilleure compréhension du langage naturel et une représentation plus riche des informations, ce qui a eu un impact significatif sur la qualité des réponses générées et des résumés créés.

Dans notre projet, nous explorons ces avancées en utilisant BERT pour à la fois répondre aux questions des utilisateurs dans le contexte d'articles scientifiques et générer des résumés extractifs de haute qualité. Cette approche état de l'art renforce la pertinence et l'efficacité de notre ChatBot dans le traitement et la synthèse d'informations complexes en IA.

2.6 Conclusion

Dans ce chapitre, nous avons exploré le cadre théorique et la revue de littérature sous-tendant notre projet axé sur le développement d'un ChatBot pour l'analyse d'articles scientifiques en NLP. En examinant les fondements du Traitement Automatique du Langage Naturel (NLP) et en mettant l'accent sur les modèles basés sur les Transformers tels que BERT, nous avons discuté de leur application au Question Answering (QA) et à la génération de résumés. Nous avons également abordé les approches extractives et abstractives du Résumé Automatique, tout en soulignant l'évolution du domaine du QA et du Résumé Automatique grâce à ces avancées. Ce chapitre fournit un socle conceptuel pour comprendre l'intégration de BERT et d'autres techniques dans notre ChatBot, préparant ainsi le terrain pour les phases ultérieures du projet.

Chapitre 3

Analyse et conception du projet

3.1 Introduction

Ce chapitre se consacre à l'analyse approfondie et à la conception détaillée du projet "NLP pour l'analyse d'articles scientifiques en IA". En vue de concrétiser cette solution, plusieurs étapestechniques clés ont été entreprises pour répondre aux besoins fonctionnels et aux objectifs du projet. Chaque étape, de l'extraction du texte à partir de sources variées à la création d'une interface utilisateur interactive, sera détaillée, expliquant comment les différentes technologies et méthodes ont été combinées pour réaliser un ChatBot capable d'extraire des données, de générer des résumés et de répondre aux questions des utilisateurs. Ce chapitre mettra en lumière le processus d'analyse des composants existants, suivi d'une présentation détaillée de la conception et de l'implémentation de chacune des étapes techniques.

3.2 Analyse de l'existant

L'analyse de l'existant, première étape cruciale de notre projet, nous a conduit à évaluer différentes approches pour le Question Answering (QA) et la génération de résumés. Parmi ces approches, nous avons porté une attention particulière à BERT (Bidirectional Encoder Representations from Transformers), une technologie révolutionnaire basée sur les Transformers. Lors de notre analyse comparative, nous avons également examiné l'approche traditionnelle de la recherche de mots-clés pour le Question Answering (QA) et la génération de résumés. Cette méthode consiste à identifier et extraire des mots ou des phrases clés du texte source pour construire la réponse ou le résumé. Cependant, cette approche présente des limitations importantes. Elle peut être inefficace pour gérer la complexité et la richesse sémantique des articles scientifiques, où les informations importantes peuvent être disséminées dans tout le texte.

Contrairement à cette approche, BERT offre une compréhension plus profonde du contexte et de la sémantique des textes. Grâce à son architecture basée sur les Transformers, BERT est capable de saisir les relations entre les mots et de prendre en compte la structure globale d'un document. Par conséquent, il est capable de fournir des réponses précises en considérant la signification du texte dans son ensemble, ce qui fait défaut à la recherche de mots-clés.

Un autre avantage majeur de BERT réside dans son processus d'entraînement préalable sur de vastes quantités de données textuelles diverses. Cela lui permet de capturer des connaissances et des associations linguistiques subtiles, rendant ses prédictions plus contextuellement pertinentes et précises. De plus, l'utilisation de BERT élimine le besoin de conception manuelle d'heuristiques ou de règles spécifiques pour chaque tâche, contrairement à certaines approches traditionnelles.

En somme, notre analyse comparative a mis en évidence que BERT surpasse les méthodes traditionnelles en offrant une compréhension contextuelle supérieure, une meilleure gestion de la complexité des articles scientifiques et une performance globalement plus élevée pour le Question Answering et la génération de résumés. C'est pourquoi nous avons choisi BERT comme pierre angulaire de notre ChatBot pour l'analyse d'articles scientifiques et la création de résumés.

3.3 Extraction de Texte depuis des Sources Variées et traitement de texte

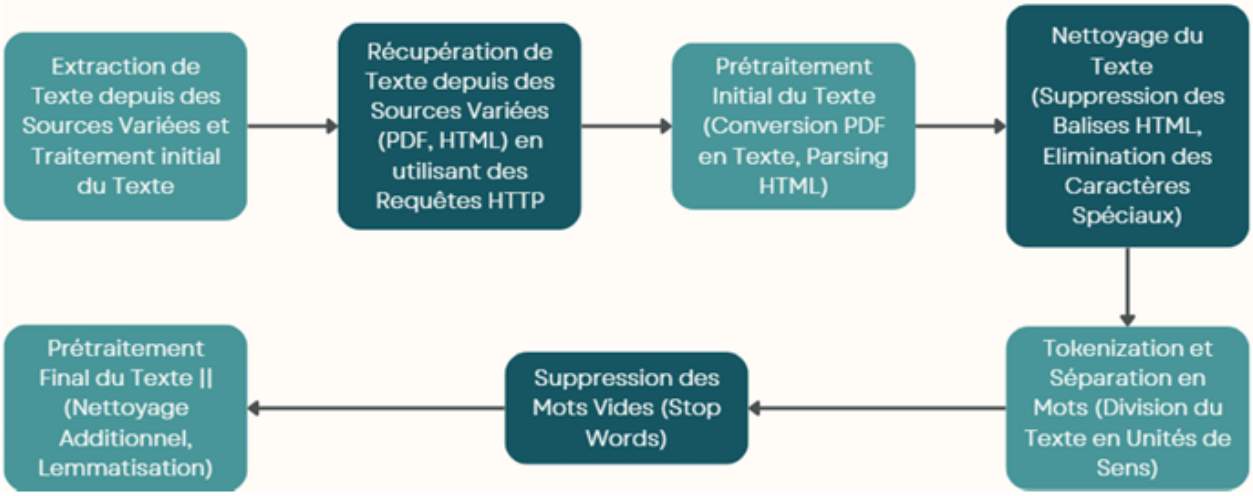

FIGURE 3.1 : Extraction de Texte depuis des Sources Variées et traitement de texte

Dans ce schéma détaillé, nous commençons par extraire le texte à partir de sources variées telles que des articles scientifiques en format PDF ou HTML. Cette extraction est réalisée en utilisant des requêtes HTTP pour récupérer le contenu des liens fournis. Ensuite, le prétraitement initial du texte est effectué pour convertir le contenu PDF en texte brut et parser le HTML pour extraire le texte.

Suite au prétraitement initial, le nettoyage du texte est effectué pour supprimer les balises HTML et éliminer les caractères spéciaux indésirables. Ensuite, le texte est soumis à la tokenization, où il est divisé en unités de sens telles que les mots. Les mots vides (stopwords) sont ensuite éliminés pour réduire le bruit et concentrer l'analyse sur les mots clés pertinents.

Enfin, le texte subit un prétraitement final, qui peut inclure des étapes supplémentaires de nettoyage et de normalisation telles que la lemmatisation pour ramener les mots à leur forme de base.

Cette conception détaillée de l'étape d'extraction et de traitement initial du texte permet de préparer efficacement les données textuelles pour les étapes ultérieures du ChatBot, comme le Question Answering et la génération de résumés.

3.4 Question Answering par BERT

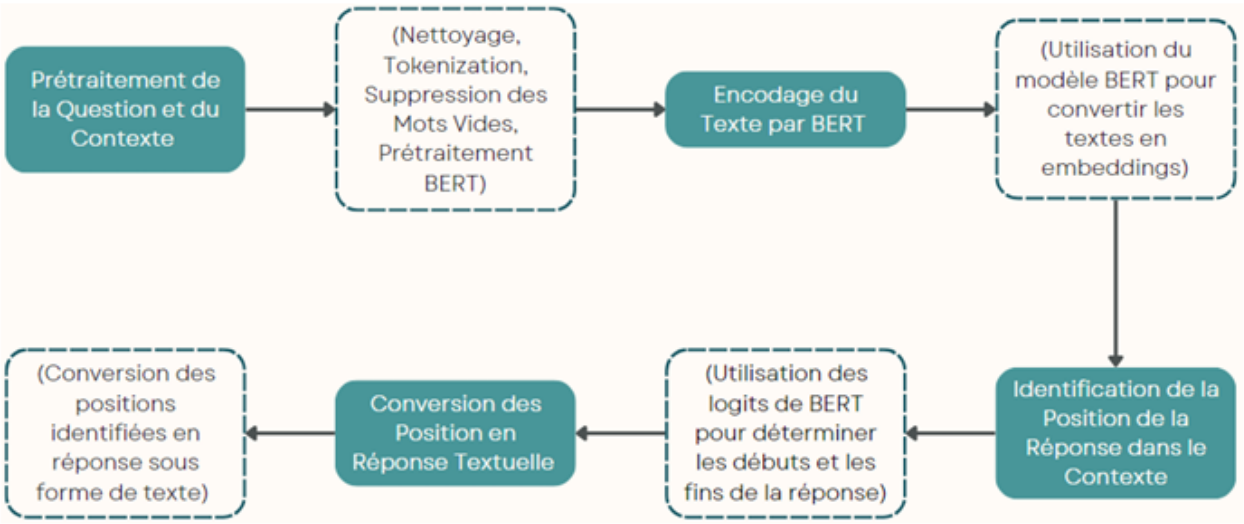

FIGURE 3.2 : Question Answering par BERT

Dans ce schéma détaillé, nous commençons par le prétraitement de la question posée par l'utilisateur ainsi que du contexte dans lequel la question est posée. Ce prétraitement inclut le nettoyage, la tokenization, la suppression des mots vides et un prétraitement spécifique à BERT pour préparer les données en vue de l'encodage par le modèle BERT.

Ensuite, le texte prétraité est encodé par le modèle BERT pour obtenir des embeddings de mots et de phrases. Ces embeddings capturent les informations sémantiques du texte. Après l'encodage, les logits générés par BERT sont utilisés pour identifier la position probable de la réponse dans le contexte. Les positions de début et de fin de la réponse sont déterminées en utilisant les informations fournies par BERT.

Enfin, les positions identifiées sont converties en une réponse textuelle qui est extraite du contexte original. Cette réponse est ensuite présentée à l'utilisateur comme la réponse générée par le ChatBot en réponse à la question posée.

Cette conception détaillée de l'étape de Question Answering par BERT illustre comment le modèle BERT est utilisé pour répondre aux questions posées par les utilisateurs en analysant le contexte et en identifiant les positions de réponse appropriées.

3.5 Génération de Résumés par BERT Summarizer

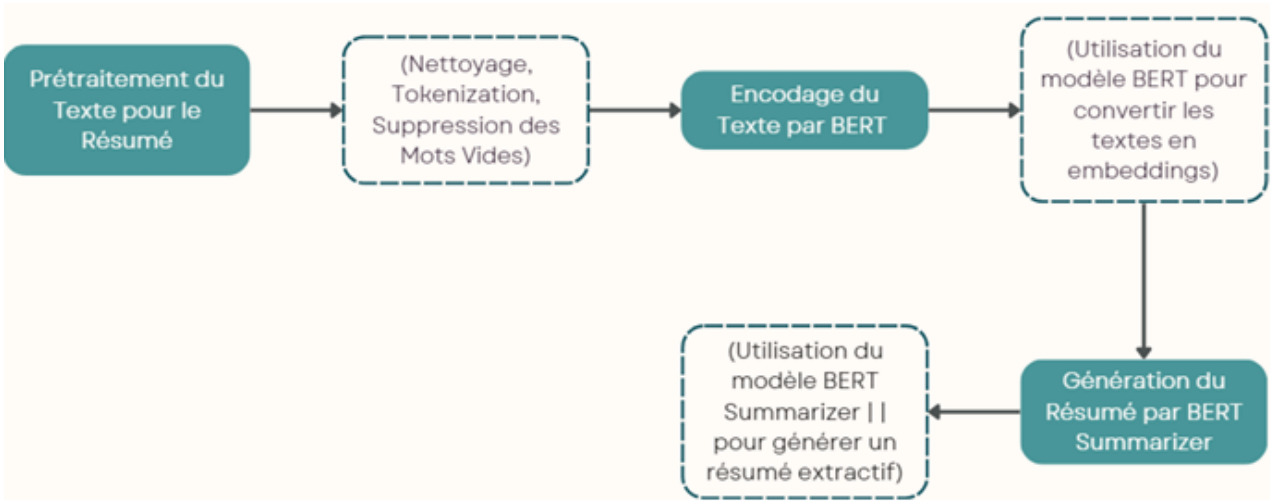

FIGURE 3.3 : Génération de Résumés par BERT Summarizer

Dans ce schéma détaillé, nous commençons par le prétraitement du texte original afin de le préparer pour la génération de résumés. Cela implique le nettoyage du texte, la tokenization et la suppression des mots vides.

Ensuite, le texte prétraité est encodé par le modèle BERT pour obtenir des embeddings de mots et de phrases. Ces embeddings représentent les informations sémantiques du texte.

Une fois que le texte est encodé, le modèle BERT Summarizer est utilisé pour générer le résumé. Le modèle sélectionne les phrases les plus importantes du texte original en fonction des embeddings et de leur pertinence pour le contenu global. Le résumé généré est donc extractif, c'est-à-dire qu'il est composé de phrases extraites du texte d'origine.

Cette conception détaillée de l'étape de Génération de Résumés par BERT Summarizer met en évidence comment le modèle BERT Summarizer utilise les embeddings du modèle BERT pour identifier et extraire les phrases clés du texte original, générant ainsi un résumé représentatif et concis du contenu.

3.6 Interface Utilisateur avec Streamlit

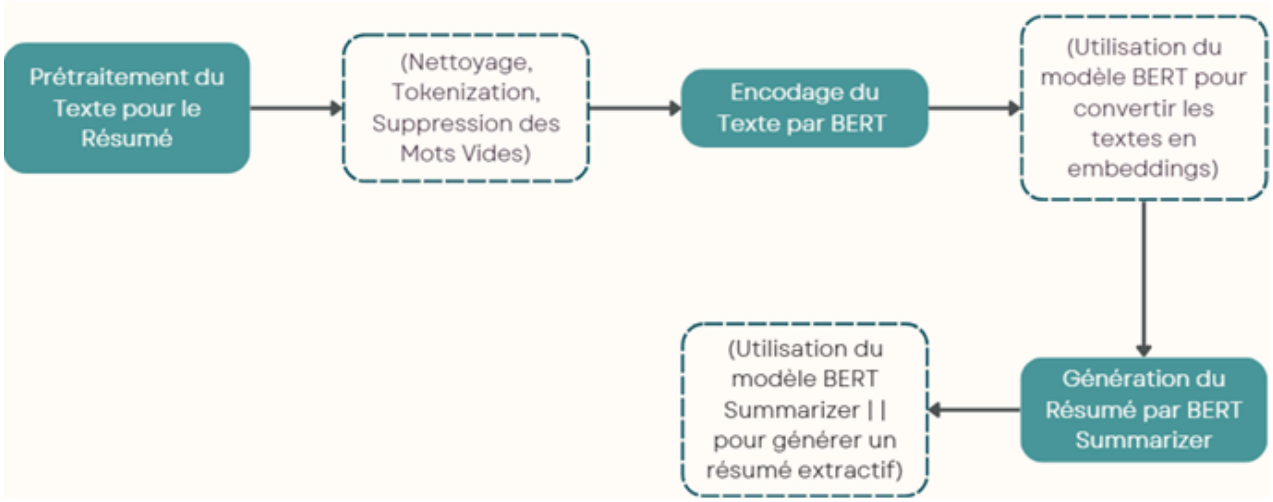

FIGURE 3.4 : Interface Utilisateur avec Streamlit

Dans ce schéma détaillé, nous commençons par créer l'interface utilisateur en utilisant Streamlit. Streamlit permet de créer facilement une interface interactive en Python sans nécessiter une expertise en développement d'interface.

Ensuite, nous intégrons les différentes fonctionnalités du projet dans l'interface utilisateur. Cela comprend l'intégration des fonctionnalités d'extraction de texte à partir d'articles, de question réponse par BERT et de génération de résumés par BERT Summarizer.

Une fois les fonctionnalités intégrées, l'interface utilisateur permet à l'utilisateur d'interagir avec le projet. L'utilisateur peut entrer un lien vers un article, extraire le texte de cet article, poser des questions et obtenir des réponses en utilisant le modèle de question-réponse par BERT, ainsi que générer des résumés à partir du texte de l'article.

Cette conception détaillée de l'étape d'Interface Utilisateur avec Streamlit met en évidence comment Streamlit est utilisé pour créer une interface interactive qui offre des fonctionnalités d'extraction, de question-réponse et de génération de résumés, permettant ainsi à l'utilisateur d'interagir facilement avec le projet.

3.7 Conception du fonctionnement Global du ChatBot

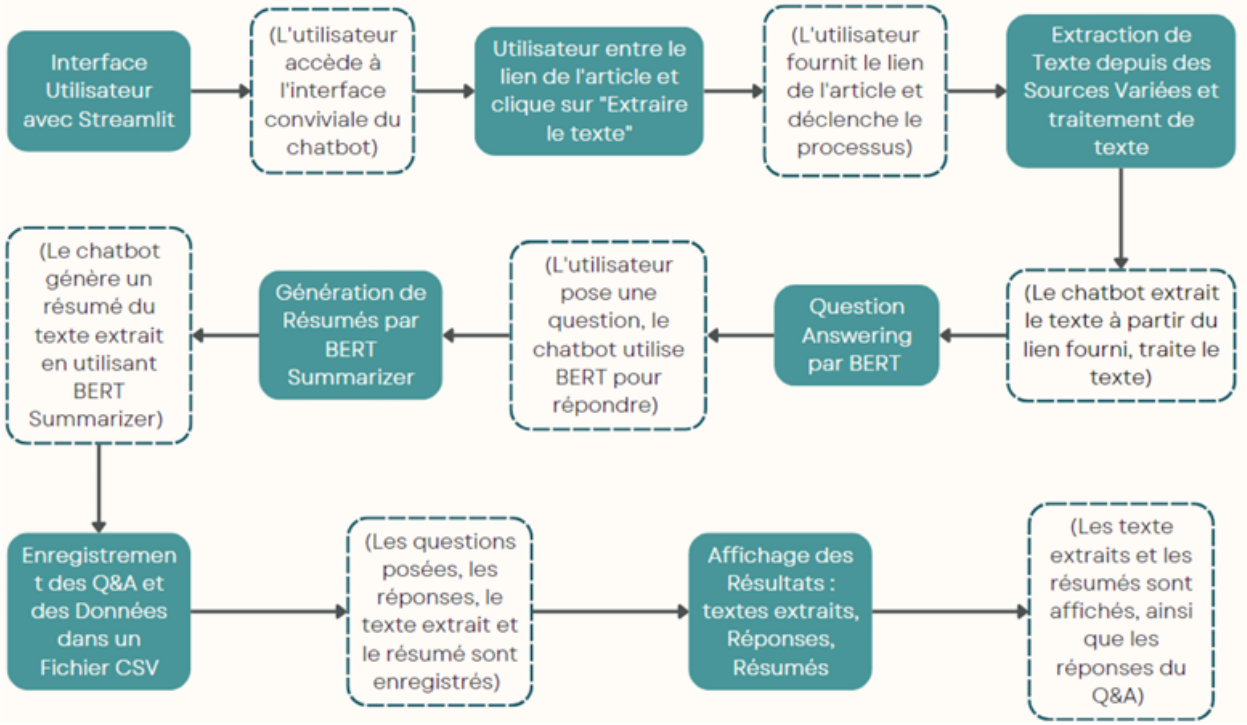

FIGURE 3.5 : Conception Global Du projet

Le schéma décrit le fonctionnement global du ChatBot dans diverses étapes. Tout commence par l'interface utilisateur conviviale via Streamlit, où l'utilisateur entre le lien d'un article.En cliquant sur "Extraire le texte", le processus démarre. Le ChatBot procède à l'extraction du texte à partir du lien fourni, le soumet à un traitement de texte pour une meilleure compréhension. Ensuite, l'utilisateur peut poser des questions, et le ChatBot utilise la puissance de BERT pour fournir des réponses précises. De plus, le ChatBot génère un résumé du texte extrait à l'aide de BERT Summarizer.

Pour conserver une trace de ces interactions, les questions posées, les réponses du ChatBot, le texte extrait et les résumés sont enregistrés dans un fichier CSV. Cette étape d'enregistrement permet de stocker les données pour référence future. Finalement, les résultats, comprenant le texte extrait, le résumé et les QA, sont affichés à l'utilisateur pour qu'il puisse facilement consulter les informations pertinentes extraites du contenu source.

3.8 Conclusion

Ce chapitre a détaillé les étapes clés de la réalisation du ChatBot. Nous avons comparé les méthodes traditionnelles avec l'approche BERT pour le Question Answering et la Génération de Résumés, mettant en évidence les avantages de BERT. Le fonctionnement du ChatBot, de l'extraction de texte à l'interface utilisateur, a été conçu pour fournir des réponses pertinentes et des résumés précis. L'intégration de BERT et des techniques NLP crée un outil performant pour

extraire des informations et répondre aux questions, tout en générant des résumés significatifs à partir de divers contenus.

Chapitre 4

La Réalisation de la solution

4.1 Introduction

Le chapitre suivant de ce rapport se penche sur la concrétisation de la solution envisagée. Nous allons explorer en détail les outils, les technologies et les étapes mises en œuvre pour réaliser le chatbot d'analyse d'articles scientifiques en utilisant les méthodes de traitement automatique du langage naturel (NLP). Cette partie du rapport mettra en lumière les librairies et frameworks essentiels qui ont été utilisés pour créer et mettre en place la solution, ainsi que les différentes étapes du processus de mise en œuvre, allant de l'extraction de texte à partir de diverses sources jusqu'à la génération de résumés pertinents et la mise en place d'une interface utilisateur intuitive. Chaque étape sera détaillée, illustrée par des exemples et des résultats obtenus, offrant ainsi une vue complète de la réalisation de ce projet.

4.2 Outils de Réalisation

La mise en œuvre réussie du chatbot d'analyse d'articles scientifiques repose sur l'utilisation d'outils technologiques appropriés. Ce chapitre se focalise sur les principaux outils qui ont été employés pour développer la solution NLP. En particulier, nous explorerons le rôle central du langage de programmation Python, ainsi que les librairies et frameworks spécifiques qui ont été choisis pour tirer parti des fonctionnalités avancées du traitement automatique du langage naturel (NLP).

4.2.1 Python : Langage de Programmation Principal

Python a été le langage de programmation fondamental utilisé tout au long du projet. Sa simplicité, sa flexibilité et sa vaste communauté de développeurs en font un choix idéal pour la mise en œuvre de solutions NLP complexes. L'écosystème Python offre une multitude de bibliothèques et d'outils dédiés au traitement du langage naturel, facilitant ainsi le développement d'algorithmes sophistiqués et la manipulation de données textuelles.

FIGURE 4.1 : Logo- Python

4.2.2 Librairies et Frameworks Utilisés

Pour exploiter les capacités du traitement automatique du langage naturel, diverses librairies et frameworks ont été intégrés dans le processus de développement. Parmi les éléments clés figurent :

- **Hugging Face Transformers :** Une bibliothèque populaire pour l'utilisation de modèles de langage pré-entraînés, y compris BERT, pour diverses tâches de traitement de texte comme la question-réponse et la génération de résumés.
- **Beautiful Soup :** Une bibliothèque Python pour extraire des informations à partir de pages web HTML et XML.
- **Requests :** Une bibliothèque Python pour effectuer des requêtes HTTP vers des sites web et récupérer le contenu.
- **Streamlit :** Un framework pour créer facilement des interfaces utilisateur interactives pour les projets de data science et de machine learning.
- **Pandas :** Une bibliothèque Python pour la manipulation et l'analyse de données.
- **NLTK (Natural Language Toolkit) :** Une bibliothèque pour le traitement du langage naturel, offrant diverses fonctionnalités pour l'analyse de texte.
- **Scikit-learn :** Une bibliothèque Python pour l'apprentissage automatique et l'exploration des données.
- **PyTorch ou TensorFlow :** Des frameworks d'apprentissage automatique populaires pour l'entraînement et l'utilisation de modèles de machine learning, y compris les modèles de traitement de texte.

Ces outils ont été soigneusement sélectionnés pour leurs performances, leur adaptabilité et leur contribution à la réalisation de notre solution de chatbot NLP.

4.3 Mise en Œuvre

Dans cette section dédiée à la mise en œuvre de notre solution, nous plongeons dans les détails pratiques de la réalisation du chatbot basé sur BERT. Nous allons explorer étape par étape les différentes phases de mise en place, depuis la sélection des sources d'articles pour nos tests jusqu'à la gestion des données. Nous examinerons de près les résultats de l'extraction de texte, de la génération de résumés et des réponses aux questions, en illustrant concrètement les réalisations obtenues pour différents types d'articles. Cette démarche nous permettra de mieux appréhender l'efficacité et les performances de notre solution à chaque étape.

Pour démarrer, nous vous présentons tout d'abord la page d'accueil de l'interface utilisateur,où le chatbot vous invite à entrer un lien d'article pour commencer le processus.

FIGURE 4.2 : La page d'accueil de l'interface utilisateur

4.3.1 Sélection des Sources d'Articles pour les Tests

Dans le cadre de la mise en œuvre et de la validation de notre solution de chatbot NLP, la sélection des sources d'articles pour les tests revêt une importance cruciale. L'objectif est de choisir des articles représentatifs et diversifiés, permettant ainsi d'évaluer l'efficacité de notre chatbot dans la résolution de problèmes complexes de question-réponse et de génération de résumés. Pour cela, deux sources ont été soigneusement choisies afin de garantir une couverture adéquate des sujets et des structures de texte.

La première source sélectionnée est un site web au contenu riche en informations : "BERT (language model)". Ce site, accessible via le lien https ://en.wikipedia.org/wiki/BERT_(language_model), offre un article complet sur le modèle de langage BERT et son architecture. L'extrait de l'article utilisé contient des informations pertinentes sur la naissance, le développement et les caractéristiques de BERT, notamment sa structure basée sur l'architecture des transformers.

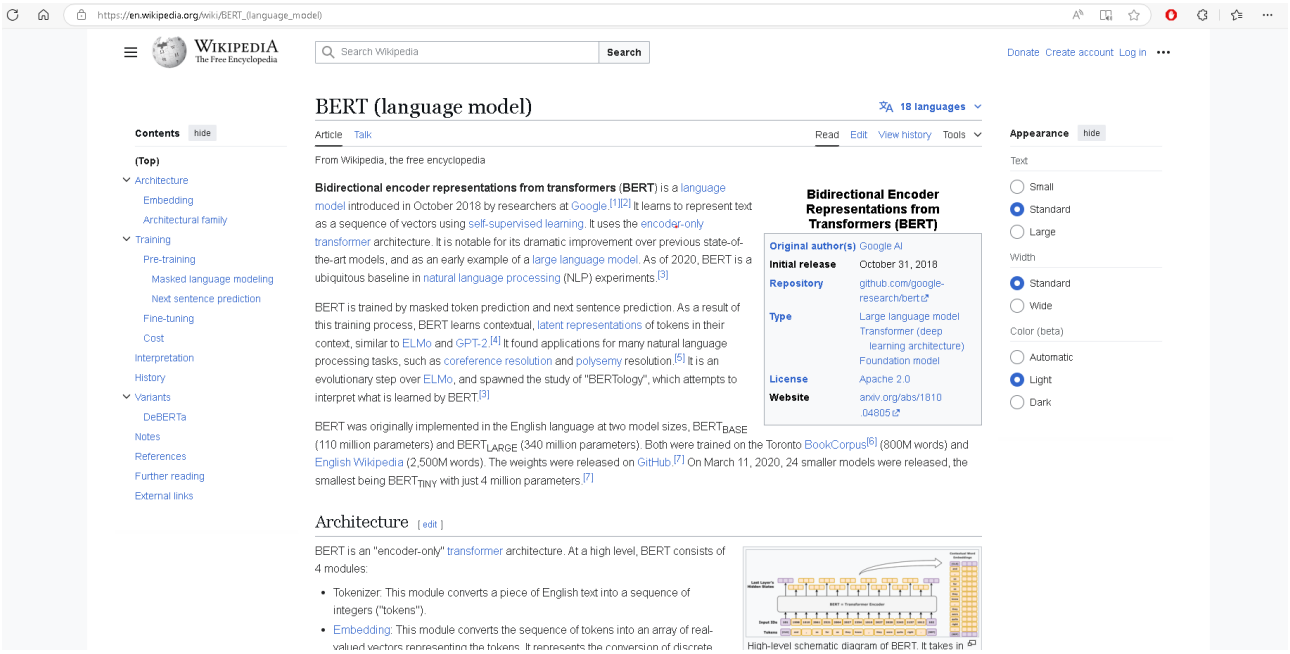

FIGURE 4.3 : Capture d'écran du site de test HTML

La seconde source choisie pour les tests provient d'un document PDF : "Utilizing Bidirectional Encoder Representations from Transformers for Answer Selection". Ce document, disponible à l'adresse https ://arxiv.org/pdf/2011.07208.pdf, présente une étude approfondie sur l'utilisation des représentations d'encodeurs bidirectionnels à partir des transformers pour la tâche de sélection de réponses. L'extrait du PDF extrait couvre l'introduction et les objectifs de l'étude, mettant en évidence l'efficacité de l'approche adoptée pour la résolution de tâches spécifiques de traitement du langage naturel.

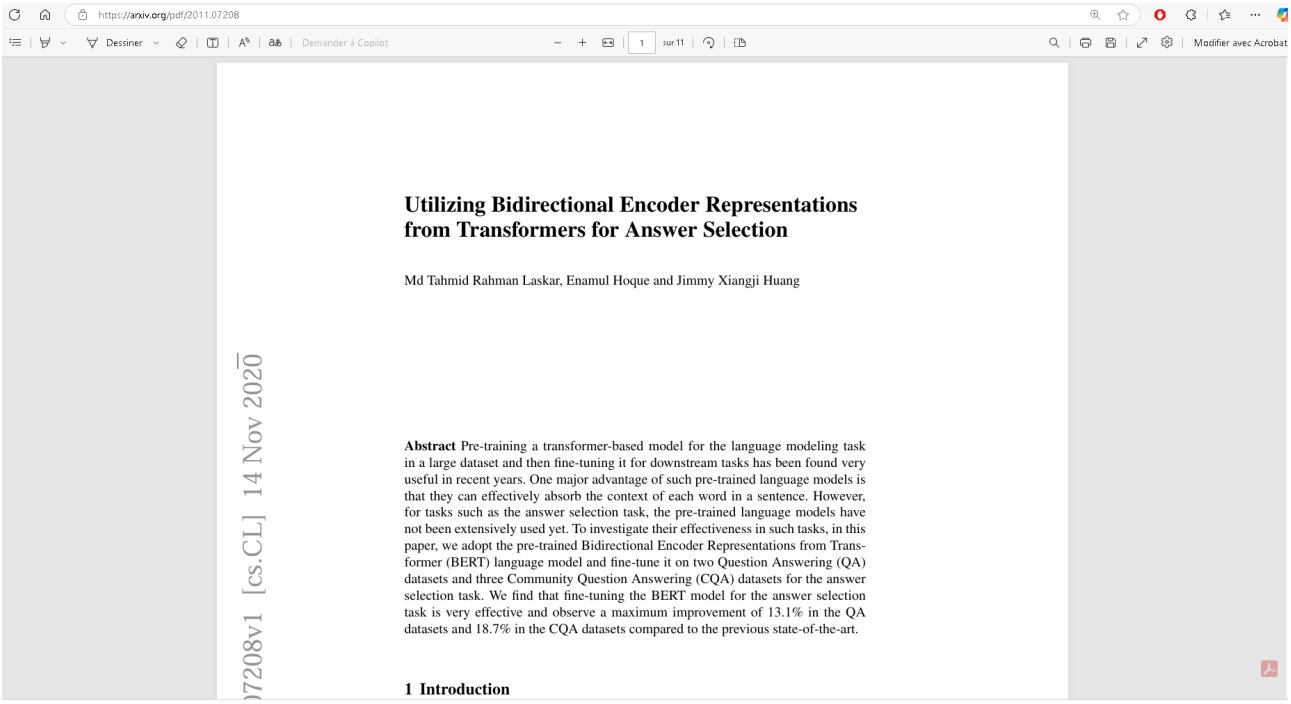

FIGURE 4.4 : Capture d'écran du site de test PDF

Ces deux sources ont été choisies pour leur pertinence et leur contribution à la diversification

des tests. Le site web fournit un aperçu global du modèle BERT et de son architecture, tandis que le document PDF se concentre sur l'utilisation de techniques de transformers pour la sélection de réponses. Ces sources sont représentatives de la variété des articles scientifiques que notre chatbot serait amené à traiter dans le monde réel. Dans les sections suivantes, nous détaillerons les résultats obtenus à partir de l'extraction, du question-réponse et de la génération de résumés pour ces sources sélectionnées.

4.3.2 Résultats de l'Extraction de Texte et de la génération de Résumés

Dans cette section, nous présentons les résultats de l'étape d'extraction de texte et de génération de résumés à partir des sources d'articles sélectionnées. Il est important de noter que, en raison des limitations de capacité de traitement du modèle BERT, l'extraction de texte est réalisée en prenant en compte la quantité maximale de mots que le modèle peut traiter efficacement. Par conséquent, le texte complet des articles n'est pas extrait, mais une portion significative qui respecte les contraintes de BERT.

4.3.2.1 Pour l'article du site HTML

Cette capture d'écran illustre le texte extrait ainsi que le résumé généré pour l'article issu du site HTML.

FIGURE 4.5 : Résultats de l'Extraction de Texte et de la génération de Résumés pour le site HTML

Voici un apperçu sur le texte extrait : <Bidirectional encoder representations from transformers (BERT) is a language model introduced in October 2018 by researchers at Google.[1][2] It learns to represent text as a sequence of vectors using self-supervised learning. It uses the encoder-only transformer architecture. It is notable for its dramatic improvement over previous state-of-the-art models, and as an early example of a large language model. As of 2020[update], BERT is a ubiquitous baseline in natural language processing (NLP) experiments.[3]... >

Voici le résumé généré : <Bidirectional encoder representations from transformers (BERT) is a language model introduced in October 2018 by researchers at Google.[1][2] It learns to represent text as a sequence of vectors using self-supervised learning. It uses the encoder-only

transformer architecture. The other one, BERTLARGE, is similar, just larger. By varying these two numbers, one obtains an entire family of BERT models.[9] For BERT The notation for encoder stack is written as L/H. For example, BERTBASE is written as 12L/768H, BERT-LARGE as 24L/1024H, and BERTTINY as 2L/128H. BERT was pre-trained simultaneously on two tasks.[10] In masked language modeling, 15% of tokens would be randomly selected for masked-prediction task, and the training objective was to predict the masked token given its context. It would first be divided into tokens like "my1 dog2 is3 cute4". Then a random token in the sentence would be picked. BERT is meant as a general pretrained model for various applications in natural language processing. BERT considers the words surrounding the target word fine from the left and right side. However it comes at a cost : due to encoder-only architecture lacking a decoder, BERT can't be prompted and can't generate text, while bidirectional models in general do not work effectively without the right side, thus being difficult to prompt. As an illustrative example, if one wishes to use BERT to continue a sentence fragment "Today, I went to", then naively one would mask out all the tokens as "Today, I went to [MASK] [MASK] [MASK] ... [MASK] ." The small model aims to fool the large model. Its key idea is to treat the positional and token encodings separately throughout the attention mechanism..>

4.3.2.2 Pour l'article du site PDF

Cette capture d'écran illustre le texte extrait ainsi que le résumé généré pour l'article issu du site PDF.

FIGURE 4.6 : Résultats de l'Extraction de Texte et de la génération de Résumés pour un article pdf

4.3.3 Résultats des Réponses aux Questions

Dans cette section, nous présentons les résultats obtenus lors de la génération de réponses aux questions posées à l'aide du modèle de Question Answering BERT. Il est important de noter que les réponses sont générées en fonction de la compréhension du modèle à partir du texte extrait et du contexte de la question. Les captures d'écran ci-dessous montrent les exemples de questions posées, les réponses générées par le modèle BERT, ainsi que la partie du texte extrait qui contient la réponse sélectionnée par le curseur.

4.3.3.1 Pour l'article du site HTML

Les captures d'écran du Questions Answering pour l'article extrait depuis l'HTML

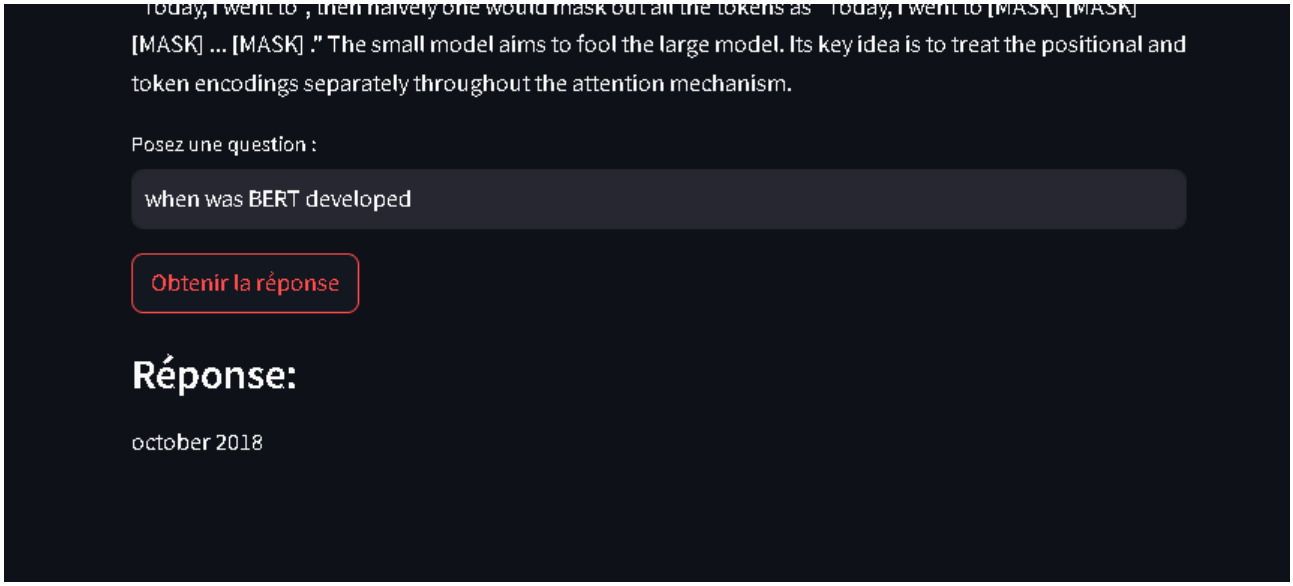

FIGURE 4.7 : question 1 - page HTML

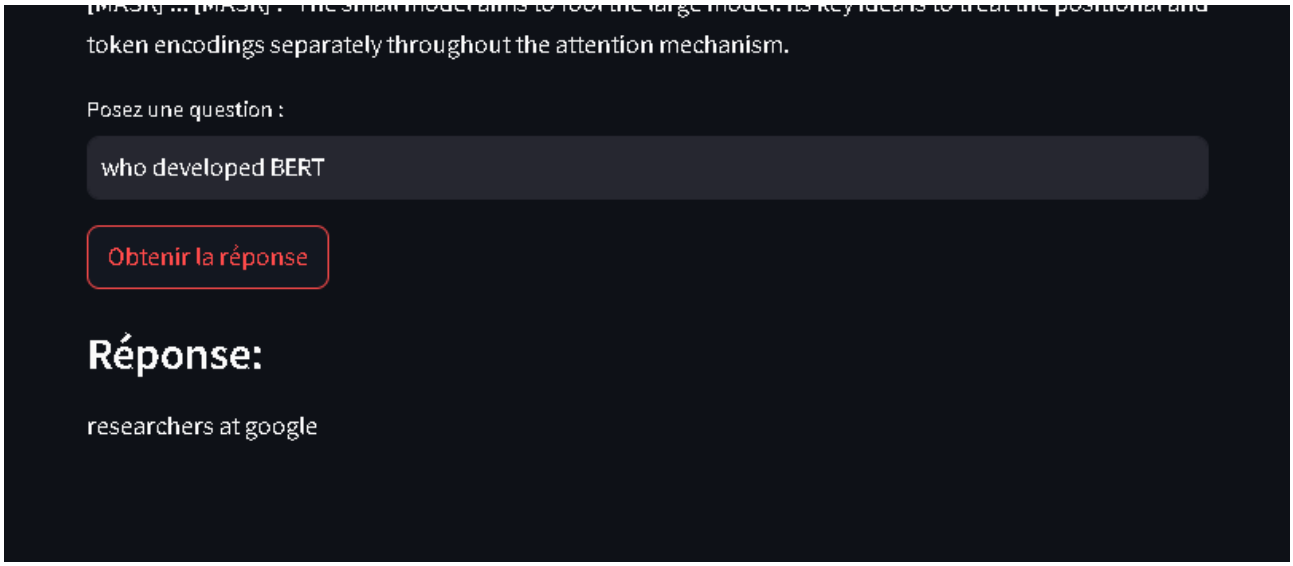

FIGURE 4.8 : question 2 - page HTML

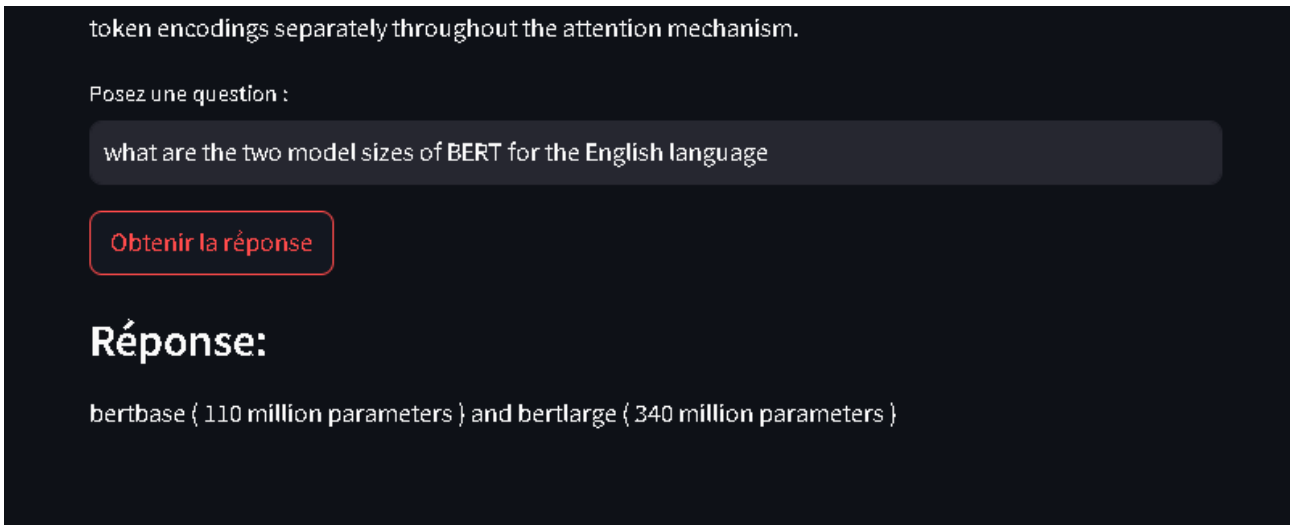

FIGURE 4.9 : question 3 - page HTML

4.3.3.2 Pour l'article du site PDF

Les captures d'écran du Questions Answering pour l'article extrait depuis le PDF

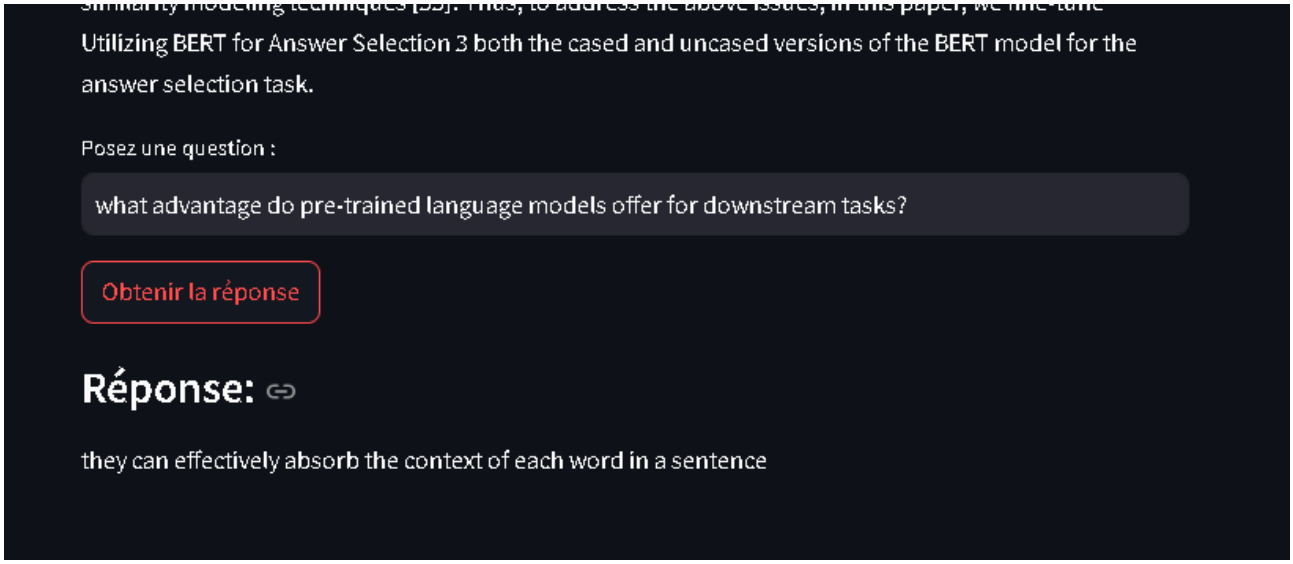

FIGURE 4.10 : question 1 - article PDF

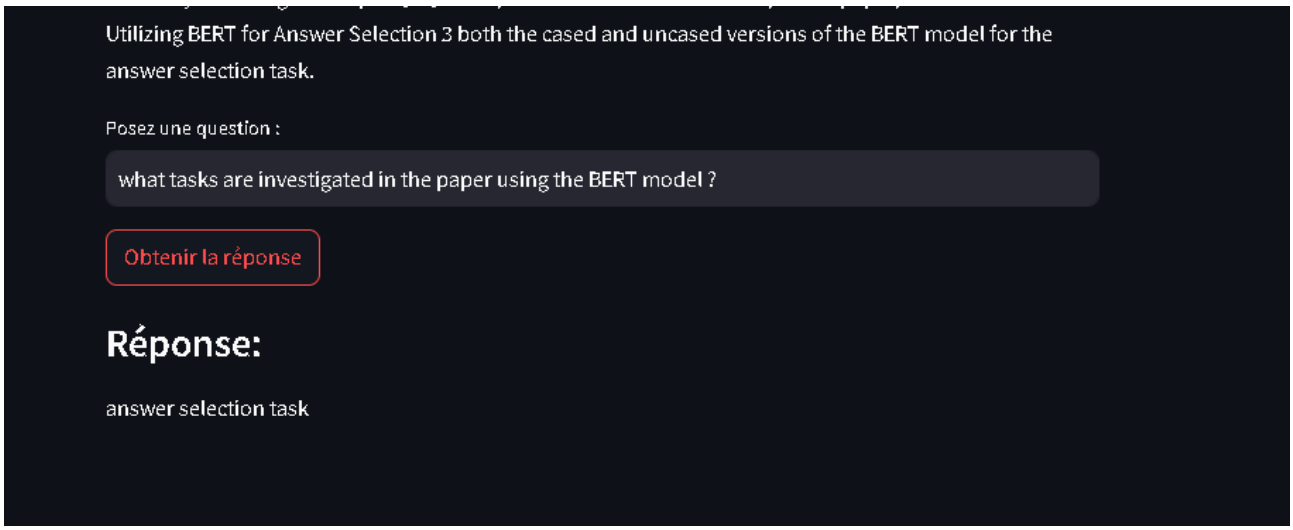

FIGURE 4.11 : question 2 - article PDF

Conclusion

En conclusion, ce rapport a exploré en détail le développement d'un chatbot basé sur le modèle BERT pour le Question Answering et la génération de résumés automatiques. Nous avons présenté le contexte général du projet, en mettant en avant les objectifs de recherche et les différentes étapes de planification. En nous appuyant sur un cadre théorique solide, nous avons discuté des fondements du NLP, notamment de la puissance des modèles BERT et de leur utilisation pour le Question Answering et la génération de résumés.

Le chapitre d'analyse et de conception a exposé la méthodologie suivie, mettant en avant les différentes étapes techniques telles que l'extraction de texte, le Question Answering par BERT et lagénération de résumés avec BERT Summarizer. Ces étapes ont été illustrées par des schémas concrets et des exemples de résultats obtenus à partir d'articles variés.

La réalisation de la solution a été décrite en détail, couvrant l'utilisation d'outils tels que Python et diverses bibliothèques pour la mise en œuvre. Les étapes de mise en œuvre ont été accompagnées de captures d'écran illustrant les résultats d'extraction de texte, les résumés générés et les réponses aux questions.

En somme, ce projet a démontré que les modèles basés sur les Transformers, tels que BERT, ont une portée significative dans l'amélioration de la compréhension automatique du langage naturel. Ils ouvrent des opportunités prometteuses pour les applications de Question Answering et de génération de résumés, offrant ainsi une solution pratique et efficace pour extraire des informations essentielles à partir de documents textuels.

Bibliographie

[1] Bert (language model).
https://en.wikipedia.org/wiki/BERT_(language_model).

[2] Génération de résumés par abstraction.
https://papyrus.bib.umontreal.ca/xmlui/handle/1866/10335.

[3] Lstm, transformers, gpt, bert : guide des principales techniques en nlp.
https://france.devoteam.com/paroles-dexperts/lstm-transformers-gpt-bert-guide-des-pri

[4] Traitement naturel du langage : tout savoir sur le natural language processing.
https://www.lebigdata.fr/traitement-naturel-du-langage-nlp-definition#:
~:text=Quelles%20sont%20les%20diff%C3%A9rentes%20techniques,d'en%20d%C3%
A9chiffrer%20le%20sens.

[5] Transformers
https://towardsdatascience.com/transformers-89034557de14

[6] Utilizing bidirectional encoder representations from transformers for answer selection.
https://arxiv.org/pdf/2011.07208.pdf

[7] Bert extractive summarizer
https://pypi.org/project/bert-extractive-summarizer/

[8] Extractive summarization with bert extractive summarizer.
https://www.holisticseo.digital/python-seo/summarize/