UNIVERSITÉ DE BLIDA 1

Faculté des sciences

Département d'informatique

MÉMOIRE DE MASTER

En Informatique

Option : Ingénierie Des Logiciels

Un système pour l'indexation automatique des vidéos

Réalisé par :
Bounadjar Faiçal
Kartali Mohamed

Supervisé par
Dr. BACHA Siham

Devant les membres du jury

Président : Dr. RIALI ISHAK Blida 1

Examinatrice : Dr. FERDI IMENE Blida 1

juin 2023

Résumé

Les plateformes de partage de vidéos ont révolutionné l'accès à l'information et les interactions en ligne. Cependant, en Algérie, la recherche de vidéos pertinentes et de haute qualité pose problème en raison de la concurrence des contenus étrangers et de l'absence de stratégies de recherche et de promotion efficaces, malgré la richesse du patrimoine culturel local.

Notre projet propose une plateforme de partage de vidéos nationale en Algérie, dotée d'un modèle de traitement du langage naturel. Cette plateforme permettra d'indexer automatiquement les vidéos et offrira aux utilisateurs la possibilité de publier et de partager leurs propres contenus locaux, ainsi qu'une expérience optimisée pour la recherche et la découverte de contenus vidéo locaux.

Mots clés: Plateforme de partage de vidéos nationale, modèle de traitement du langage naturel, indexer automatiquement les vidéos, publier et partager le contenu local, expérience optimisée pour la recherche.

Abstract

Video-sharing platforms have revolutionized access to information and online interactions. However, in Algeria, finding relevant and high-quality videos is problematic due to competition from foreign content and the lack of effective search and promotion strategies, despite the rich local cultural heritage.

Our project proposes a national video sharing platform in Algeria, equipped with a natural language processing model. This platform will automatically index videos and provide users with the ability to publish and share their own local content, as well as an optimized experience for searching and discovering local video content.

Keywords: National video sharing platform, natural language processing model, automatically index videos, publish and share local content, search-optimized experience.

الملخص

أحدثت منصات مشاركة الفيديو ثورة في الوصول إلى المعلومات والتفاعلات عبر الإنترنت. ومع ذلك، في الجزائر، يمثل العثور على مقاطع فيديو ذات صلة وعالية الجودة مشكلة بسبب المنافسة من المحتوى الأجنبي وعدم وجود استراتيجيات فعالة للبحث والترويج، على الرغم من التراث الثقافي المحلي الغني.

يقترح مشروعنا منصة وطنية لمشاركة الفيديو في الجزائر، مزودة بنموذج معالجة اللغة الطبيعية. ستقوم هذه المنصة تلقائيًا بفهرسة مقاطع الفيديو وتزويد المستخدمين بالقدرة على نشر ومشاركة المحتوى المحلي الخاص بهم، بالإضافة إلى تجربة محسّنة للبحث عن محتوى الفيديو المحلي واكتشافه.

الكلمات المفتاحية: منصة مشاركة الفيديو الوطنية، نموذج معالجة اللغة الطبيعية، فهرسة مقاطع الفيديو تلقائيًا، نشر المحتوى المحلي ومشاركته، تجربة البحث المحسّنة.

Remerciements

Avant tout, nous remercions Allah, le Tout-Puissant, de nous avoir donné le courage et la volonté nécessaires pour mener à bien cette humble réalisation.

Nos premiers remerciements sont adressés à notre promotrice, Dr. BACHA Siham, qui nous a proposé ce thème et nous a fait confiance malgré notre connaissance limitée dans le domaine du traitement du langage naturel. Nous la remercions pour son aide précieuse, ses judicieux conseils et le temps qu'elle nous a accordé.

Nos remerciements vont également aux enseignants Mr. BALA Mahfoud et Mr. HEMINA Karim, ainsi qu'à nos collègues et amis, notamment BRAHIM Aimen et NACHEF Abdelkrim, pour leur disponibilité, leurs conseils et leur capacité d'écoute et d'échange d'informations.

Nos remerciements vont également aux membres du jury qui ont généreusement accepté d'évaluer notre travail et de le enrichir par leurs précieuses propositions.

Nous tenons également à adresser nos remerciements chaleureux au corps professoral et administratif de l'Université de BLIDA, dont la contribution a grandement contribué à notre réussite dans nos études universitaires.

Dédicaces

Je dédie ce modeste travail à :

À mes chers parents, aucun mot de dédicace ne saurait exprimer pleinement mes sincères sentiments envers vous.

Je vous suis infiniment reconnaissant pour votre patience inépuisable, votre soutien constant et votre assistance précieuse. C'est avec un profond amour et un respect profond que je rends hommage à vos grands sacrifices.

À ma tante qui m'a soutenu tout au long de ma vie et de mon parcours académique, en particulier en m'enseignant la langue française jusqu'à la dernière minute et en m'aidant à corriger mes erreurs d'écriture, à ma sœur, à mon frère qui m'encouragent toujours à me surpasser.

À mon binôme Faiçal, à tous mes amis qui ont été toujours à mes côtés surtout dans les moments difficiles et à tous mes collègues du club scientifique ITC.

Dédicaces

Ce humble travail est dédié à:

mes chers parents, en commençant par mon père que Dieu lui fasse miséricorde, à ma mère que Dieu la protège. Aucune dédicace ne peut véritablement rendre justice à l'ampleur de mes sincères sentiments envers vous.

Mon père, que Dieu lui fasse miséricorde, je suis infiniment reconnaissant pour ta patience, ton amour et tout ce que tu as fait pour moi , en termes d'expériences, de valeurs, de soutien constant et d'une aide inestimable. J'aurais souhaité que tu sois à mes côtés aujourd'hui, en cet instant, pour voir les fruits de ton assistance paternelle, ami et père tu étais le meilleur. J'aurai aimé que tu sois présent en cet événement tant attendu mais Dieu a décidé autrement. J'espère que tu es fier de moi de la ou tu es. Merci pour tout, mon père.

À ma chère mère qui m'a accompagnée tout au long de ma vie et de mon parcours académique, je tiens à t'exprimer ma gratitude infinie pour ta patience inépuisable, ta soutien constant et votre précieuse assistance. Tes sacrifices extraordinaires qui sont porteurs d'un amour profond et d'un respect incommensurable de ma part. Je te remercie du fond du cœur pour tout ce que tu as fait.

À mon frère "Hani" , sa femme et leur fille "melek" ,qui m'encouragent toujours à me surpasser, que dieux vous protège.

Je souhaite exprimer ma profonde gratitude envers mon binôme Mohamed, qui a été une source constante d'encouragement grâce à son travail, ses idées et sa patience inébranlable.

Un grand dédicace également à tous mes collègues du club scientifique ITC, que je considère comme ma deuxième famille. Ils ont toujours été présents à mes côtés, dans les moments de joie et de peine. Un énorme merci à eux. Je souhaite également remercier tous mes amis qui m'ont apporté leur aide, même avec un simple mot, dans ma vie.

Tables des matières

Liste des tableaux

Liste des figures

Introduction Générale

Les plateformes de partage de vidéos comme YouTube sont devenues très populaires dans le monde [10], ils ont considérablement changé la façon dont les gens accèdent aux informations et interagissent en ligne, offrant aux utilisateurs la possibilité de télécharger et de partager facilement des vidéos avec une audience mondiale. En Algérie, une grande majorité de la population est connectée à internet, YouTube est particulièrement prisé [11]. Cependant, la recherche de vidéos pertinentes et de haute qualité pour les utilisateurs algériens peut être difficile, surtout lorsque la plateforme ne dispose pas d'un système de recherche adéquat. Cette situation est confrontée par des problématiques par exemple : la visibilité limitée du contenu algérien, la concurrence des contenus étrangers et l'absence de stratégies de promotion et de diffusion efficaces.

Malgré la richesse et la diversité du patrimoine culturel algérien, le contenu en ligne reste limité et peu visible, en raison de cette concurrence des contenus étrangers. Pour pallier cette situation, des créateurs et producteurs algériens ont néanmoins pris l'initiative de créer des chaînes YouTube pour rendre accessible les archives filmiques algériennes, couvrant tous les aspects de la culture. Des initiatives offrent des enregistrements historiques sur l'Algérie et son patrimoine culturel, mais malgré cela, ce contenu reste souvent relégué au second plan sur les plateformes de partage de vidéos. Les algorithmes de recommandation et de mise en avant des contenus favorisent souvent les produits étrangers.

Face à ces défis, on propose de résoudre les difficultés actuelles liées à la recherche et à la découverte de contenu sur les plateformes de partage de vidéos existantes en Algérie en développant une plateforme de partage de vidéos nationale. Cette plateforme sera dotée d'un modèle de traitement du langage naturel permettant d'indexer automatiquement les vidéos, tout en offrant aux utilisateurs algériens la possibilité de publier et de partager leurs propres contenus locaux. La mise en place d'une telle plateforme impliquera de relever des défis techniques et de ressources, tels que la collecte et l'analyse de données linguistiques et culturelles locales pour entraîner les modèles de traitement du langage naturel.

Afin de mieux appréhender l'étendue de la problématique et de détailler les différentes

étapes nécessaires pour le développement de notre solution, ce document est structuré en plusieurs chapitres. Dans le premier chapitre, nous examinerons le contexte et l'importance des plateformes de partage de vidéos dans le paysage numérique mondial et les spécificités liées à l'Algérie ainsi les différentes problématiques auxquelles ces plateformes sont confrontées. Le deuxième chapitre abordera l'analyse et la conception de notre solution, en détaillant chaque module du système. Enfin, le troisième chapitre présentera la réalisation de notre projet, y compris l'environnement et les outils de travail utilisés, les résultats obtenus et l'interface graphique développée.

Notre objectif est de créer une plateforme de partage de vidéos nationale en Algérie, qui intègre un modèle de traitement du langage naturel. En combinant une analyse approfondie des problématiques actuelles et cette solution innovante, nous visons à offrir aux utilisateurs algériens une expérience optimisée pour la recherche et la découverte de contenu vidéo local. Nous souhaitons également mettre en valeur le riche patrimoine culturel de l'Algérie à travers cette plateforme, offrant ainsi aux utilisateurs une meilleure accessibilité et une plus grande visibilité pour les contenus nationaux.

Chapitre 1

L'état de l'art

1.1 Contexte et problématique

Les plateformes de partage de vidéos sont devenues extrêmement populaires dans le monde entier, avec des milliards d'heures de vidéo visionnées chaque jour, permettant aux utilisateurs de télécharger et de partager facilement des vidéos avec une audience mondiale. L'émergence de ces plateformes a considérablement changé la façon dont les gens accèdent aux informations et interagissent en ligne.

L'Algérie compte aujourd'hui plus de 32,09 millions d'internautes, ce qui représente près de 70,9% de la population du pays [18]. Les plateformes de partage de vidéos comme YouTube sont très populaires dans ce pays, mais la plupart d'entre elles ne prennent pas en compte les spécificités culturelles des pays et régions, même pour trouver des vidéos pertinentes et de haute qualité peut être un défi, surtout lorsque la plateforme n'a pas d'indexation ou de système de recherche adéquat. Selon les études, on distingue 5 problématiques principales qui rendent la recherche des vidéos spécifiques pour les utilisateurs algériens difficile et frustrante.

1.1.1 Visibilité limitée et manque de valorisation

La première problématique est la visibilité limitée du contenu algérien sur YouTube et sur internet en général, ce qui rend difficile la promotion et la valorisation du patrimoine culturel du pays, notamment la littérature, les documentaires et la musique, entre autres. Malheureusement, malgré la richesse et la diversité du patrimoine culturel algérien il y a un manque important du contenu disponible en ligne, et ceux qui existent ont souvent

une visibilité limitée auprès de la population algérienne. Cette situation s'explique par le manque de stratégies efficaces de promotion et de diffusion du contenu, mais aussi par l'importante concurrence des contenus étrangers sur les plateformes en ligne.

Pourtant, il y a des efforts louables de la part de certains créateurs et producteurs algériens pour rendre ce contenu plus visible. Prenons par exemple la mise en ligne d'archives vidéo et de documentaires sur la culture algérienne comme 'Tourathy.elmajala' et 'Rabnass Archive Algeria2.0' (figure 1.1) entre autre, parlant de cette dernière chaîne qui est active sur Youtube, il était difficile de la trouver, elle a pour objectif la préservation et de rendre accessibles les archives filmiques algériens, couvrant tous les aspects de la culture, de la musique, des actualités, des feuilletons algériens depuis l'invention de l'image animée dans les années 1890. Il est important de mentionner également la chaîne de "Nassira Belloula" (figure 1.2), une journaliste et écrivaine algérienne, ses écrits, allant des romans aux poésies en passant par les essais, récits et nouvelles. Elle est nominée pour plusieurs prix et a été finaliste du Prix Mohamed Dib et Yamina Mechakra. Malgré la richesse de leurs contenus, la population algerienne a du mal a les trouver parmi les 51 millions chaînes présentes sur Youtube en se basant sur le site web [57].

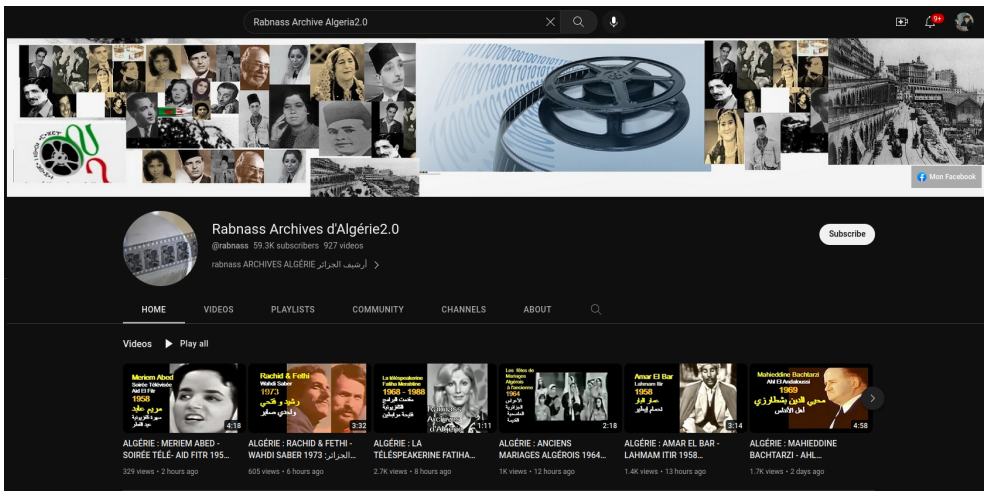


Figure 1.1: "Rabnass Archive Algeria2.0". chaîne culturel algerienne compte 59,3k abonnés.

Ces dernières permettent de rendre disponible tout enregistrement historique concernant l'Algérie et son héritage culturel pour informer, faciliter la recherche et éduquer un public international , et aussi pour honorer les célébrités algériennes du passé et du présent.

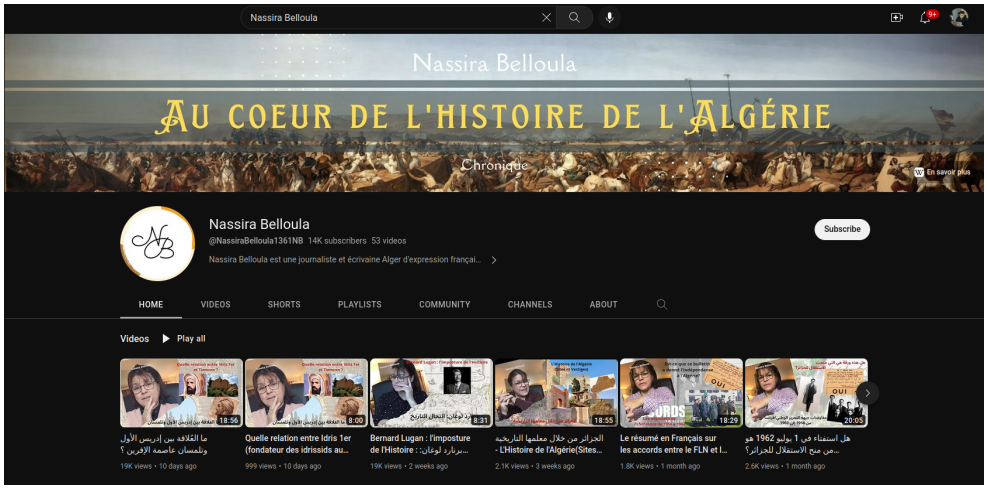


Figure 1.2: "Nassira Belloula".chaîne d'une journaliste et écrivaine algérienne

Un autre exemple de chaîne YouTube algérienne est la chaîne "Koussay Zaoui" (figure 1.3). Elle traite principalement de l'histoire de Tlemcen et de l'Andalousie, ainsi que d'autres sujets liés à l'histoire et à la culture de l'Algérie. Dans ses vidéos, Koussay Zaoui raconte l'histoire de Tlemcen et de l'Andalousie à travers des documentaires, des visites de lieux historiques et des discussions sur divers sujets relatifs à l'histoire et à la culture de l'Algérie. Il couvre également des événements historiques tels que la conquête musulmane de l'Espagne et le défi des Kouloughlis de Tlemcen.

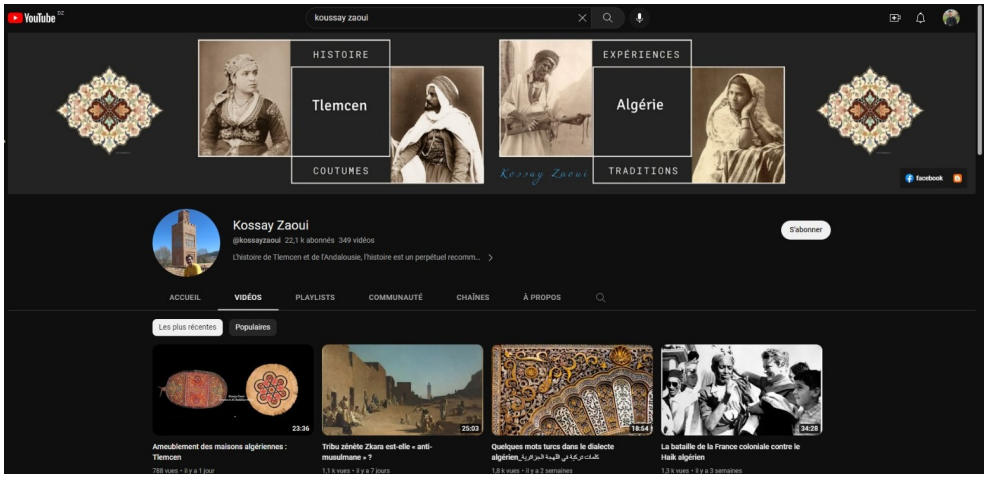


Figure 1.3: "Kossay Zaoui". chaîne pour l'histoire de Tlemcen et de l'Andalousie

Cependant, malgré ces initiatives, ce contenu reste souvent relégué au second plan sur ces plateformes. Les algorithmes de recommandation et de mise en avant des contenus favorisent souvent les produits étrangers, qui ont des budgets de promotion et de diffusion

plus importants, cela contribue à la marginalisation du contenu algérien et à la baisse de sa visibilité.

1.1.2 Préservation et protection des documents visuels algériens

La deuxième problématique concerne la suppression des chaînes sur Youtube (figure 1.4), malheureusement les chaînes culturelles algériennes ne font pas l'exception, plusieurs d'entre elles se trouvant sur Youtube ont été supprimées en raison de la violations des règles de la communauté ou les fausses plaintes de droits d'auteur.

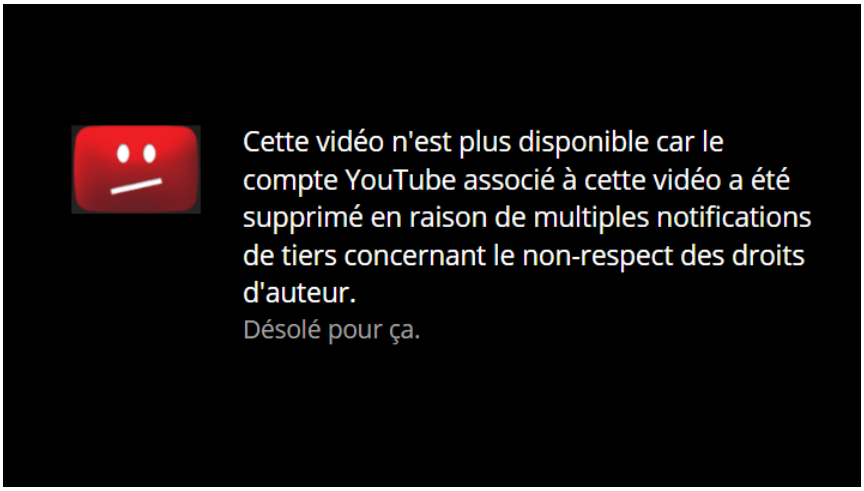


Figure 1.4: Youtube message après la suppression d'une vidéo

Cependant, une des raisons les plus courantes est le "faux signal" qui se définit en signalements abusifs malveillants de contenus qui sont effectués sans raison valable ou justifier souvent dans le but de nuire à la chaîne ou à son créateur. Ces signalements abusifs peuvent entraîner la suppression de la chaîne sans qu'il y est eu une vérification minutieuse du contenu signalé (figure 1.5).

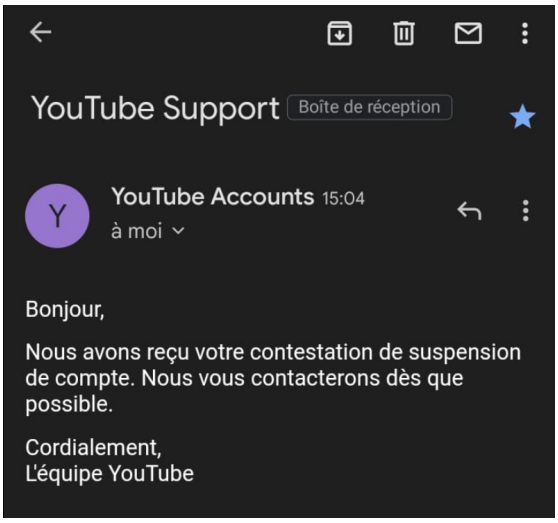


Figure 1.5: La réponse du Youtube après la réclamation d'un créateur de contenu.

Cette pratique est non seulement injuste pour les créateurs du contenu, mais elle a également un impact négatif sur la diversité des ceux disponibles sur la plateforme, par exemple la chaîne intitulée "Realities Dz حقائق" (figure 1.6) qui a malheureusement été supprimée de YouTube, une nouvelle chaîne portant le même nom a été créée le 2 juillet 2022.

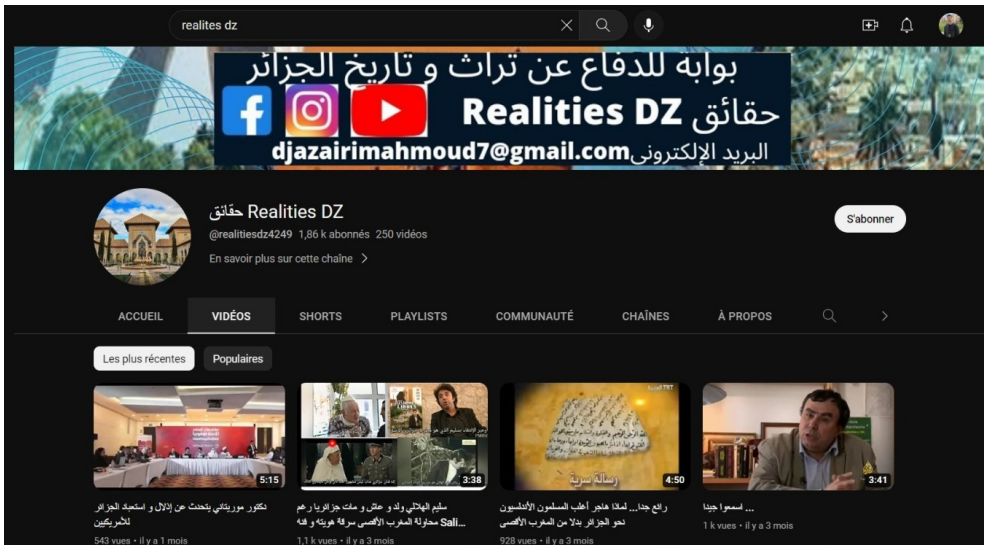


Figure 1.6: La chaîne Youtube "Realities Dz حقائق".

La suppression de ces chaînes peut entraîner la perte de documents précieux pour les chercheurs, les historiens et les amateurs de la culture algérienne .

1.1.3 Les défis des droits d'auteur

La troisième problématique est liée aux droits d'auteur sur YouTube, qui affectent de nombreuses chaînes et créateurs de contenu algériens et étrangers en général. Bien que le droit d'auteur soit une protection importante pour eux, il peut être utilisé de manière abusive pour supprimer le contenu légitime qui ne viole pas les droits d'auteur.

Dans certains cas, des personnes réclament les droits d'auteur sur des vidéos historiques ou d'archives qui ne leurs appartiennent pas réellement. Cela peut être dû à une confusion ou à une intention malveillante pour supprimer le contenu d'une chaîne concurrente.

De plus les algorithmes de détection automatique des droits d'auteur utilisés ne sont pas toujours précis et peuvent signaler à tort des vidéos qui ne violent pas réellement ces droits et cela peut conduire à des erreurs de suppression de ces enregistrements imagés qui ont un réel intérêt historique ou éducatif comme nous l'avons vu précédemment.

Dans un exemple plus explicite, il y a eu des cas où des chaînes Youtube qui partageaient des contenus historiques ont été confrontées à des revendications de droits d'auteur abusives. Par exemple, la chaîne 'Public Domain Movies' qui partageait des films du domaine public, a été confrontée à ce type de revendications sur des films qui étaient censés être dans le domaine public.

1.1.4 Adaptation à la spécificité de la culture algérienne

Malgré ces difficultés, les chaînes d'archives algériennes continuent de travailler pour préserver le patrimoine culturel du pays et le rendre accessible à un public mondial d'où la nécessite d'avoir une plateforme d'archivage nationale qui permet au peuple de contribuer à l'enrichir.

En plus, les utilisateurs ont des besoins spécifiques en matière de partage de vidéos, notamment les intérêts qui distinguent la population algérienne. Donc il existe un besoin croissant d'une base de partage de vidéos destinée typiquement aux Algériens, avec un accent mis sur la culture et les contenus locaux, une plateforme qui peut répondre aux besoins.

1.1.5 Les limites du partage sur les plateformes d'archive en Algérie

En effectuant nos recherches sur les plateformes d'archives numériques, nous avons constaté la présence de plusieurs plateformes à l'échelle mondiale. Certaines plateformes d'archives

sont gérées par des institutions gouvernementales telles que les archives nationales [24], tandis que d'autres sont gérées par des organisations non gouvernementales [49], des bibliothèques [23] ou des associations [14].

Nous avons remarqué que seule la plateforme "Portail du patrimoine culturel algérien" [1] est disponible en Algérie. Ils ont indiqué (figure 1.7) que, dans la phase actuelle de construction, seuls les corpus du Malouf de Constantine et de la Sanaâ d'Alger sont disponibles. Quant à celui de l'école de Tlemcen, il est actuellement en cours d'enrichissement.

Figure 1.7: Portail Du Patrimoine Culturel Algérien

Les partenaires de cette plateforme sont : Algérie Presse Service, la Radio Algérienne, la Télévision Algérienne et le Centre de recherche en information scientifique et technique (Cerist). Une part non négligeable de leurs archives intégrera le portail pour être mise à la disposition de ses usagers.

Le problème de cette plateforme est qu'elle ne permet l'accès qu'à ses partenaires pour ajouter leurs archives, alors que des archives, en particulier des vidéos, pourraient être trouvées au sein de la population elle-même. Cependant, ces personnes ne sont pas en mesure de contribuer leurs archives sur cette plateforme, ce qui les pousse à se diriger vers des plateformes mondiales comme YouTube pour partager leurs archives nationales, étant donné qu'il n'y a pas encore de solution algérienne pour cela.

Le principal défi donc est de créer une plateforme de partage de vidéos pour les Algériens, qui soit adaptée à leurs besoins et leurs préférences, et qui propose des vidéos pertinentes et de haute qualité . Toutefois, l'indexation de ces dernières de manière efficace et précise pour permettre aux utilisateurs de les trouver facilement est un autre

défi important. Les méthodes traditionnelles d'indexation sont souvent longues et fastidieuses, nécessitant une intervention humaine pour étiqueter chaque vidéo. C'est pourquoi l'utilisation de l'intelligence artificielle et des modèles de traitement du langage naturel peut être utile pour automatiser le processus d'indexation de vidéos.

1.2 Travaux existants

1.2.1 Les plateformes open source pour le partage de vidéos

Les plateformes open source pour le partage de vidéos sont des outils en ligne qui permettent aux utilisateurs de partager, visionner et échanger des vidéos en ligne. Contrairement aux plateformes de partage de vidéos traditionnelles, ces plateformes open source sont développées et maintenues par des communautés de développeurs bénévoles plutôt que par des entreprises privées. Cela signifie que le code source de la plateforme est accessible à tous et que les utilisateurs peuvent contribuer au développement de la plateforme en soumettant des suggestions et des modifications. Les plateformes open source pour le partage de vidéos sont de plus en plus populaires car elles offrent une alternative libre et transparente aux plateformes de partage de vidéos commerciales.

Au cours de notre exploration de ces plateformes, nous avons identifié plusieurs options pertinentes. Cependant, nous avons restreint notre sélection à celles qui répondent spécifiquement à nos exigences, tels que :

- La facilité d'utilisation par n'importe quelle personne.

- Une interface utilisateur offrant la possibilité de partager des vidéos et de consulter celles d'autres personnes sur notre plateforme.

- La sécurité des données et les performances de la plateforme basées sur les technologies utilisées et la manière d'écrire le code source.

- Un code propre et une architecture propre pour la possibilité de modifier le code par n'importe quel développeur de notre équipe.

Suite à notre recherche, nous avons rassemblé les plateformes identifiées dans un tableau. Ce dernier 1.1 contient les informations pertinentes et notre évaluation de chaque option, facilitant ainsi la prise de décision quant à la plateforme à adopter pour notre projet.

Après cela, on a choisi d'utiliser les plateformes qui ont comme langage le PHP dans le backend, car c'est sécurisé, stable, facile à apprendre et dispose d'une grande communauté de support.

Ensuite, nous avons commencé à tester des plateformes telles que BriskLimbs, Clipbucket et Vsw. Bien qu'ils aient des problèmes qui sont mentionnés dans le tableau sous la colonne "Négative", nous avons choisi de travailler avec Vsw car son code est propre et simple à utiliser, ce qui nous permet de corriger ses erreurs techniques et de l'adapter pour l'utiliser comme plateforme de partage de vidéos.

Nom	Code	Technologie	Positive	Négative
BriskLimbs	https://github.com/sakydev/briskLimbs	Html - Css - Javascript - Php - Mysql	Extensible, mise à jour quotidien, contrôle administrateur et bonne documentation	Après avoir essayé la plate-forme, elle a eu des problèmes pour télécharger des vidéos. Nous avons envoyé un e-mail au support de la communauté, mais ils n'ont pas répondu.
Brisklimbs Headless	https://github.com/sakydev/brisklimbs-headless	Html - Css - Php - Laravel - Mysql	Extensible, mise à jour quotidien, contrôle administrateur et bonne documentation	Nécessairement modifier tous les projets à cause de la version laravel.
Clipbucket	https://github.com/arslancb/clipbucket	Html - Css - Javascript - Php - Mysql	Documentation détaillée, extensible (intégration vidéo) et base de données bien structurée.	Plus mis à jour
Vsw	https://github.com/mohabmes/vsw	Html - Css - Php - Mysql	Code propre et simple à utiliser.	Obligation de refactoriser le code php car c'est une ancienne version et pas de support communautaire.
MediaDrop	https://github.com/mediadrop/mediadrop	Html - Css - Javascript - Python - Mysql	Interface administrative, stocker la vidéo n'importe où, plus rapide, évolutive et bonne documentation.	Code et structure complexes, nous ne sommes pas à l'aise avec python pur dans le web.
Klopix	https://github.com/RiatTahiri/Klopix	Html - Css - Javascript - ReactJs - NodeJs - ExpressJs - MongoDb	Pile MERN pour javascript, facile à basculer entre client et serveur (fluide), modèles architecturaux MVC	Projet en cours.
PeerTube	https://github.com/Chocobozzz/PeerTube	Html - Css - Javascript - Typescript - Mysql	Plus de 4 versions, communautés, Peer2Peer, transcodage vidéo et hautement configurable	Nous ne connaissons pas tous la technologie Typescript

Table 1.1: Plateformes open source pour le partage de vidéos.

1.2.2 Les approches d'indexation automatique

Avec la croissance exponentielle de la quantité de données multimédia disponibles, la recherche et l'identification d'informations pertinentes au sein de cette masse de données sont devenues des défis majeurs pour les utilisateurs. Cette difficulté est exacerbée dans le cas des données vidéo.

Pour relever ce défi, l'indexation automatique des bases de données est essentielle pour faciliter la recherche et réduire le temps de réponse des systèmes de recherche d'information. En effet, l'indexation automatique permet d'attribuer des index aux documents multimédia afin de les catégoriser et de faciliter leur recherche. L'utilisation de méthodes d'indexation automatique permet également d'améliorer l'efficacité et la précision de la recherche d'informations dans les bases de données multimédia.

Plusieurs approches d'indexation de vidéos ont été proposées dans la littérature [36, 56, 15, 44, 21, 31], dont l'indexation basée sur le contenu qui utilise des caractéristiques de bas niveau pour décrire le contenu de la vidéo, parmi lesquelles on compte les plus récentes :

Machine learning approches

Une autre étude a été réalisée par Alex Beutel et al [6], ils ont proposé une approche qui se base sûre le *Machine learning*, Les bases de données utilisent un *BTree-Index* pour rechercher tous les éléments dans une plage spécifique de clés. Les bases de données appliquent un *Hash-Map* pour rechercher l'enregistrement d'une seule clé. Un *BitMap-Index* est utilisé pour déterminer si une clé est présente,ces dernière année, ces structures de données ont été étudiées et améliorées [26, 20] mais ces méthodes supposent que la distribution des données dans le pire des cas,ce qui ne s'applique pas aux données du monde réel, c'est l'un de leurs principaux inconvénients.

Pour cela ils sont proposés d'utiliser un *machine learning* modèle pour apprendre le modèle de données, corrélations, etc. Pour synthétiser automatiquement une structure d'index et remplacer les *B-TREE*, les *hashmaps* et les filtres bloom par *learned indexes*. Ils ont utilisé *Regression model* avec *squared error* pour prédire la position du point initial et après exécuter une recherche binaire classique pour localiser la position exacte cela peut remplacer *B-Tree*.Apres Ils ont modifier *hash function* d'une manière qu'elle dépend du model et cela peut remplacer *hashmap*. Ensuite Ils ont considéré la recherche si un élément existait comme un problème de classification binaire dans lequel les éléments sont étiquetés "1" s'ils sont présents et "0" lorsqu'ils ne le sont pas.Pour la prédiction, de simples *RNN* et

CNN sont formés. Pour éviter les faux négatifs, un filtre bloom est utilisé. Pour remplacer *BitMap-Index*.

Reinforcement Learning approches

Basu et al. [4] ont proposé une technique de réglage utilisant *Reinforcement Learning* pour le problème de sélection d'index. Ils appliquent une méthode de réduction de l'espace d'état pour étendre leur algorithme aux bases de données et aux charges de travail plus importantes et traitent le problème de sélection d'index comme un *Markovian Decision Process*.Lan et al. [29] proposent une approche de conseil d'indexation utilisant des règles heuristiques et *Deep Reinforcement Learning*.Sadri et al. [45] utilisent l'apprentissage par *Reinforcement Learning* pour choisir les index pour les bases de données clusterisées. Les études précédentes [29] et [45] ont utilisé une diminution du coût d'exécution de la requête pour évaluer la performance de l'index choisi.

Vishal Sharma et al [47] ont proposé une approche qui s'appelle MANTIS un système autogéré qui génère automatiquement des index presque parfaits. En prenant en compte les multi-attributs et les différent types d'indexation sous une contrainte de taille de stockage finie. Il existes plusieurs anciens approches qui se base sûre l'indexation automatique [4, 17, 29, 38] mais ils ont soit limités leur recherche sur un type B-TREE ou concentre sur un seul attribue d'indexation ,soit ils n'ont pas considéré la taille de la base de données. MANTIS a deux phases : la première est INDEX TYPE SÉLECTION ils ont combiné "*supervised*" et "*reinforcement learning*", un *DNN* pour choisir le meilleur type d'indexation en convertans les requêtes *sql* en une représentation vectorielle.

et INDEX RECOMMANDATION,ils ont formuler ce problème comme *Markovian decision process* (MDP) et le *Deep Q-Learning network* qui recommande l'aspect du multi-attributs.

Transformers approches

Après une année, Vishal Sharma et al [46] ont proposé une autre approche qui s'appelle *indexer++* pour régler d'index en ligne tenant compte de *workload*. Les indexeurs en ligne sont confrontés à plusieurs défis non partagés par leurs homologues hors ligne comme la résilience au bruit, les frais généraux, la détection des tendances et la réactivité.

Pour cela *Indexer++* propose une solution à deux phases : la première phase identifie les tendances de *workload*, Ils sont utilisés *embedding*, les techniques NLP pour la

représentation vectorielle de workload, les modèles NLP qui ont été pré-entraînés pour détecter les changements dans les modèles de *workload* et les *K-medoids* sont utilisés pour agréger les tendances historiques et futures de *workload*.Et la deuxième phase ils sont utilisés *deep reinforcement learning* pour la sélection d'index, ils sont utilisés aussi des contraintes de taille de disque pour améliorer les performances de *deep Q-Networks* (DQN) [34] et ils sont gérés les *Noisy Rewards* à l'aide d'un filtre de convolution 1D et la priorité.

Cependant, ces approches souffre d'un gap sémantique, qui peut être pallié par l'indexation sémantique basée sur le contenu. Néanmoins, l'extraction automatique d'informations sémantiques à partir du contenu de la vidéo est une tâche complexe qui nécessite de prendre en compte plusieurs problèmes tels que la nature des indexes à associer aux documents vidéo en fonction des requêtes possibles de l'utilisateur.

En 2022 Muennighoff Niklas [35] ont proposent le modèle SGPT pour utiliser *decoder transformers* pour l'indexation sémantique et la recherche sémantique d'information via *prompting* ou *finetuning*.

ils ont proposé des modifications aux modèles GPT [43, 2] pour les utiliser comme *Cross-Encoders* ou *Bi-Encoders*, ils ont donc deux approches (figure 1.8) :

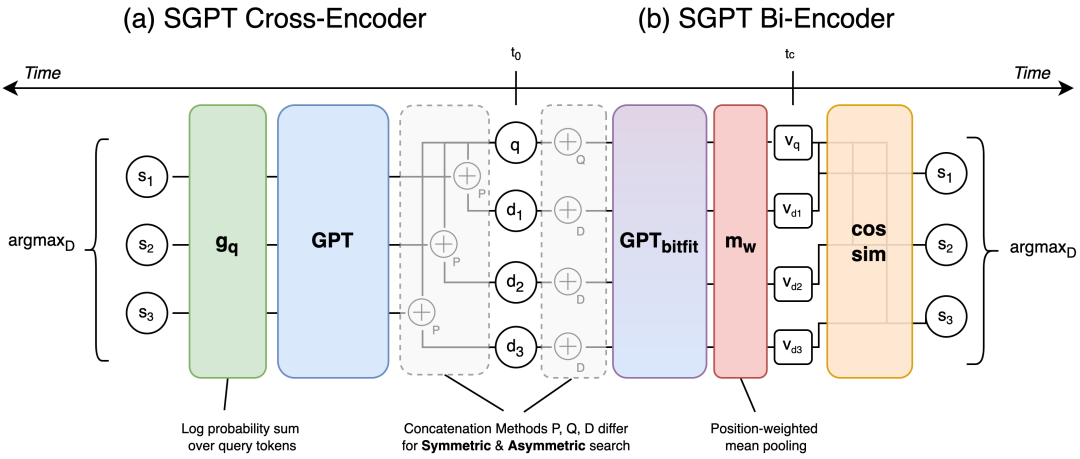


Figure 1.8: SGPT model

(i) SGPT-BE utilise *position-weighted mean pooling* et *fine-tuning* des seuls tenseurs de biais. Le modèle peut être utilisé pour la recherche sémantique ou l'indexation sémantique.

(ii) SGPT-CE extrait les log-probabilités des modèles GPT pré-entraînés pour produire des résultats de recherche de pointe non supervisés. La configuration présentée ne peut être utilisée que pour la recherche sémantique.

Enfin, cette recherche a choisi l'approche SGPT (GPT Sentence Embeddings for Semantic Search) comme point d'ancrage afin de démontrer l'importance de l'indexation automatique sémantique pour améliorer les performances de recherche dans les bases de données en utilisant les informations contextuelles qui sont le titre et la description pour avoir un embedding qui sera utiliser comme index.

En effet, Les résultats obtenus sont convaincants : avec ses 5,8 milliards de paramètres, SGPT surpasse les meilleurs plongements de phrases précédemment connus de 7% et dépasse une méthode concurrente utilisant 175 milliards de paramètres, comme mesuré sur le benchmark de recherche BEIR [33].

En résumé, l'utilisation de l'approche SGPT offre des performances améliorées, une précision accrue et des résultats remarquables dans le domaine de l'indexation automatique sémantique, et vu que les transformers se base sure le fine tuning également connu sous le nom de *transfert learning*, qui est une approche populaire en apprentissage par transfert qui consiste à adapter un modèle pré-entraîné à une nouvelle tâche avec des données étiquetées limitées. Dans le domaine du Traitement du Langage Naturel (NLP), le *fine-tuning* se réfère spécifiquement à l'ajustement des paramètres d'un modèle pré-entraîné, à une nouvelle tâche spécifique, en utilisant un ensemble de données spécifique à la tâche. Les modèles basés sur les transformers, tels que BERT [15], GPT [16] et RoBERTa [30], ont atteint des performances de pointe sur diverses tâches de NLP grâce à leur capacité à capturer les contextes et les significations des mots et des phrases. Cependant, l'entraînement de ces modèles à partir de zéro sur une nouvelle tâche peut être extrêmement coûteux en termes de ressources. Le fine-tuning d'un modèle pré-entraîné basé sur les transformers sur un ensemble de données spécifique à la tâche peut réduire considérablement le temps d'entraînement et améliorer les performances, surtout lorsque l'ensemble de données est petit ou que la tâche est similaire à la tâche de pré-entraînement., justifiant ainsi son choix dans cette recherche. Cette technique aussi permet de réduire le temps de recherche tout en augmentant la précision des résultats obtenus.

1.3 Conclusion

Dans ce chapitre, on a d'abord souligné l'importance de créer une plateforme de partage de vidéos avec un modèle de traitement de langage pour l'indexation automatique des vidéos, afin de résoudre les problèmes que nous avons identifiés. Ensuite, nous avons présenté les résultats de notre recherche sur les plateformes de partage de vidéos open source, ainsi que les études qui ont traité du problème d'indexation automatique. Nous avons ainsi exposé

notre choix pour d'entre elles.

Chapitre 2

Conception de la solution

2.1 Introduction

Avant de réaliser une solution ou un système informatique, une étape d'analyse et de conception est obligatoire. Cette étape a pour objectif de définir et formaliser les étapes nécessaires pour développer l'application, afin de répondre au mieux aux besoins des utilisateurs.

L'objectif de notre travail est de proposer un outil permettant à l'utilisateur de partager ses archives vidéo sur une plateforme dédiée, avec un mécanisme de recherche sémantique efficace pour trouver rapidement les vidéos pertinentes. Pour atteindre cet objectif, notre logiciel enchaîne les processus de partage, d'extraction des informations pertinentes et de recherche sémantique des vidéos.

Dans ce chapitre, nous présenterons le schéma global du système ainsi que notre conception détaillée de chaque module du système, afin de clarifier et préciser le fonctionnement de notre outil.

2.2 Schéma global de notre système

Notre système repose sur une architecture globale qui comporte trois processus principaux: Extraction des informations contextuelles d'une vidéo, prétraitement et vectorisation comme on peut le voir dans la figure 2.1 :

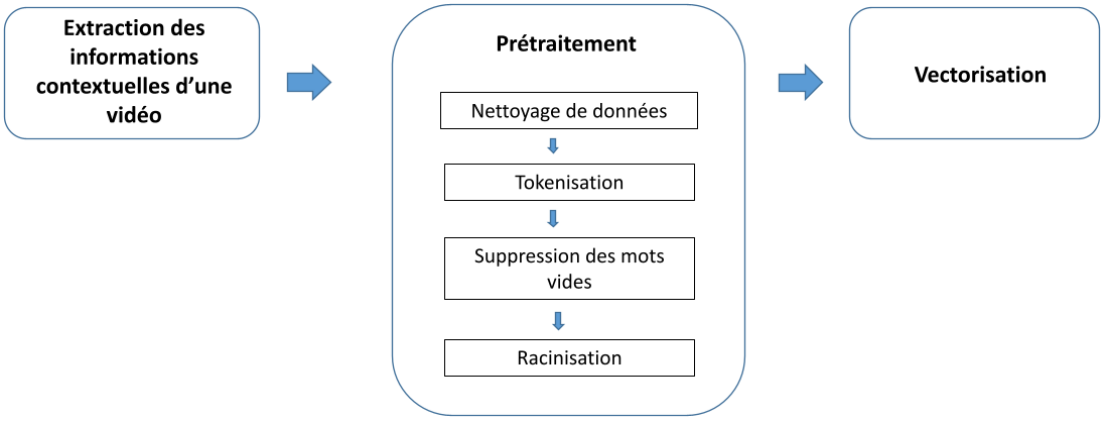


Figure 2.1: Schéma global du fonctionnement de notre système

Dans les sections suivantes, nous allons commencer par détailler les différents modules du système.

2.3 Conception du modèle d'indexation

Cette section explore la conception du modèle d'indexation . Elle est divisée en trois sous-sections : prétraitement des données, fine-tuning et le modèle de recherche sémantique, chacune abordant des aspects spécifiques liés à la recherche sémantique. Ces sous-sections offrent des aperçus précieux et des analyses approfondies.

2.3.1 Pré-traitement des données

Des opérations de prétraitement ont été appliquées à chaque article afin de rendre les données utilisables avec divers algorithmes d'apprentissage automatique. Prenons le texte suivant comme illustration :

مَرْحَبًا! كَيْفَ حَالُكُمْ؟ ، هل تستمتعون باليوم ؟ .

Les étapes de prétraitement suivantes ont été effectuées :

Nettoyage des données

Le nettoyage des données est une étape cruciale du prétraitement dans le domaine du traitement du langage naturel. Il permet d'éliminer les informations indésirables, les erreurs et le bruit des données textuelles, afin de préparer les données pour une utilisation optimale dans des algorithmes d'apprentissage automatique. Prendre l'exemple précédent, La première étape consisterait à supprimer les caractères spéciaux tels que les symboles de ponctuation. Donc, le texte deviendrait : مرحبا كيف حالكم هل تستمتعون باليوم .

Ensuite, nous pourrions éliminer les balises HTML, les numéros de référence, d'URL ou d'adresses e-mail, même si elles ne sont pas présentes dans cet exemple spécifique. Enfin, nous pourrions vérifier et corriger les erreurs orthographiques éventuelles.

Ces étapes de nettoyage des données sont essentielles pour obtenir des données textuelles de qualité et pour garantir des résultats précis et fiables lors de l'application d'algorithmes d'apprentissage automatique.

Tokénisation

Après avoir nettoyé les données textuelles en éliminant les informations indésirables, les erreurs et le bruit, la tokénisation est une étape essentielle du prétraitement. Elle consiste à diviser le texte en unités plus petites appelées "tokens" afin de structurer le texte et le préparer pour des analyses plus avancées. Après avoir appliqué la tokénisation sur l'exemple précédent, le texte pourrait être divisé en tokens individuels tels que "مرحبا" "كيف" "حالكم" "هل" "تستمتعون" "باليوم" et ".". Chaque mot constitue un token distinct qui peut être utilisé pour l'analyse ultérieure. La tokenisation peut également prendre en compte la gestion des signes de ponctuation, tels que la virgule et le point d'interrogation seraient traités comme des tokens séparés.

Dans certains cas, la tokenisation peut également inclure la création de tokens composés de plusieurs mots consécutifs, appelés n-grammes. Par exemple, si nous voulons considérer des tokens composés, le texte "machine learning" pourrait être tokenisé en "machine", "learning" et "machine learning" en tant que token composé.

Enfin, après la tokenisation, des étapes supplémentaires de normalisation peuvent être appliquées, telles que la conversion en minuscules, la lemmatisation ou la suppression des mots vides. Cela permet d'obtenir des tokens plus cohérents et significatifs pour faciliter les analyses ultérieures.

En résumé, la tokenisation après le nettoyage des données permet de structurer le texte en unités significatives, favorisant ainsi l'analyse et le traitement des données textuelles. Les tokens obtenus peuvent être utilisés dans différentes tâches de traitement du langage naturel, comme la classification de texte, la génération de texte, l'extraction ou la recherche d'informations ce qui nous intéresse dans cette recherche.

Suppression des mots vides (*stopwords*)

Une fois que les données ont été nettoyées et tokenisées, la suppression des *stopwords* devient une étape essentielle du prétraitement des données textuelles. Après avoir divisé le texte en *tokens*, l'objectif de cette étape est d'éliminer les mots qui font partie de la liste des *stopwords*.

Les *stopwords* sont des mots couramment utilisés tels que des articles, des prépositions, des pronoms et des conjonctions, qui n'apportent pas de signification spécifique et apparaissent fréquemment dans les documents. Leur présence ne contribue que peu à la compréhension globale du texte.

La liste des *stopwords* peut être prédéfinie dans des bibliothèques de traitement du langage naturel ou personnalisée selon les besoins du projet. Les mots correspondant à cette liste sont alors supprimés des *tokens*, réduisant ainsi leur impact sur les résultats finaux.

Il convient de souligner qu'il est important d'utiliser la suppression des *stopwords* avec prudence, car certains *stopwords* peuvent contenir des informations pertinentes dans certains contextes. Il est donc recommandé de vérifier attentivement la liste et de l'adapter en fonction des exigences spécifiques du projet.

Considérons le texte suivant en arabe : في يوم جميل، ذهبت إلى المكتبة لقراءة كتاب جديد عن الثقافة العربية وتاريخها. كانت القصة مثيرة ومفيدة لفهم التطورات الاجتماعية والثقافية في المنطقة.

Pour mettre en évidence l'importance de l'élimination des stopwords, nous pouvons appliquer cette étape de prétraitement. Les stopwords couramment utilisés en arabe tels que في (fi) et عن (an) peuvent être éliminés pour se concentrer sur les mots clés.

Après l'élimination des stopwords, le texte ressemblera à ceci : يوم جميل، ذهبت المكتبة لقراءة كتاب جديد الثقافة العربية وتاريخها. القصة مثيرة ومفيدة لفهم التطورات الاجتماعية والثقافية المنطقة.

En supprimant les stopwords cela nous permet de mettre l'accent sur les mots clés tels que يوم (jour), مكتبة (bibliothèque)...ect

En conclusion, la suppression de ces mots, constitue une étape cruciale du prétraitement. Son objectif est de réduire le bruit et de se concentrer sur les mots significatifs pour les analyses ultérieures, notamment dans le cadre de la recherche d'informations.

Racinisation

La racinisation est une étape essentielle du pré traitement des données. Elle vise à extraire la racine des mots afin de les ramener à leur forme de base. les mots peuvent avoir différentes formes et conjugaisons en fonction de leur utilisation dans une phrase. Cependant, en identifiant la racine d'un mot, on peut faciliter la recherche, la classification et l'analyse ultérieure des données.

Par exemple, prenons le mot arabe كتبت (katabtu), qui signifie "j'ai écrit" en français. En utilisant la racinisation, nous pouvons extraire la racine كتب (katab), qui représente le verbe "écrire". Cela nous permet de regrouper ce mot avec d'autres formes conjuguées du même verbe, facilitant ainsi l'analyse de son utilisation et de son contexte.

En conclusion, la racinisation est une étape clé du prétraitement des données en arabe, permettant de ramener les mots à leur forme de base en identifiant leur racine. Cela facilite

l'analyse et l'utilisation ultérieure des données, contribuant ainsi à des applications plus efficaces dans le domaine de la langue arabe.

2.3.2 Transfer learning (*Fine tuning*)

Le fine-tuning présente plusieurs avantages. Tout d'abord, il permet de réduire le temps et les ressources nécessaires à l'entraînement d'un modèle de haute qualité. De plus, il permet d'adapter le modèle aux particularités de la tâche cible, conduisant ainsi à de meilleures performances par rapport à l'entraînement à partir de zéro. En outre, le fine-tuning est particulièrement utile lorsque les données d'entraînement sont limitées, car il permet d'exploiter efficacement les informations contenues dans le modèle pré-entraîné.

Cependant, le fine-tuning nécessite un ensemble de données suffisamment représentatif et similaire à la tâche cible pour obtenir de bons résultats. Des précautions doivent également être prises pour éviter le surapprentissage (*Over Fitting*).

En conclusion, le fine-tuning est une technique puissante qui permet d'adapter un modèle pré-entraîné à une tâche spécifique en tirant parti des connaissances préalables du modèle. Cela peut conduire à des améliorations significatives des performances, en particulier lorsque les données d'entraînement sont limitées. Le fine-tuning est devenu une pratique courante dans la communauté NLP, offrant une solution efficace pour l'adaptation des modèles pré-entraînés à de nouvelles tâches et de nouveaux domaines.

Vectorisation

La vectorisation est une étape cruciale qui permet de représenter des données textuelles sous forme de vecteurs numériques.

Cette représentation vectorielle est essentielle pour que les algorithmes d'apprentissage automatique puissent traiter les données textuelles, car ils sont conçus pour manipuler des données numériques. Parmi les différentes techniques de vectorisation, les *word embeddings* jouent un rôle majeur , ils de représenter les mots sous forme de vecteurs dans des espaces sémantiques, où des mots ayant des significations similaires sont situés à proximité les uns des autres. Grâce à cela, ses representations capturent les relations sémantiques et syntaxiques entre les mots. Ils sont appris à partir de vastes corpus de texte en utilisant des techniques telles que *Word2Vec* [32] ou *GloVe* [39]. Les *Word Embeddings* sont couramment utilisés dans les differents tâches telles que la recherche de similarité sémantique. En transformant le texte en vecteurs numériques, la vectorisation des données textuelles

permet aux algorithmes d'apprentissage automatique de traiter le texte de manière plus efficace. Cette transformation facilite également la capture d'informations sémantiques et contextuelles importantes pour les tâches d'analyse et de modélisation ultérieures. dans notre cas la représentation vectorielle se base sure deux technique illustrée comme suivant : 2.2.

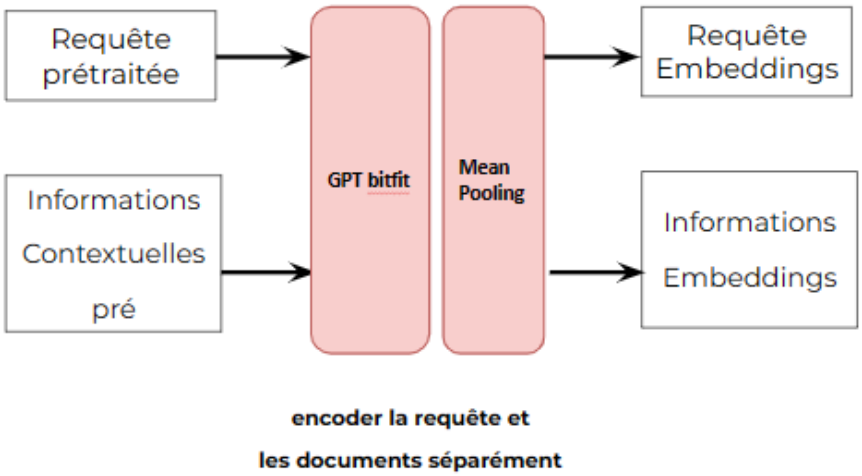


Figure 2.2: embeddings du modele.

le GPT bitfit est une méthode de fine tuning proposée par Google AI,Il vise à compresser le modèle tout en préservant son performance.

Le Position weighted mean pooling est une technique qui permet de capturer l'importance relative des mots dans une séquence en attribuant des poids en fonction de leur position.

Nous avons utilisée le tokenizer de la bibliothèque Hugging face pour la transformation des mots au représentations vectorielles ,hugging face offre des outils et des ressources pour faciliter l'utilisation et le fine-tuning des modèles pré-entraînés, ce dernier va tokenizer et encoder les inputs afin de produire les représentation vectorielle

pour bien comprendre la partie embedding et la transformation des donnée textuelle au vecteur voici un exemple dans le cas général : 2.3.

prenons la phrase "Hello world" comme input, d'abord le tokenizer va transformer cette phrase en tokens après transformer chaque token a un vecteur d'une taille de 768 ,ensuite multiplier chaque vecteur a le poids qui le correspond et a la fin calcule la moyenne de ces vecteur a fin de produire la représentation finale du input .

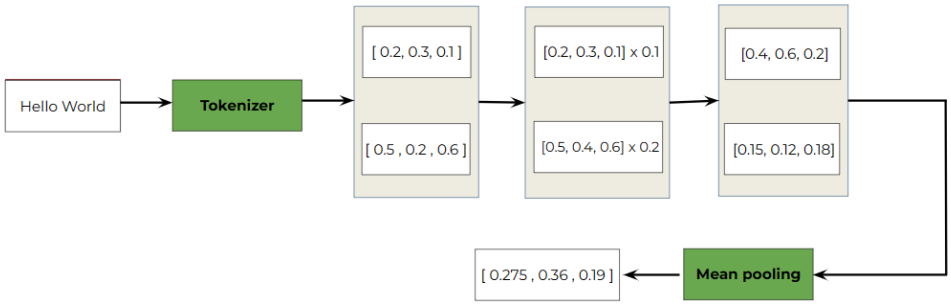


Figure 2.3: exemple du embeddings.

2.3.3 Le modèle du recherche sémantique

le SGPT est un modele pré-entraîné basée sur un décodeur transformer pour le traitement du langage naturel, développée par Niklas Muennighoff, conçue spécifiquement pour la recherche et l'indexation sémantiques. L'utilisation efficace de SGPT pour la recherche et l'indexation sémantiques repose sur deux approches : SGPT-BE et SGPT-CE.

SGPT-BE modifie les modèles GPT pour être utilisés comme Cross-Encoders ou Bi-Encoders, en utilisant la pondération moyenne positionnelle et l'ajustement fin des tenseurs de biais (BitFit [5]). Cette approche peut être utilisée pour la recherche sémantique sans avoir besoin d'encoder la requête et les vidéos ensemble. Cela est possible car les vidéos sont pré-encodés et indexés dans une étape séparée, et le modèle peut simplement utiliser les vidéos pré-encodés pour la recherche sémantique.

D'autre part, SGPT-CE extrait les log-probabilités des modèles GPT pré-entraînés. Cependant, cette approche nécessite d'encoder la requête et les vidéos ensemble, ce qui la rend moins efficace pour les grands ensembles de données.

Dans notre cas, malgré que les Cross-Encoders ont tendance à être plus performants que les Bi-Encoders[50], nous sommes intéressés par l'approche SGPT-BE (figure 2.4), car elle permet une recherche sémantique efficace sans avoir besoin d'encoder la requête et les documents ensemble. En pré-encodant les documents et en affinant le modèle SGPT en utilisant uniquement les tenseurs de biais, le modèle peut effectivement effectuer une recherche sémantique et une indexation sans avoir besoin de ré-encoder constamment tous les documents avec chaque nouvelle requête. Cela nous permet d'optimiser les performances de recherche de notre système sans compromettre l'efficacité ou la qualité des résultats de recherche.

Étant donné une requête q et des documents d1-3, SGPT classe les documents avec des

scores s1-3. SGPT-BE encode séparément les requêtes et les documents. Les vecteurs de documents résultants v1-3 peuvent être mis en cache et récupérés au moment t, lorsque qu'une nouvelle requête est soumise. Les scores sont des similarités cosinus voir figure 2.4.

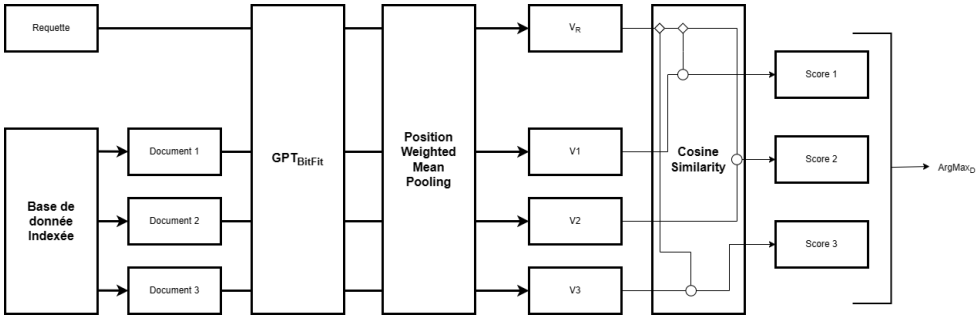


Figure 2.4: Fonctionnement de SGPT-BE.

Dans notre cas, les embeddings des vidéos sont déjà encodés dans une base de données indexée. Ensuite, la requête est encodée et la similarité cosinus est calculée entre la requête et chaque document. Enfin, la vidéo avec la similarité cosinus maximale est sélectionné comme la réponse à la requête. Cette approche permet une recherche sémantique efficace sans avoir besoin d'encoder la requête et les vidéos ensemble à chaque fois. Les vidéos encodés peuvent être stockés en cache et réutilisés pour de futures requêtes. comme expliqué dans la figure 2.5 :

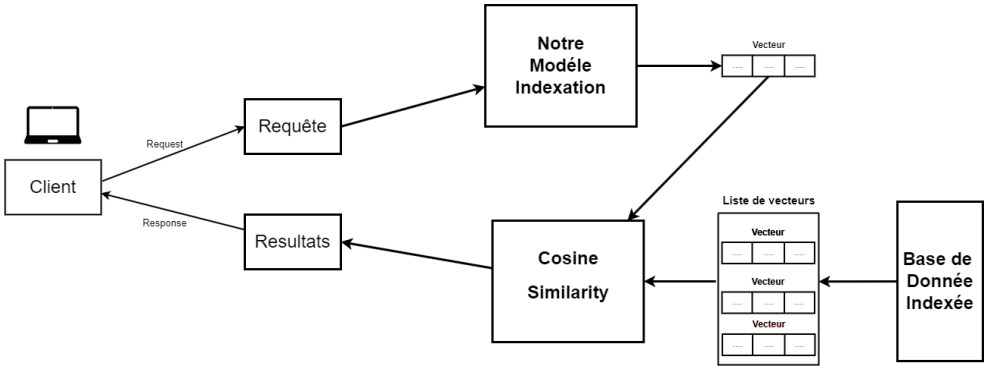


Figure 2.5: processus de recherche.

Une base de données indexée sera créée au début, et chaque fois qu'une nouvelle vidéo sera insérée, ses informations (le titre et sa description) seront encodées et le résultat sera inséré dans la base de données indexée pour une utilisation ultérieure lors de la recherche. comme expliqué dans la figure 2.6 :

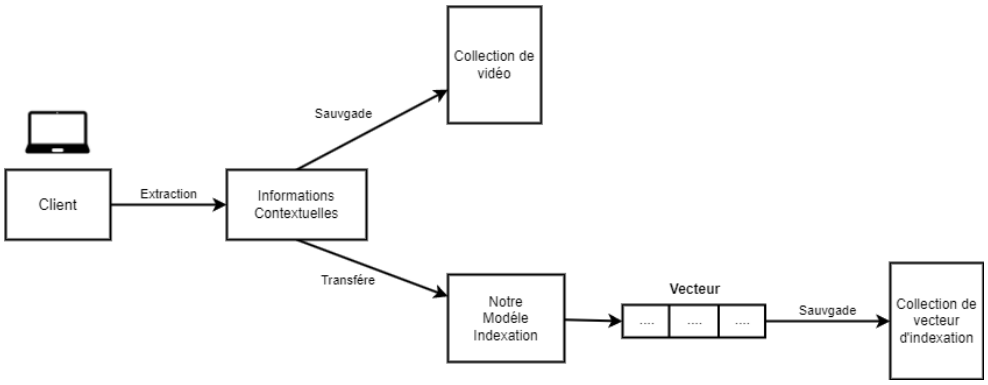


Figure 2.6: Processus de l'insertion d'une nouvelle vidéo.

La similarité cosinus est une mesure de similarité qui calcule la similarité entre deux vecteurs en mesurant l'angle entre eux dans un espace multidimensionnel. voici la formule pour calculé la similarité cosinus 2.1 :

$$Cosine\ Similarity = \frac{\sum_{i=1}^{n} A_i B_i}{\sqrt{\sum_{i=1}^{n} A_i^2}\sqrt{\sum_{i=1}^{n} B_i^2}} \tag{2.1}$$

A = Requete B = Document n= taille du documents

Plus précisément, elle mesure le cosinus de l'angle entre les deux vecteurs, donnant ainsi une valeur entre -1 et 1, où 1 indique une similarité maximale et -1 une similarité minimale. C'est une mesure couramment utilisée pour la recherche d'information et la récupération d'informations, notamment dans les moteurs de recherche. comme expliqué dans la figure 2.7 :

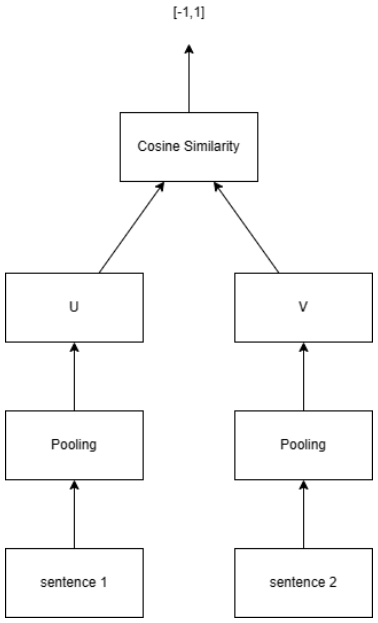


Figure 2.7: Fonctionnement de similarité cosinus.

Prenons un exemple,la phrase 1 : 'Hello World', la phrase 2 : 'Hello' et la phrase 3 : 'Hello Hello Hello'. Tout d'abord, nous calculons le nombre d'occurrences de chaque mot, donc nous trouvons pour la phrase 1 (1,1), pour la phrase 2 (1,0) et pour la phrase 3 (3,0). Nous représentons les résultats et calculons le cosinus de l'angle entre les phrases. À la fin, nous trouvons que cosine (phrase 1,phrase 2) et cosine (phrase 1,phrase 3) sont égaux à cosinus (45) = 0,71, ce qui signifie qu'elles sont très probablement similaires et que la phrase 3 et la phrase 2 sont identiques. vu que cosinus (0) = 1. Comme expliqué dans la figure 2.8 :

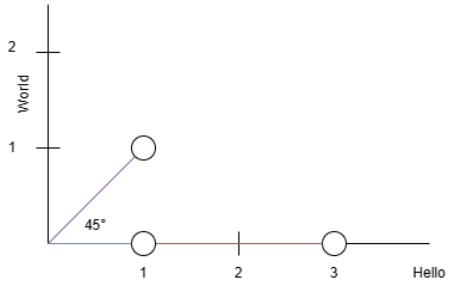


Figure 2.8: Exemple de similarité cosinus.

2.4 Le déploiement du modèle SGPT

Le modèle SGPT a été conçu pour faciliter la recherche sémantique et l'extraction d'informations pertinentes à partir d'un texte donné. Notre solution utilise le framework FastAPI pour le déploiement du modèle comme expliqué dans la figure 2.9 :

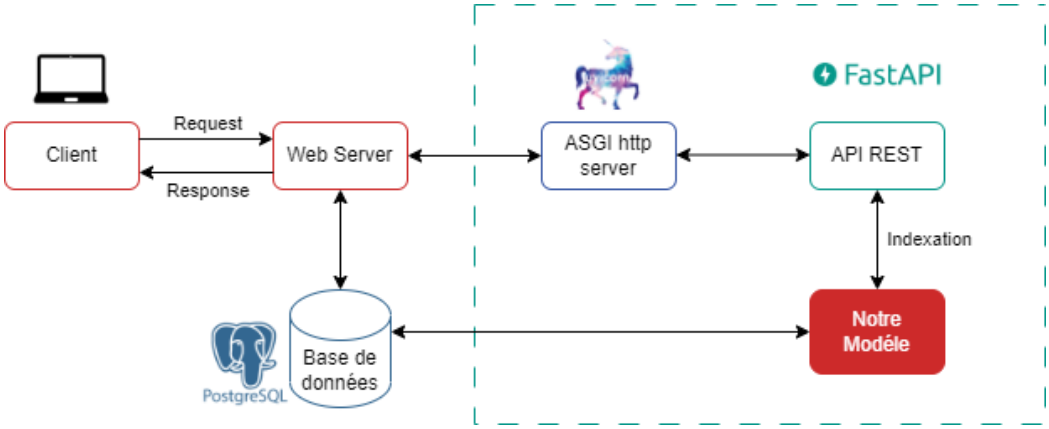


Figure 2.9: Schéma global du fonctionnement de notre système

Pour la recherche, le modèle SGPT prend en entrée une requête contenant un texte de recherche et extrait l'information pertinente de celle-ci. Ensuite, il compare le résultat avec les données d'indexation disponibles en utilisant la mesure de similarité cosinus. Le modèle utilise ensuite les relations sémantiques capturées pour identifier les documents pertinents correspondant à la requête.

Pour l'indexation, le modèle SGPT prend en entrée un texte pour en extraire l'information pertinente. En utilisant son approche, il peut créer un vecteur d'indexation de ce texte et le garder avec les données d'indexation disponibles.

FastAPI [12] est un framework Web léger et rapide pour la création d'APIs basées sur Python. Il utilise une syntaxe déclarative pour les routes et les modèles de données, ce qui facilite la création rapide d'APIs. De plus, il offre des performances supérieures à celles des autres frameworks tels que Flask et Django.

Le framework prend également en charge l'utilisation de types d'annotations Python pour la validation des données, ce qui facilite la détection précoce des erreurs et le développement d'APIs plus fiables.

En outre, FastAPI prend en charge la norme OpenAPI pour la génération de documentation interactive, ce qui permet de générer automatiquement une documentation pour

notre solution, ce qui facilite grandement le processus de développement et de test de notre solution.

En conclusion, la solution basé sur l'utilisation de FastAPI pour le déploiement de notre modèle SGPT offre une approche efficace pour la recherche sémantique et l'extraction d'informations pertinentes à partir d'un texte donné.

2.5 Conception de la plateforme de partage de vidéo

Dans cette section, nous allons expliquer comment le logiciel offre ses diverses fonctionnalités. Nous allons décrire le fonctionnement du système à l'aide d'un langage de modélisation. Notre choix s'est porté sur la méthode UML (Unified Modeling Language), une notation graphique conçue pour représenter, spécifier et construire des systèmes logiciels [25]

2.5.1 Diagramme de cas d'utilisation

Les cas d'utilisation sont une façon efficace de décrire les fonctionnalités de notre plateforme de partage de vidéos. Chaque cas d'utilisation devrait décrire une interaction spécifique entre l'utilisateur et le système, et peut être accompagné d'un diagramme de cas d'utilisation pour faciliter la compréhension.

Dans notre plateforme de partage de vidéos, nous avons identifié plusieurs cas d'utilisation qui nous avons regroupés dans le diagramme de cas d'utilisation global (figure 2.10) suivant:

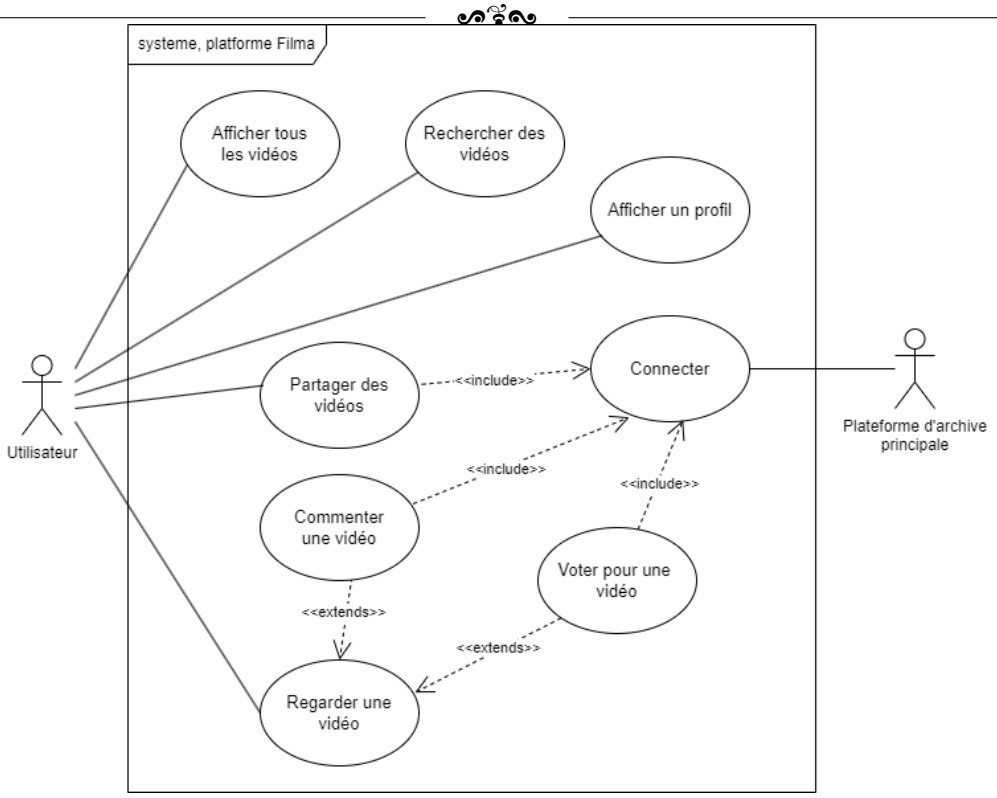


Figure 2.10: Diagramme de cas d'utilisation global.

Pour documenter les détails de chaque interaction utilisateur-système, nous avons créé des tableaux de cas d'utilisation pour chacun d'eux.

Chaque tableau (2.1 2.2 2.3 2.4 2.5 2.6 2.7) contient des informations clés sur le cas d'utilisation, telles que l'auteur, la description, les préconditions et les postconditions.

En utilisant ces tableaux, nous pouvons décrire chaque étape spécifique impliquée dans chaque cas d'utilisation, clarifier les exigences fonctionnelles et nous assurer que tous les scénarios possibles ont été pris en compte.

Auteur	Utilisateur.
Description	L'utilisateur peut partager des vidéos depuis son ordinateur sur la plateforme et lui donner un titre ainsi qu'une description. Le système vérifie alors le format de la vidéo et la taille du titre avant de la rendre disponible sur la plateforme.
Précondition	L'utilisateur doit être connecté à la plateforme.
Postcondition	Lors de la phase de téléchargement de la vidéo, le titre et la description sont fusionnés et les informations pertinentes sont extraites par notre modèle.

Table 2.1: Cas d'utilisation partager des vidéos.

Auteur	Utilisateur.
Description	L'utilisateur peut effectuer une recherche sémantique sur des vidéos en entrant des mots-clés dans la barre de recherche. Le système affichera alors les vidéos pertinentes pour l'utilisateur.
Précondition	La barre de recherche est visible sur la page de recherche.
Postcondition	Les résultats de recherche pertinents sont affichés à l'utilisateur. L'utilisateur peut visionner les vidéos trouvées ou modifier les termes de recherche pour une nouvelle recherche.

Table 2.2: Cas d'utilisation rechercher des vidéos.

Auteur	Utilisateur.
Description	Lorsqu'il accède à la page de lecture de la vidéo, l'utilisateur peut lire la vidéo, mettre en pause, reprendre la lecture, avancer ou reculer dans la vidéo, ajuster le volume, activer ou désactiver les sous-titres, et activer ou désactiver le mode plein écran. L'utilisateur peut également voir les commentaires associés à la vidéo et les votes attribués à la vidéo.
Précondition	La vidéo à regarder est disponible sur la plateforme.
Postcondition	L'utilisateur a regardé la vidéo.

Table 2.3: Cas d'utilisation regarder une vidéo.

Auteur	Utilisateur.
Description	L'utilisateur peut donner un vote positif ou négatif à une vidéo.
Précondition	L'utilisateur est connecté à la plateforme.
Postcondition	Le système enregistre le vote et met à jour les statistiques de la vidéo.

Table 2.4: Cas d'utilisation voter pour une vidéo.

Auteur	Utilisateur.
Description	L'utilisateur peut ajouter un commentaire à une vidéo et, par la suite, la supprimer.
Précondition	L'utilisateur doit être connecté à la plateforme.
Postcondition	Le système affiche le commentaire sur la page de la vidéo.

Table 2.5: Cas d'utilisation commenter une vidéo.

Auteur	Utilisateur.
Description	Afficher la page de profil de l'utilisateur et visualiser toutes ses vidéos.
Précondition	L'utilisateur doit être connecté pour voir son propre profil. Sinon, n'importe qui pourrait accéder au profil d'autres personnes.
Postcondition	La page de profil de l'utilisateur avec la liste des vidéos qu'il a téléchargées est affichée.

Table 2.6: Cas d'utilisation afficher un profil.

Auteur	Utilisateur.
Description	Visualiser une liste de toutes les vidéos disponibles sur la plateforme.
Précondition	Il n'y a pas de précondition pour ce cas d'utilisation.
Postcondition	L'utilisateur peut regarder une liste de toutes les vidéos disponibles sur la plateforme.

Table 2.7: Cas d'utilisation afficher tous les vidéos

2.5.2 Diagramme de classe

La conception de la plateforme de partage de vidéos a été réalisée en utilisant une approche orientée objet. Dans le cadre de cette approche, un diagramme de classe a été créé pour identifier les entités principales du système et leurs relations.

Ce diagramme de classe (figure 2.11) est utilisé comme point de départ pour la conception de la solution.

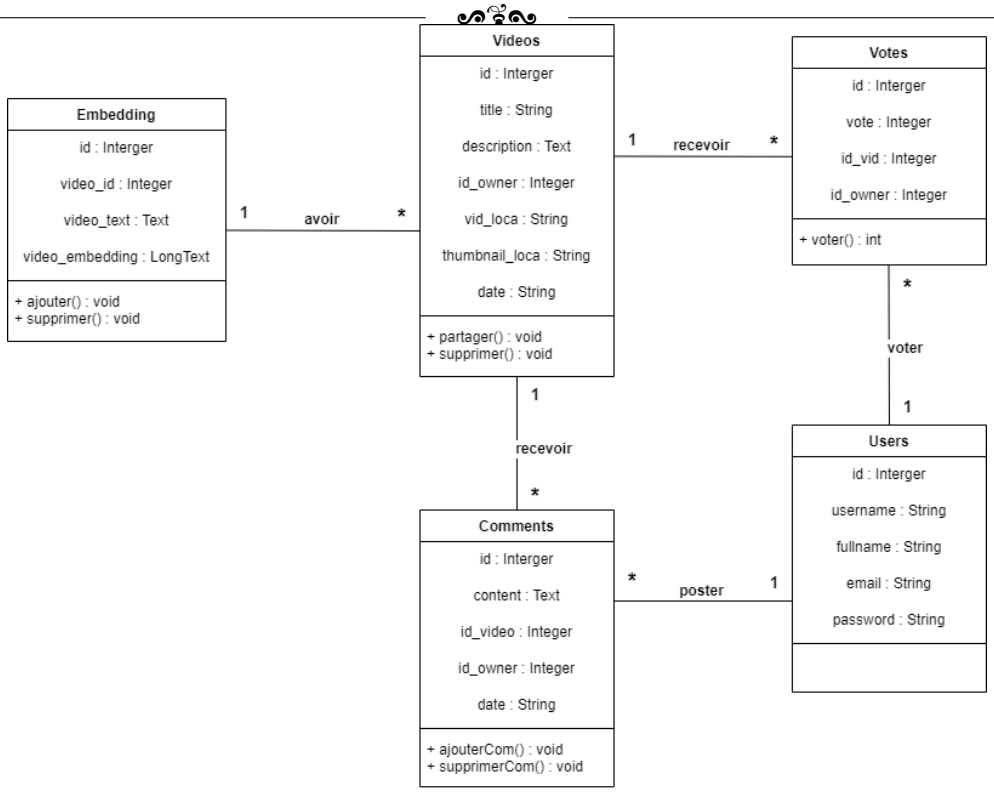


Figure 2.11: Diagramme de classe.

La relation entre "vidéo" et "embedding" est de un à un. Cela signifie qu'une vidéo peut avoir un seul embedding et qu'un embedding peut être associé à une seule vidéo. Cette relation peut être utile pour garder l'information pertinente décrivant le contenu d'une vidéo.

La relation entre "vote" et "vidéo" est de un à plusieurs. Cela signifie qu'une vidéo peut recevoir plusieurs votes et qu'un vote est associé à une seule vidéo. Cette relation permet de capturer les opinions des utilisateurs sur les vidéos.

La relation entre "comment" et "vidéo" est de un à plusieurs. Cela signifie qu'une vidéo peut avoir plusieurs commentaires et qu'un commentaire est associé à une seule vidéo. Cette relation permet aux utilisateurs de commenter les vidéos et de partager leurs opinions.

La relation entre "user" et "comment" est de un à plusieurs. Cela signifie qu'un utilisateur peut poster plusieurs commentaires et qu'un commentaire est posté par un seul utilisateur. Cette relation permet d'associer les commentaires aux utilisateurs qui les ont postés.

La relation entre "user" et "vote" est de un à plusieurs. Cela signifie qu'un utilisateur

peut voter pour plusieurs vidéos et qu'un vote est effectué par un seul utilisateur. Cette relation permet de suivre les préférences des utilisateurs.

2.6 Conception de la plateforme d'archivage principale

"*Mahroussa tech*" est considérée comme la plateforme principale d'archive algérienne et qui contient les mêmes utilisateurs pour les différents plateformes intégrées à celle-ci.

La relation entre "*Mahroussa tech*" et les différents plateformes, telles que "*Filma*" (plateforme de partage de vidéos) et d'autres, est expliquée par le schéma suivante 2.12 :

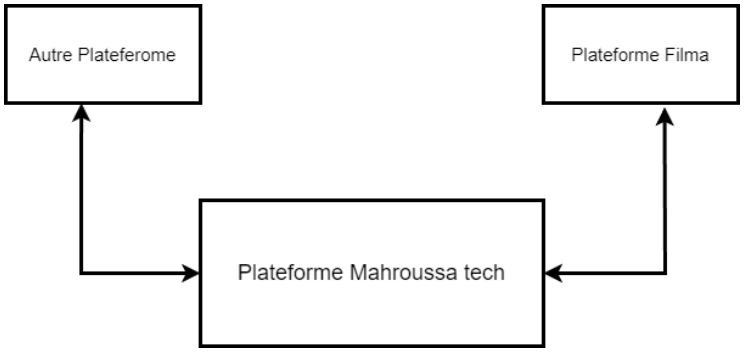


Figure 2.12: Schéma explique la relation entre les plateformes.

Lorsque l'utilisateur accède à la plateforme d'archivage principale "*Mahroussa tech*", il a la possibilité de créer un compte ou de s'authentifier s'il en possède déjà un. Les informations du compte sont sauvegardées dans la base de données principale de "*Mahroussa tech*" et sont également transférées vers les bases de données des différentes plateformes intégrées à celle-ci.

2.7 Conclusion

Dans ce chapitre, on a présenté une conception semi-formalisée des solutions proposées, en incluant des schémas, des diagrammes et des justifications. La prochaine étape consistera à mettre en œuvre ces solutions et à évaluer leur efficacité

Chapitre 3

Implémentation et résultats

3.1 Introduction

Dans ce chapitre, nous allons présenter la partie de réalisation de notre projet. Notre objectif dans le développement était de parvenir à créer un produit final utilisable par les utilisateurs.

Tout d'abord, nous présenterons l'environnement et les outils de travail utilisés. Ensuite, nous présenterons la métrique d'évaluation, ainsi que les résultats et les tests. Nous décrirons également en détail les étapes d'adaptation affectées à la plateforme de partage de vidéos, afin d'expliquer l'intégration du modèle à cette plateforme. Enfin, nous clôturerons avec la présentation de l'interface graphique.

3.2 Environnement et outils de travail

3.2.1 Matériels

Le matériel utilisé consiste en 2 ordinateurs personnels.

Le premier poste de travail dispose d'un système d'exploitation Windows 11 Famille 64 bits. Il est équipé d'un processeur AMD Ryzen 5 4500U avec Radeon Graphics @ 2,38 GHz, et d'une mémoire RAM de 8 Go.

Le deuxième poste de travail a utilisé Kaggle Notebooks [28], un environnement de calcul cloud qui permet une analyse reproductible et collaborative, ainsi que l'écriture et l'exécution de code Python. Cet environnement offre des ressources informatiques gratuites avec une accélération GPU pour les tâches d'apprentissage.

3.2.2 Langages de programmation et logiciels

Au cours du développement de notre système, nous avons employé divers langages de programmation et logiciels. Voici une courte description de ces outils utilisés :

Langages de programmation

Figure 3.1: Logos des langages de programmation

- **Python** : un langage de programmation de haut niveau qui prend en charge la programmation impérative structurée, fonctionnelle et orientée objet. Il se distingue par son typage dynamique solide et sa gestion automatique de la mémoire. De plus, il est livré avec plusieurs bibliothèques qui simplifient le processus de développement [53].
- **PHP** : acronyme de "Hypertext Preprocessor", est un langage de script côté serveur généraliste et Open Source, spécialement conçu pour le développement d'applications web. Il peut être intégré facilement au HTML. PHP offre une syntaxe simple et familière, ce qui facilite aux développeurs de manipuler des données, gérer des formulaires, interagir avec des bases de données et créer des fonctionnalités avancées pour offrir une expérience utilisateur enrichissante [27].

Outils et bibliothèques

Figure 3.2: Logos de quelques librairies

- **PyTorch** : PyTorch est une bibliothèque open source pour l'apprentissage machine, développée par Meta et basée sur Torch. Elle permet d'effectuer des calculs tensoriels optimisés pour le CPU et le GPU compatible avec CUDA. PyTorch peut être programmé en Python, et leur fonctionnalités incluent la manipulation de tenseurs avec une intégration facile avec Numpy, ainsi que des calculs efficaces sur CPU et GPU tels que les produits de matrices et les convolutions [Wikipédia].

- **Transformers** : est une architecture révolutionnaire de réseaux neuronaux largement utilisée dans le domaine du traitement du langage naturel (NLP) et de la vision par ordinateur. Basés sur l'idée de l'attention, Transformers captent les relations à longue distance entre les éléments d'une séquence. Leur capacité à traiter les éléments de manière parallèle améliore l'efficacité et les performances du modèle. Transformers se composent d'encodeurs et de décodeurs qui travaillent ensemble [19].

- **Uvicorn** : est une implémentation de serveur Web ASGI pour Python. Jusqu'à récemment, Python manquait d'une interface serveur/application minimale de bas niveau pour les frameworks asynchrones. La spécification ASGI comble cette lacune et signifie que nous sommes désormais en mesure de commencer à créer un ensemble commun d'outils utilisables dans tous les frameworks asynchrones [51].

- **Numpy** : qui signifie "Numerical Python", est une bibliothèque Python extrêmement populaire conçue principalement pour les calculs mathématiques et scientifiques. Elle fournit une interface permettant de stocker et d'effectuer des opérations sur les données. Les tableaux NumPy sont similaires aux listes Python, mais ils offrent des performances nettement améliorées, en particulier pour les tableaux de grande taille qui jouent un rôle central dans l'écosystème de la Data Science [13].

- **SciPy** : est une bibliothèque de calcul scientifique qui repose sur NumPy. Elle offre une gamme étendue de fonctionnalités pour l'optimisation, les statistiques et le traitement du signal. SciPy, qui signifie Python scientifique, a été développé par Travis Olliphant, le créateur de NumPy lui-même. Tout comme NumPy, SciPy est une bibliothèque open source, ce qui signifie qu'elle est disponible gratuitement et peut être utilisée librement. Avec sa richesse de fonctionnalités et son intégration étroite avec NumPy, SciPy constitue un outil essentiel pour les tâches scientifiques et d'analyse de données en Python [52].

- **Psycopg2** : Psycopg est l'adaptateur de base de données PostgreSQL le plus populaire pour le langage de programmation Python. Ses principales fonctionnalités

sont l'implémentation complète de la spécification Python DB API 2.0 et la sécurité des threads (plusieurs threads peuvent partager la même connexion). Psycopg2 est compatible avec Unicode et Python 3 [42].

Formats de données

- **Json** : acronyme de "JavaScript Object Notation", est un format spécialement conçu pour les types de données du langage JavaScript. Au fil des dernières années, JSON s'est imposé comme l'un des formats les plus utilisés pour l'échange et le stockage de données, en particulier dans le domaine du développement web [40].

Logiciels et éditeurs de textes

Figure 3.3: Logos des logiciels

- **Visual Studio Code** : est un éditeur de code open-source développé par Microsoft supportant un très grand nombre de langages grâce à des extensions. Il supporte l'autocomplétion, la coloration syntaxique, le débogage, et les commandes git [22].
- **Git** : est un système de gestion de versions décentralisé, développé par Linus Torvalds, créateur du noyau Linux. Il est distribué sous les termes de la licence publique générale (GPL) en tant que logiciel libre. En 2016, Git était le logiciel de gestion de versions le plus largement adopté, utilisé par plus de douze millions de personnes [37].
- **GitHub** : est une plateforme open source de gestion de versions et de collaboration destinée aux développeurs de logiciels. Livrée en tant que logiciel à la demande (SaaS, Software as a Service), la solution GitHub a été lancée en 2008. Elle repose sur Git, un système de gestion de code open source créé par Linus Torvalds dans le but d'accélérer le développement logiciel [48].

3.3 Métriques d'évaluation

Afin d'évaluer le système d'indexation sémantique, nous allons évaluer nos modèles en utilisant différentes métriques, telles que la précision, le rappel, la F1-mesure et l'accuracy.

Matrice de confusion

Pour une analyse plus approfondie de la qualité des classes générées par le modèle d'indexation, les tables de confusion peuvent être d'une grande utilité. Une matrice de confusion, aussi connue sous le nom de tableau de contingence, permet d'évaluer la précision d'une classification. Elle est obtenue en comparant les données classées par le modèle avec des données de référence distinctes de celles utilisées pour l'entraînement du modèle. Voici les termes clés associés à une matrice de confusion :

Vrais positifs (TP) : le nombre de cas où le modèle prédit correctement la classe positive.

Faux positifs (FP) : le nombre de cas où le modèle prédit à tort la classe positive.

Faux négatifs (FN) : le nombre de cas où le modèle prédit à tort la classe négative.

Vrais négatifs (TN) : le nombre de cas où le modèle prédit correctement la classe négative.

En analysant ces éléments, nous sommes en mesure d'évaluer plus précisément la performance du modèle de catégorisation, en identifiant les cas où il prédit correctement ou incorrectement chaque classe. Les tables de confusion fournissent donc une vision détaillée de la qualité de la classification réalisée par le modèle.

Accuracy

Cette mesure évalue l'efficacité globale de l'algorithme en comparaison avec les données de test 3.1 :

$$Accuracy = \frac{T_p + T_n}{T_p + T_f + F_p + F_n} \tag{3.1}$$

Précision

Elle évalue la capacité prédictive du modèle en mesurant sa capacité à effectuer des prédictions correctes des classes. 3.2 :

$$Precision = \frac{T_p}{T_p + F_p} \tag{3.2}$$

Rappel

Cette mesure évalue l'efficacité globale de l'algorithme en comparaison avec les données de test 3.3 :

$$rappel = \frac{T_p}{T_p + F_n} \tag{3.3}$$

F1-Mesure

Cette mesure évalue l'efficacité globale de l'algorithme en comparaison avec les données de test 3.4 :

$$F1\text{-}Score = \frac{2 * Precision * Rappel}{Precision + Rappel} \tag{3.4}$$

3.4 Résultats et Tests

Afin de tester la validité de notre modèle et de mettre en œuvre notre approche, nous avons utilisé plusieurs *Datasets*. Parmi ces ensembles, nous avons utilisé AG News [55], DBpedia [3] et NATCAT [7]. Concernant NATCAT, pour le modèle SGPT nous avons entraîné le modèle avec 4 ensembles de données seulement et l'avons testé avec le 5ème, ce qui nous a permis d'obtenir les résultats suivants. De plus, nous avons également exploité les ensembles de données PAWS-X [54] pour le français et XNLI [9] pour l'arabe. Ensuite les résultats trouve seront compare avec XLMRoberta [8] fine-tuned sur XNLI.

3.4.1 NATCAT Dataset

NATCAT est un ensemble de données divisé en trois ensembles : Wikipedia, Stack Exchange et Reddit. Chacun de ces ensembles comprend environ 27 à 28 ensembles d'entraînement. Étant donné la taille importante de NATCAT, qui compte environ 2,7 millions d'échantillons, nous avons choisi de fine-tuner le modèle SGPT uniquement sur 4 ensembles d'entraînement, soit environ 400 000 échantillons, plutôt que d'utiliser l'ensemble complet.

Structure du dataset

L'exemple ci-dessous montrent la structure de la dataset NATCAT .


```
{
    "label": 3,
    "text": "Shelby Farms Shelby Farms, located in Memphis, Shelby County, Tennessee,
    is one of the twenty largest urban parks in the United States . At a size of, it
    covers more than five times the area of Central Park in New York City with .
    Lakes, natural forests, and wetlands provide natural habitats for many smaller
    species close to an urban metropolitan area . ",
    "positive": "tourist attractions in tennessee by count",
    "negative1": "1980 in soviet sport",
    "negative2": "jews and judaism in appalachia",
    "negative3": "hungarian latter day saints",
    "negative4": "office holders in grenada",
    "negative5": "history of agriculture in the united kingdom",
    "negative6": "avaya products",
    "negative7": "blues albums by scottish artists"
}
```


La structure des données a été modifiée afin de permettre l'entraînement du modèle. On ajoute un champ "label" avec une valeur numérique (1 dans le cas positive et 0 dans le cas négative), un champ "texte" contenant le texte du document, et un autre champs contenant le texte "négative ou positive" négatifs, une nouvelle structure a été adoptée.


```
{
    "label": 0,
    "sentence1": "text",
    "sentence2": "negative text"
}
{
    "label": 1,
    "sentence1": "text",
    "sentence2": "positive text"
}
```

Résultats

les résultats obtenu sont les suivants :

Modelés	SGPT	XLM-RoBeRta
Accuracy	0.90	0.90
Rappel	0.60	0.61
Précision	0.52	0.59
f1 Score	0.56	0.60

Table 3.1: Tableau comparative des résultats.

Notre modèle a donné des résultats similaires à XLM-RoBERTa pour l'ensemble de données NatCat. Malgré que notre modelé a été entraîne sure un échelon de la dataset natcat par contre XLM roberta a été entraîner sure toutes la dataset

3.4.2 AG News Dataset

Le dataset AG News est utilisé pour la classification des sujets. Il est construit en sélectionnant les 4 plus grandes classes du corpus d'origine. Chaque classe comprend 30 000 échantillons d'entraînement et 1 900 échantillons de test. Au total, il y a 120 000 échantillons d'entraînement et 7 600 échantillons de test.

Le fichier classes.txt contient une liste de classes correspondant à chaque label. Les fichiers train.csv et test.csv contiennent tous les échantillons d'entraînement sous forme de valeurs séparées par des virgules. Ils sont constitués de 2 colonnes représentant l'indice de classe (0:world,1:Sports,2:Business et 3:Sci/tech) et la colonne du texte.

Structure du dataset

La structure de la dataset AG News est illustrée par l'exemple ci-dessous.

```
{
    "label": 3,
    "text": "New iPad released Just like every other September, this one is no
    different. Apple is planning to release a bigger, heavier, fatter iPad "
}
```

Afin de faciliter l'entraînement du modèle, une modification a été apportée à la structure des données. Cette modification consiste à inclure dans le cas positif un champ "label" avec une valeur numérique 1 , un champ "texte" contenant le texte du document, ainsi qu'un champ supplémentaire contenant l'ancien label correspondant . Dans le cas négatif la valeur du champ label est définit par 0 et le texte et un ancien label dont la valeur qui ne correspond pas a l'exemple.

```
{
    "label": 1,
    "sentence1": "New iPad released Just like every other September, this one is no
    different. Apple is planning to release a bigger, heavier, fatter iPad ",
    "sentence2": "Business"
}
{
    "label": 0,
    "sentence1": "New iPad released Just like every other September, this one is no
    different. Apple is planning to release a bigger, heavier, fatter iPad ",
    "sentence2": "Sports"
}
```

Résultats

les résultats obtenu sont les suivants :

Modelés	SGPT	XLM-RoBeRta
Accuracy	0.54	0.63
Rappel	0.90	0.90
Précision	0.50	0.54
f1 Score	0.66	0.68

Table 3.2: Tableau comparative des résultats.

Cependant, XLM-RoBERTa a surpassé notre modèle pour l'ensemble de données AG News. ce qui est logique car la dataset AG news est destine a être utilise dans les taches de classification et pas la recherche sémantique car elle contiens que 4 requêtes et plus de 120k de documents ce qui rend le taux d'erreur plus élevé .

3.4.3 Paws-x Dataset

Le dataset PAWS-X comprend 23 659 paires d'évaluation de PAWS traduites par des humains et 296 406 paires d'entraînement traduites par des machines. Ces paires sont disponibles dans six langues typologiquement distinctes : français, espagnol, allemand, chinois, japonais et coréen. Toutes les paires traduites proviennent d'exemples de PAWS-Wiki.

Structure du dataset

Cette dataset est bien structurée et nécessite pas de modifications, sa structure est illustrée par l'exemple ci-dessous.

```
{
sentence1 : À Paris, en octobre 1560, il rencontra secrètement l'ambassadeur
d'Angleterre, Nicolas Throckmorton, lui demandant un passeport pour retourner
en Angleterre en passant par l'Écosse.
sentence2 : En octobre 1560, il rencontra secrètement l'ambassadeur d'Angleterre,
Nicolas Throckmorton, à Paris, et lui demanda un passeport pour retourner en Écosse
par l'Angleterre.
label : 0
}
```

Résultats

les résultats obtenu sont les suivants :

Modèles	SGPT	XLM-RoBeRta
Accuracy	0.75	0.63
Rappel	0.82	0.94
Précision	0.63	0.50
f1 Score	0.71	0.65

Table 3.3: Tableau comparative des résultats.

En revanche, notre modèle a surpassé XLM-RoBERTa pour l'ensemble de données PAWS-X.

3.4.4 XNLI Dataset

Le dataset XNLI est un sous-ensemble de quelques milliers d'exemples de MNLI traduits dans 14 langues différentes, dont certaines sont peu disponibles en ressources. Comme pour MNLI, l'objectif de XNLI est de prédire l'implication textuelle entre deux phrases, en les classant comme impliquant, contredisant ou étant neutres l'une par rapport à l'autre.

Structure du dataset

La structure de cette dataset est déjà bien organisée, mais elle nécessite simplement la modification des noms de colonnes. Ainsi, les colonnes devraient être renommées en "sentence1", "sentence2" et "label", plutôt que "hypothesis", "premise" et "label". L'illustration de cette structure est présentée dans l'exemple ci-dessous.

Avant la modification :

{

"hypothesis":أتصل بأمه حالما أوصلته حافلة المدرسية ,

"label": 1,

"premise": آوقال، ماما، لقد عدت للمنزلا

}

Apres la modefication :

{

"Sentence 1": أتصل بأمه حالما أوصلته حافلة المدرسية ,

"Sentence 2": آوقال، ماما، لقد عدت للمنزلا,

"label": 1

}

Résultats

Puisque que XLM-RoBERta a déjà été entraîné sur XNLI, nous allons uniquement présenter les résultats sur SGPT car l'utilisation de cette comparaison aurait probablement conduit à un Overfiting des résultats. les résultats obtenu sont les suivants :

Modèles	SGPT
Accuracy	0.52
Rappel	0.90
Précision	0.50
f1 Score	0.66

Table 3.4: Tableau des résultats.

3.5 Adaptation de la plateforme de partage vidéo

Nous sommes ravis de présenter les dernières évolutions de notre plateforme de partage de vidéos "*VSW*" choisi. Tout d'abord, nous avons décidé de changer le nom de notre plateforme, passant de "*VSW*" à "*Filma*". Ce nouveau nom a été choisi dans le but de rendre notre plateforme plus lisible, accessible et facilement compréhensible pour notre communauté d'utilisateurs passionnés de vidéos.

Au fur et à mesure de l'utilisation de la plateforme "*Filma*", nous avons rencontré des problèmes. Nous avons pris ces problèmes à cœur et nous nous sommes engagés à améliorer continuellement notre plateforme. Par conséquent, nous avons entrepris des efforts considérables pour résoudre ces problèmes et offrir une expérience encore plus agréable et fluide.

Dans cette partie, nous souhaitons présenter les modifications que nous avons apportées à notre plateforme "*Filma*", ainsi que les solutions mises en place pour répondre à nos besoins.

- Dans le login et l'enregistrement : nous avons modifié le filtre utilisé pour les données dans le champ d'entrées car il était obsolète en php version 8.

- Nous avons modifié le script sql en php car dans la dernière version 8, la méthode d'écriture de sql en php a changé, nous avons donc mis à jour toutes les requêtes sql.

- Nous avons résolu le problème d'obtenir un id d'utilisateur après la connexion, car nous en avons besoin dans la page de profil et dans sql pour télécharger les vidéos du même utilisateur.

- Nous avons résolu quelques problèmes dans les codes HTML pour suivre les changements dans les codes php modifiées.

- Nous avons résolu le problème de visibilité du mot de passe dans la base de données avec la fonction md5(), elle est utilisée pour calculer le hachage md5 (le hachage sous forme de nombre hexadécimal de 32 caractères) d'une chaîne.

- Nous avons résolu le problème de suppression de vidéos dans le stockage de notre plateforme. Avant, seules les informations relatives à la vidéo étaient supprimées de la base de données, mais la vidéo elle-même restait dans le stockage.

- Nous avons résolu le problème de téléchargement de vidéos. Auparavant, il était impossible de télécharger des vidéos de grande taille de plus de 2 Mo et d'une durée approximative de 40 secondes. Maintenant, nous sommes en mesure de télécharger des vidéos de grande taille allant jusqu'à 256 GB. Et même la taille du titre vidéo, de 70 caractères à 100 caractères, et la taille du description vidéo doit comporter moins de 32500 caractères, conformément aux normes de YouTube.

- Nous avons corrigé le problème de suppression des commentaires par la personne qui partage la vidéo uniquement. Nous avons modifié cela en permettant à la personne qui a écrit le commentaire de le supprimer uniquement.

- Nous avons effectué la traduction de la plateforme de l'anglais vers le français.

- Nous avons changé notre système de gestion de base de données relationnelle de MySQL à PostgreSQL. En effet, MySQL est généralement reconnu pour être plus rapide avec des commandes en lecture seule, mais cela se fait au détriment de la concurrence. D'un autre côté, PostgreSQL est plus efficace pour les opérations de lecture-écriture, les ensembles de données volumineux et les requêtes complexes.

- À la fin, nous avons ajouté l'admin panel, offrant à l'administrateur la possibilité de supprimer des vidéos ou des commentaires qui ne respectent pas les règles de la plateforme.

3.6 Intégration du modéle SGPT au plateforme "*Filma*"

L'intégration du modèle SGPT se fait en utilisant le framework FastAPI pour créer une API, que l'on peut ensuite appeler dans la plateforme "*Filma*".

Dans cette partie, nous souhaitons présenter quelques fonctions utilisées pour créer cette API, ainsi que la manière dont nous l'avons intégrée dans notre plateforme.

Au début, il est nécessaire de télécharger le modèle SGPT dans le code afin de pouvoir l'utiliser dans les fonctions par la suite (figure 3.4).

```python
def get_model():
    tokenizer = AutoTokenizer.from_pretrained("bounedjarr/sgpt-finetuned-natcat")
    model = AutoModel.from_pretrained("bounedjarr/sgpt-finetuned-natcat")
    return tokenizer,model

tokenizer, model = get_model()
```

Figure 3.4: Fonction de chargement du modèle

Nous avons deux fonctionnalités essentielles dans cette API. La première consiste à extraire les informations pertinentes à partir du titre et de la description de la vidéo, et la deuxième fonctionnalité permet de faire une recherche sémantique sur un texte donné.

Pour la première fonctionnalité, nous avons créé une méthode appelée "get-embedding()" dans le point de terminaison suivant :

"`http://127.0.0.1:8000/embedding/text?id=2&title=testTitre&desc=testDesc`"

Cette méthode (figure 3.5) prend l'ID, le texte et la description de la vidéo. Ensuite, elle concatène le texte et la description, extrait les informations pertinentes à partir de ceux-ci, puis les enregistre dans un vecteur d'index au format JSON. Enfin, les résultats sont sauvegardés dans la base de données.

```python
@app.post("/embedding/text")
async def get_embedding(id: int, title: str , desc: str):
    text = title.lower() + ' . ' + desc.lower()
    doc = []
    doc.append(text)
    embedding = get_weightedmean_embedding(tokenize_with_specb(doc, is_queries=False),
model)
    x_np = embedding.numpy()
    x_str = json.dumps(x_np.tolist())
    sql = "INSERT INTO embedding (video_id, video_text, video_embedding) VALUES (%s,%s,%s)"
    val = (id, text, x_str)
    mycursor.execute(sql,val)
    dbcon.commit()
    return {"status" : "success"}
```

Figure 3.5: Fonction d'extraire les informations pertinentes

Pour la deuxiéme fonctionnalité, nous avons créé une méthode appelée "get-search()" dans le point de terminaison suivant :

"`http://127.0.0.1:8000/search?text=searchText`"

Mais d'abord, il est nécessaire de récupérer tous les vecteurs d'indexation qui sont sauvegardés dans la base de données. Pour cela, nous avons créé la méthode présentée dans la figure 3.6.

```python
def get_all_docs():
    sql = "SELECT video_id,video_embedding FROM embedding"
    mycursor.execute(sql)
    result = mycursor.fetchall()
    return result
```

Figure 3.6: Fonction d'obtenir tous les vecteurs d'indexation

Ensuite, nous utilisons les vecteurs d'indexation obtenus dans la méthode (figure 3.7), ainsi que le vecteur d'indexation extrait à partir du texte de recherche pour calculer la similarité cosinus entre ces vecteurs. Cela nous permet de retourner les IDs des vidéos correspondantes en résultat.

```python
@app.get("/search")
def get_search(text:str):
    ids = []
    docs = []
    queries = []
    queries.append(text.lower())
    queries_embeddings = get_weightedmean_embedding(tokenize_with_specb(queries, is_queries=True),
    model)
    ids, docs = get_fixed_docs()
    doc_embeddings = docs
    # Calculate cosine similarities
    # Cosine similarities are in [-1, 1]. Higher means more similar
    results =assign_scores_to_docs(ids, get_cosine_similarities(queries_embeddings, doc_embeddings))
    x= sorted(results, key=lambda x: list(x.values())[0], reverse=True)
    k = extract_ids(x)
    return k
```

Figure 3.7: Fonction de recherche sémantique

Maintenant nous disposons d'une API avec deux points de terminaison, ce qui nous permet de l'utiliser facilement dans notre plateforme "*Filma*".

Dans la plateforme "*Filma*", nous avons utilisé la fonction cURL de PHP, qui signifie "Client URL" [41]. Cette fonction nous permet d'effectuer des requêtes HTTP en PHP, ce qui nous permet d'interagir avec notre API et d'envoyer des données ou de récupérer des résultats.

La figure 3.8 présente la première utilisation de l'API, qui consiste à extraire les informations pertinentes depuis le titre et la description envoyés, et à les sauvegarder dans la base de données.

```php
$url = 'http://127.0.0.1:8000/embedding/text?
id='.$video_id['id'].'&title='.urlencode($title).'&desc='.urlencode($description);
$r = curl_init($url);
curl_setopt($r, CURLOPT_POST, true);
curl_setopt($r, CURLOPT_RETURNTRANSFER, true);
$response = curl_exec($r);
curl_close($r);
```

Figure 3.8: Utilisation de première point de terminaison API

La figure 3.9 présente la deuxiéme utilisation de l'API, qui consiste à effectuer une recherche sémantique à partir du texte donné.

```php
    $url = 'http://127.0.0.1:8000/search?text='.urlencode($search);
    $r = curl_init($url);
    curl_setopt($r, CURLOPT_RETURNTRANSFER, true);
    $response = curl_exec($r);
    curl_close($r);

    $arr = json_decode($response);
    $search_list = implode(",", $arr);

    $q = $db->prepare("SELECT videos.id as vid, title, description, id_owner, thumbnail_loca, date,
users.id, username, fullname FROM videos JOIN users ON videos.id_owner=users.id WHERE videos.id IN
(SELECT unnest(ARRAY[$search_list])) ORDER BY POSITION(videos.id::text in '$search_list')");

    $q->execute();
    $searchResult = $q->fetchALL(PDO::FETCH_ASSOC);
```

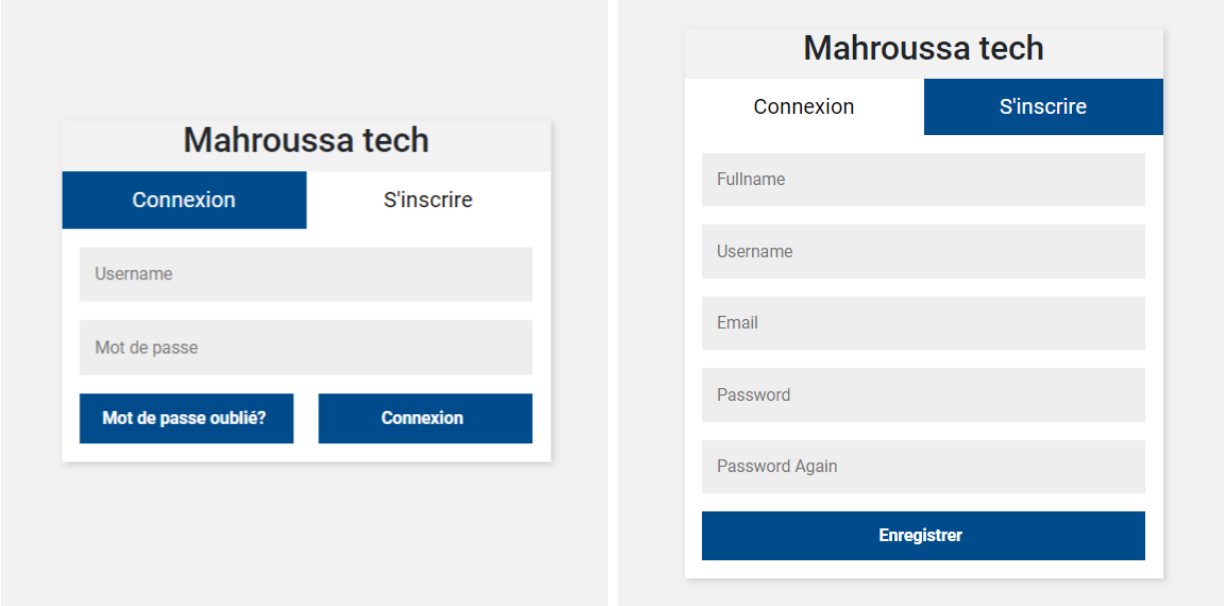
Figure 3.9: Utilisation de deuxième point de terminaison API

3.7 Interface graphique

Notre interface est composée principalement de 7 pages :

Authentification

L'utilisateur a la possibilité de créer un compte pour s'identifier ou de se connecter directement s'il en possède déjà un. Les données d'authentification sont enregistrées dans la collection des utilisateurs, présente à la fois dans la base de données de la principale plateforme d'archivage "*Mahroussa tech*" et dans la plateforme "*Filma*".



(a) La connexion (b) La création du compte

Figure 3.10: La page de l'authentification dans la plateforme d'archive principale

Choix des plateformes

Après la connexion, l'utilisateur peut choisir et accéder à toutes les plateformes reliées à notre plateforme d'archives principale avec le même compte d'utilisateur.

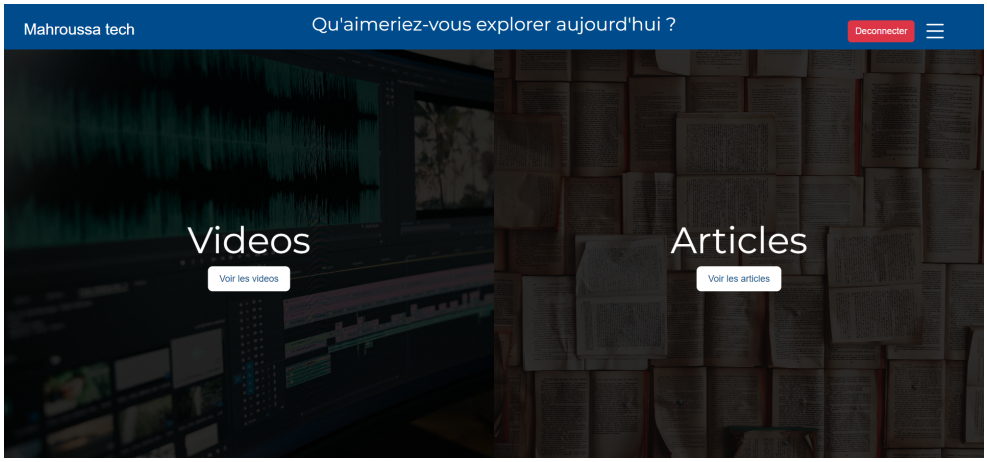


Figure 3.11: La page de choix des plateformes

Accueil de "*Filma*"

L'interface principale de "*Filma*" affiche toutes les vidéos existantes sur notre plateforme, dans l'ordre de leur date de partage. L'utilisateur peut regarder une vidéo, rechercher une vidéo, partager une vidéo ou consulter les vidéos qu'il a partagées dans son profil.

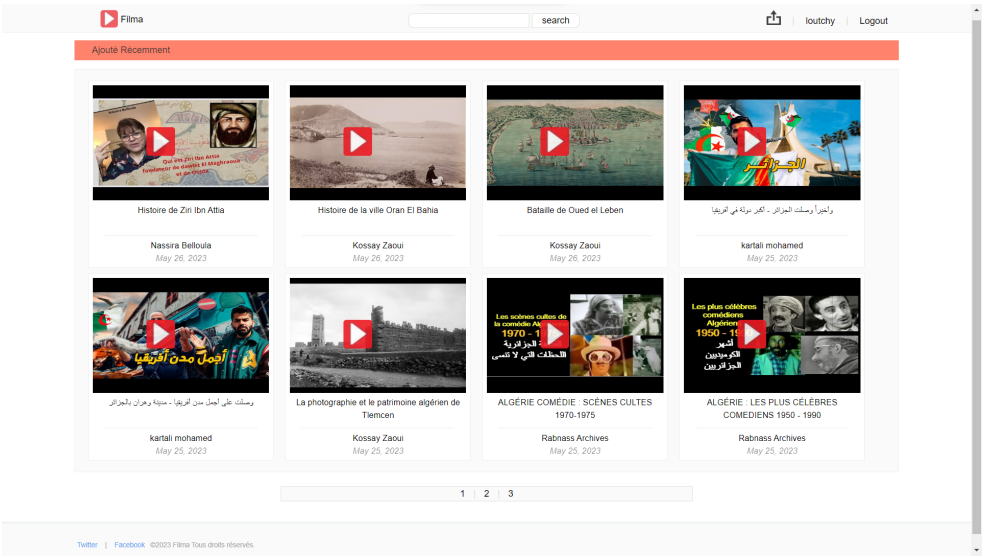


Figure 3.12: La page d'accueil de "*Filma*"

Partager une vidéo

Dans cette page, l'utilisateur doit saisir toutes les informations nécessaires de sa vidéo afin de pouvoir la partager sur notre plateforme.

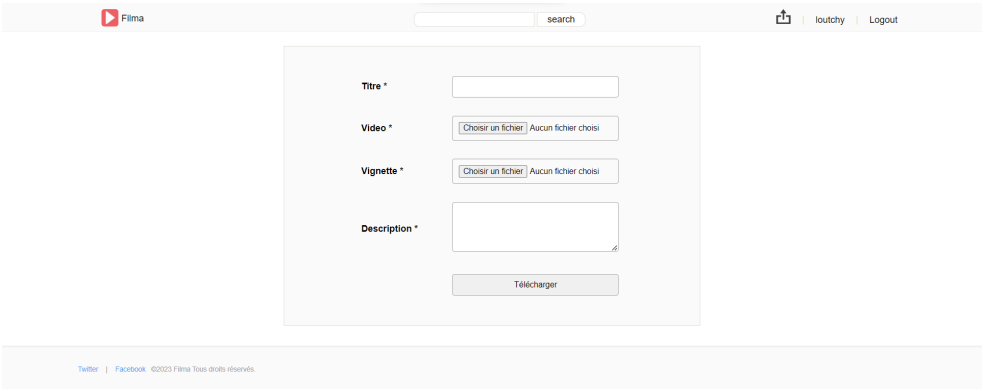


Figure 3.13: La page de partager une vidéo

Recherche des vidéos

Après la saisie du texte pour la recherche, nous affichons toutes les vidéos pertinentes à la recherche de l'utilisateur.

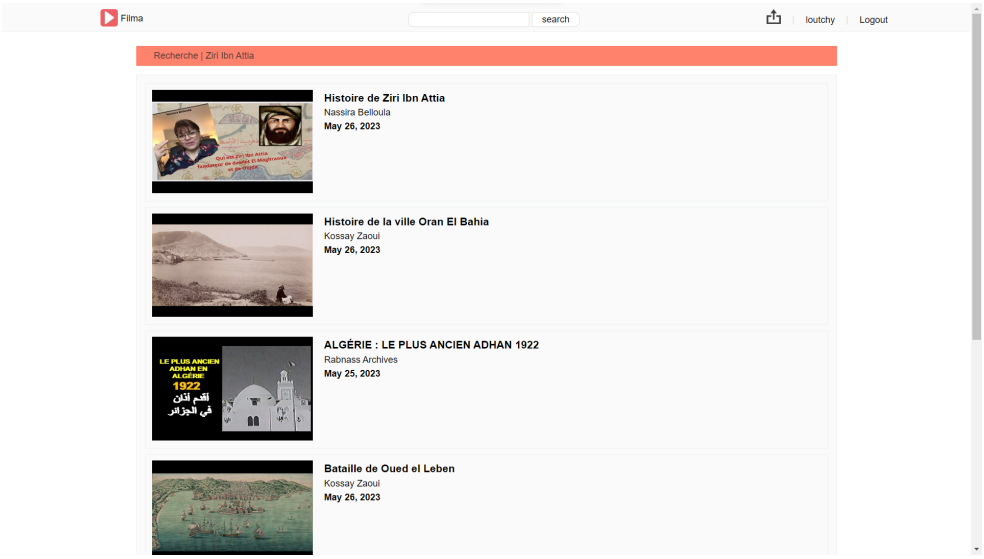


Figure 3.14: La page de recherche des vidéos

Regarder une vidéo

Si l'utilisateur souhaite regarder une vidéo, il peut également voir toutes les informations associées à cette vidéo, telles que le nom de l'auteur, la date de partage, le titre et la description. Il peut également voter pour la vidéo ou commenter.

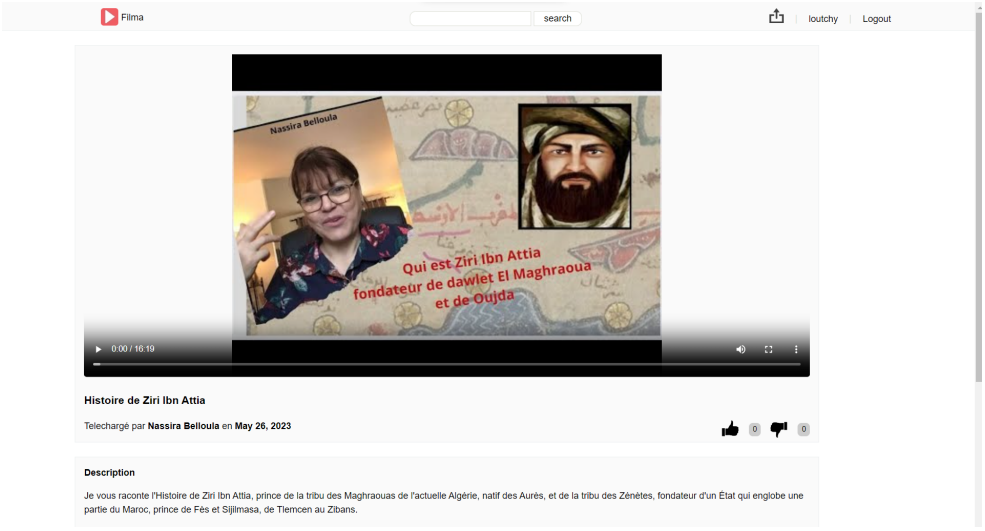


Figure 3.15: La page de regarder une vidéo

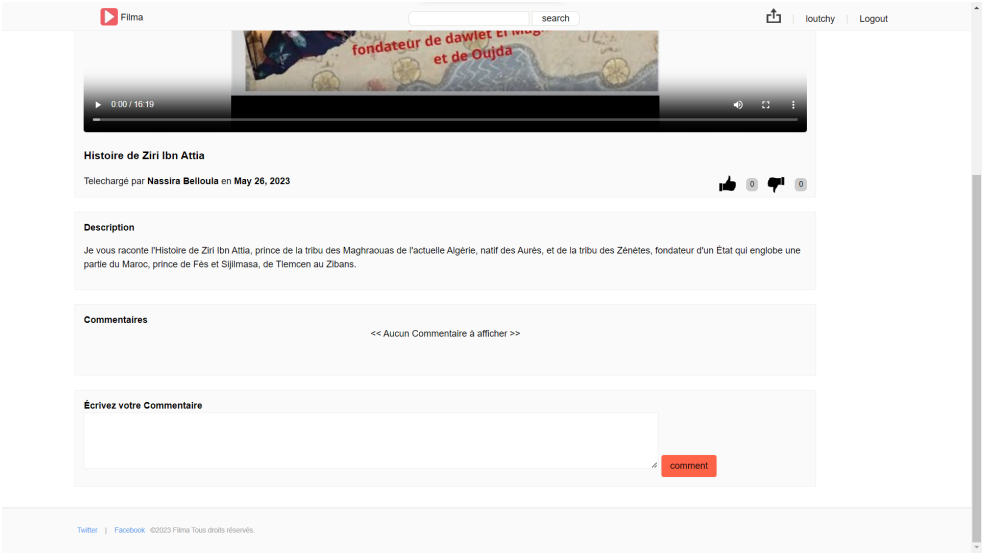


Figure 3.16: La suite du page pour voter ou commenter une vidéo

Profile utilisateur

Dans la page de profil de l'utilisateur, nous souhaitons afficher toutes ses vidéos, et il a également la possibilité de supprimer une vidéo.

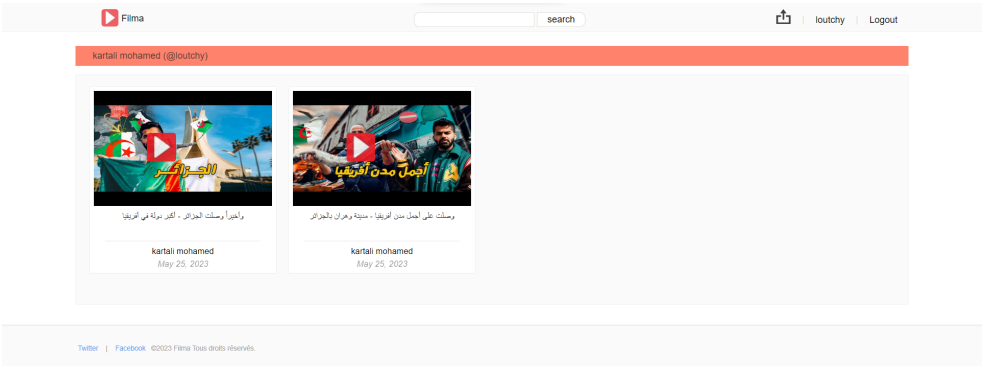


Figure 3.17: La page de profile utilisateur

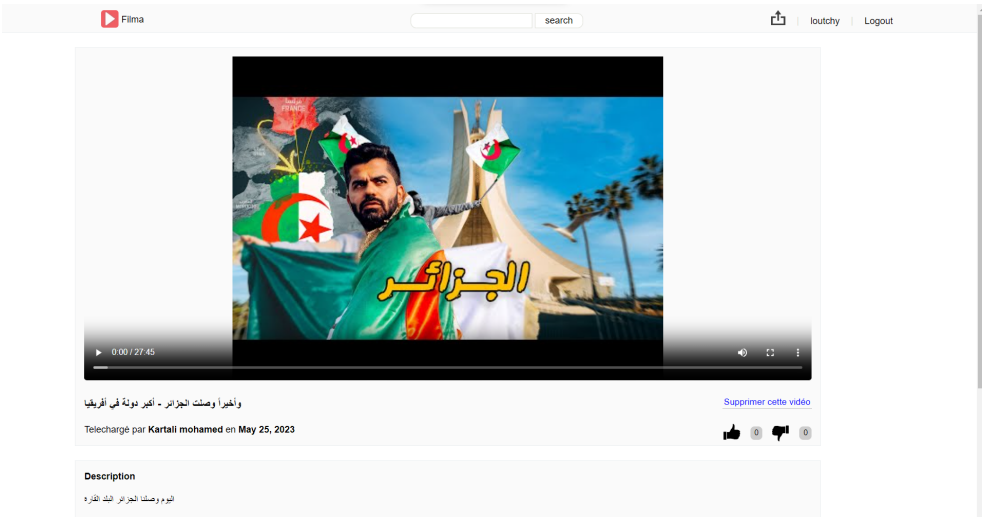


Figure 3.18: Le détails de la vidéo de l'utilisateur

3.8 Conclusion

Dans ce chapitre, nous avons commencé par présenter l'environnement de travail et les outils utilisés. Ensuite, nous avons abordé les évaluations en utilisant les métriques appropriées, ainsi que les résultats et les tests obtenus. Enfin, nous avons discuté de l'adaptation de la plateforme de partage vidéo "*Filma*", de l'intégration des modèles SGPT sur cette plateforme, ainsi que de leurs différentes interfaces.

Conclusion générale

Aujourd'hui, les internautes en général, et les chercheurs et historiens en particulier, font face quotidiennement à un flux massif d'informations et à une grande diversité de contenus, malgré leur besoin spécifique de trouver du contenu algérien uniquement. Les plateformes de partage de vidéos telles que YouTube contribuent largement à cette situation. Dans un environnement où les utilisateurs ne disposent toujours pas de moyens satisfaisants pour gérer ce flux, un nouveau besoin émerge : celui d'une expérience de lecture algérienne personnalisée, correspondant de manière pertinente et efficace aux besoins et aux centres d'intérêt de chaque utilisateur

L'objectif principal de ce projet était de développer une plateforme offrant aux lecteurs, chercheurs et historiens algériens la possibilité de mener leurs recherches sur l'histoire et les traductions algériennes de manière plus efficace et enrichissante. Nous aspirions à apporter une contribution significative et à ajouter de la valeur au patrimoine culturel et traditionnel de l'Algérie, en le préservant soigneusement.

Ce mémoire se concentre sur l'analyse de la littérature existante portant sur l'indexation automatique. Par la suite, une recherche approfondie a été entreprise afin d'explorer l'état actuel du l'indexation automatique sémantique. Cette recherche a permis d'identifier les travaux les plus fiables, les différentes techniques utilisées et les outils les plus performants dans ce domaine.

Une période prolongée du projet a été consacrée à la compréhension de ces différentes techniques, puis à la sélection d'une méthode spécifique pour l'amélioration et l'utilisation dans notre projet.

Ensuite, il a été nécessaire d'adapter le modèle pré-entraîné choisi à une tâche spécifique en effectuant un fine-tuning sur différents ensembles de données (datasets), en exploitant les connaissances préalables du modèle.

Aussi, sur la plateforme de partage de vidéos choisie, nous avons effectué plusieurs adaptations et corrigé les erreurs afin d'offrir une expérience encore plus agréable et fluide

à l'utilisateur.

La dernière étape consiste à déployer le modèle et à l'intégrer à notre plateforme afin d'atteindre notre objectif d'une plateforme qui permettra d'indexer automatiquement les vidéos et offrira aux utilisateurs la possibilité de partager et de rechercher des contenus vidéo locaux.

Les résultats obtenus pour l'indexation et la recherche sémantique des vidéos ont été hautement satisfaisants en comparaison avec d'autres modèles et travaux similaires. Malgré une période d'entraînement limitée en raison de contraintes de ressources et de performances, le modèle de recherche sémantique a démontré son efficacité en atteignant un taux de précision de 71% en français, 66% en arabe et 90% en anglais. Nous avons également obtenu des résultats remarquables dans les trois langues étudiées.

Cependant, il est essentiel de souligner que la solution proposée présente encore des possibilités d'amélioration significatives. Plusieurs aspects nécessitent une révision. Une perspective intéressante serait d'effectuer une extraction de texte à partir de toutes les vidéos en utilisant la reconnaissance automatique de la parole. Cela permettrait de mieux comprendre le thème et le contexte de chaque vidéo. De plus, l'indexation ne se limiterait pas seulement au titre et à la description, mais prendrait également en compte le contexte, le thème, les titres, les descriptions et les hashtags de chaque vidéo.

Une autre perspective intéressante serait de former le modèle sur la langue algérienne darija, afin d'améliorer l'efficacité de la plateforme et de répondre aux spécificités linguistiques du peuple algérien. Cette approche nécessiterait également la création d'une dataset en darija algérien.

En ce qui concerne notre site, afin d'attirer plus d'utilisateurs et de rendre la plateforme plus accessible, il serait nécessaire de développer une application mobile pour notre solution.

Références

[1] République Algerienne. Portail de patrimoine culturel algerien. `http://www.patrimoineculturelalgerien.com/index.php`. (Accessed on janvier 12, 2023).

[2] Alex Andonian, Quentin Anthony, Stella Biderman, Sid Black, Preetham Gali, Leo Gao, Eric Hallahan, Josh Levy-Kramer, Connor Leahy, Lucas Nestler, Kip Parker, Michael Pieler, Shivanshu Purohit, Tri Songz, Phil Wang, and Samuel Weinbach. Gpt-neox: Large scale autoregressive language modeling in pytorch. *arXiv preprint arXiv:2110.11327*, 2021.

[3] Sören Auer, Christian Bizer, Georgi Kobilarov, Jens Lehmann, Richard Cyganiak, and Zachary Ives. Dbpedia: A nucleus for a web of open data. In Karl Aberer, Key-Sun Choi, Natasha Noy, Dean Allemang, Kyung-Il Lee, Lyndon Nixon, Jennifer Golbeck, Peter Mika, Diana Maynard, Riichiro Mizoguchi, Guus Schreiber, and Philippe Cudré-Mauroux, editors, *The Semantic Web*, pages 722–735, Berlin, Heidelberg, 2007. Springer Berlin Heidelberg.

[4] Debabrota Basu, Qian Lin, Weidong Chen, Hoang Tam Vo, Zihong Yuan, Pierre Senellart, and Stéphane Bressan. Cost-model oblivious database tuning with reinforcement learning. In *Database and Expert Systems Applications: 26th International Conference, DEXA 2015, Valencia, Spain, September 1-4, 2015, Proceedings, Part I 26*, pages 253–268. Springer, 2015.

[5] Elad Ben Zaken, Shauli Ravfogel, and Yoav Goldberg. Bitfit: Simple parameter-efficient fine-tuning for transformer-based masked language-models. *arXiv preprint arXiv:2106.10199*, 2021.

[6] Alex Beutel, Tim Kraska, E Chi, Jeffrey Dean, and Neoklis Polyzotis. A machine learning approach to databases indexes. In *Proceedings of the ML Systems Workshop at NIPS*, 2017.

[7] Zewei Chu, Karl Stratos, and Kevin Gimpel. Natcat: Weakly supervised text classification with naturally annotated datasets, 2020.

[8] Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzman, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. Unsupervised cross-lingual representation learning at scale. pages 8440–8451, 01 2020.

[9] Alexis Conneau, Ruty Rinott, Guillaume Lample, Adina Williams, Samuel R. Bowman, Holger Schwenk, and Veselin Stoyanov. Xnli: Evaluating cross-lingual sentence representations. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 2018.

[10] DataReportal. Social media in the workplace. `https://datareportal.com/reports/digital-2023-april-global-statshot#:~:text=Social%20media%20in%20the%20workplace`.

[11] DataReportal. Youtube users in algeria in 2023. `https://datareportal.com/reports/digital-2023-algeria#:~:text=YouTube%20users%20in%20Algeria%20in,in%20Algeria%20in%20early%202023`.

[12] DataScientest. Fastapi : tout savoir sur le framework web python le plus utilisé pour le machine learning. `https://datascientest.com/fastapi`. (Accessed on avril 3, 2023).

[13] DataScientest. Numpy : la bibliothèque python la plus utilisée en data science. `https://datascientest.com/numpy`.

[14] Association des archivistes français. BibliothÈque d'archives. `https://www.archivistes.org/-Bibliotheque-d-archives-BA-`.

[15] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

[16] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Improving language understanding by generative pre-training. *arXiv preprint arXiv:1810.04805*, 2018.

[17] Bailu Ding, Sudipto Das, Ryan Marcus, Wentao Wu, Surajit Chaudhuri, and Vivek R Narasayya. Ai meets ai: Leveraging query executions to improve index recommendations. In *Proceedings of the 2019 International Conference on Management of Data*, pages 1241–1258, 2019.

[18] DZ Entreprise. data reportal : le nombre d'internautes en algérie a augmenté en 2023. `https://www.dzentreprise.net/data-reportal-internautes-algerie/`, 2023.

[19] Hugging Face. les transformers. `https://huggingface.co/learn/nlp-course/fr`.

[20] Bin Fan, Dave G Andersen, Michael Kaminsky, and Michael D Mitzenmacher. Cuckoo filter: Practically better than bloom. In *Proceedings of the 10th ACM International on Conference on emerging Networking Experiments and Technologies*, pages 75–88, 2014.

[21] Thibault Formal, Carlos Lassance, Benjamin Piwowarski, and Stéphane Clinchant. Splade v2: Sparse lexical and expansion model for information retrieval. *arXiv preprint arXiv:2109.10086*, 2021.

[22] Framalibre. Visual studio code. `https://framalibre.org/content/visual-studio-code`.

[23] BIBLIOTHÈQUE NATIONALE DE FRANCE. Archives de l'internet. `https://www.bnf.fr/fr/archives-de-linternet`.

[24] République Française. Francearchives portail national des archives. `https://francearchives.gouv.fr/fr/`.

[25] J. Gabay and D. Gabay. *UML 2 Analyse et conception: Mise en oeuvre guidée avec études de cas*. UML. Dunod, 2008.

[26] Goetz Graefe and P-A Larson. B-tree indexes and cpu caches. In *Proceedings 17th International Conference on Data Engineering*, pages 349–358. IEEE, 2001.

[27] The PHP Group. Qu'est ce que php? `https://www.php.net/manual/fr/intro-whatis.php`.

[28] Kaggle. How to use kaggle. `https://www.kaggle.com/docs/notebooks`.

[29] Hai Lan, Zhifeng Bao, and Yuwei Peng. An index advisor using deep reinforcement learning. In *Proceedings of the 29th ACM International Conference on Information & Knowledge Management*, pages 2105–2108, New York, NY, USA, 2020. ACM.

[30] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*, 2019.

[31] Krisha Mehta. A machine learning approach to databases indexes. Medium, August 2021.

[32] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*, 2013.

[33] Bhaskar Mitra, Matthew Richardson, Nicolas A. Kraft, and Nick Craswell. BEIR: A heterogeneous benchmark for zero-shot evaluation of information retrieval models. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, July 2020.

[34] Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Alex Graves, Ioannis Antonoglou, Daan Wierstra, and Martin Riedmiller. Playing atari with deep reinforcement learning. *Nature*, 518(7540):529–533, 2015.

[35] Niklas Muennighoff. Sgpt: Gpt sentence embeddings for semantic search. *arXiv preprint arXiv:2202.08904*, 2022.

[36] Pandu Nayak. Understanding searches better than ever before. Google Blog, 2019.

[37] Marius Nestor. Git 2.8.2 popular source code management system released with over 18 bug fixes. `https://news.softpedia.com/news/git-2-8-2-popular-source-code-management-system-released-with-over-`.

[38] Gabriel Paludo Licks, Julia Colleoni Couto, Priscilla de Fátima Miehe, Renata De Paris, Duncan Dubugras Ruiz, and Felipe Meneguzzi. Smartix: A database indexing agent based on reinforcement learning. *Applied Intelligence*, 50:2575–2588, 2020.

[39] Jeffrey Pennington, Richard Socher, and Christopher D. Manning. Glove: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2014.

[40] Felipe Pezoa, Juan L Reutter, Fernando Suarez, Martín Ugarte, and Domagoj Vrgoč. Foundations of json schema. In *Proceedings of the 25th international conference on World Wide Web*, pages 263–273, 2016.

[41] PHP. Fonctions curl. https://www.php.net/manual/fr/ref.curl.php.

[42] Psycopg. Psycopg – adaptateur de base de données postgresql pour python. https://www.psycopg.org/docs/.

[43] Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. Improving language understanding by generative pre-training. *URL https://s3-us-west-2. amazonaws. com/openai-assets/research-covers/language-unsupervised/language$_u$nderstanding$_p$aper.pdf*, 2018.

[44] Nils Reimers and Iryna Gurevych. Sentence-bert: Sentence embeddings using siamese bert-networks. *arXiv preprint arXiv:1908.10084*, 2019.

[45] Zahra Sadri, Le Gruenwald, and Eleazar Lead. Drlindex: Deep reinforcement learning index advisor for a cluster database. In *Proceedings of the 24th Symposium on International Database Engineering & Applications*, New York, NY, USA, 2020. ACM.

[46] Vishal Sharma and Curtis Dyreson. Indexer++: Workload-aware online index tuning with transformers and reinforcement learning. In *Proceedings of the 37th ACM/SIGAPP Symposium on Applied Computing*, SAC '22, page 372–380, New York, NY, USA, 2022. Association for Computing Machinery.

[47] Vishal Sharma, Curtis Dyreson, and Nicholas Flann. Mantis: Multiple type and attribute index selection using deep reinforcement learning. In *Proceedings of the 25th International Database Engineering Applications Symposium*, IDEAS '21, page 56–64, New York, NY, USA, 2021. Association for Computing Machinery.

[48] La Rédaction TechTarget. DÉfinition github. https://www.lemagit.fr/definition/GitHub.

[49] TEMPEST. Digital archive platform. Retrieved from https://digitalpreservation.sk/en/#advantages. (Accessed on June 15, 2023).

[50] Nandan Thakur, Nils Reimers, Johannes Daxenberger, and Iryna Gurevych. Augmented sbert: Data augmentation method for improving bi-encoders for pairwise sentence scoring tasks. *arXiv preprint arXiv:2010.08240*, 2020.

[51] Uvicorn. Un serveur web asgi, pour python. https://www.uvicorn.org/.

[52] w3schools. Qu'est-ce que scipy ? https://www.w3schools.com/python/scipy/scipy_intro.php#.

[53] WIKILIVRES. Programmation python/introduction. https://fr.wikibooks.org/wiki/Programmation_Python/Introduction. (Accessed on dec 28, 2022).

[54] Yinfei Yang, Yuan Zhang, Chris Tar, and Jason Baldridge. PAWS-X: A Cross-lingual Adversarial Dataset for Paraphrase Identification. In *Proc. of EMNLP*, 2019.

[55] Xiang Zhang, Junbo Jake Zhao, and Yann LeCun. Character-level convolutional networks for text classification. In *NIPS*, 2015.

[56] Jeffrey Zhu, Mingqin Li, Jason Li, and Cassandra Oduola. Bing delivers more contextualized search using quantized transformer inference on NVIDIA GPUs in Azure. In *Proceedings of the 2021 ACM International Conference on Conference on Information and Knowledge Management*, CIKM '21, New York, NY, USA, 2021. ACM.

[57] Zippia. 35+ youtube statistics [2023]: How popular is youtube in 2023? https://www.zippia.com/advice/youtube-statistics/, Mar. 15, 2023.