

|  |
| --- |
|  |
| EXPLAIN : PATENTCLASSIFIER |
|  |
| Présenter par :  BENOIT NZIENGUI MOUANDA  YANNICK BISTON DOSSA  MADELEINE LESLIE TAYA TIAKOU  FRANCK NGAKOU  RAPH ANDRE TAKAM TENE |

SOMMAIRE

Introduction …………………………………………………………………………………………..3

Documentation et Terminologie ………………………………………………………………………4

2.1 Documents de Référence …………………………………………………………………………4

2.2 Glossaire ………………………………………………………………………………………….4

Description du Projet ………………………………………………………………………………….5

3.1 Origine du Projet ………………………………………………………………………………….6

3.2 État de l'Art ……………………………………………………………………………………….6

3.3 Description du Produit/Solution ………………………………………………………………….8

3.4 Nature Innovante de la Technologie/Solution ……………………………………………………9

3.5 Liberté d'Exploitation et Risque de Contrefaçon …………………………………………………10

3.6 Aspect Réglementaire ……………………………………………………………………………11

3.7 Étude de Marché et Création de Valeur …………………………………………………………11

3.8 Positionnement sur le Marché et Avantages Concurrentiels …………………………………….11

3.9 Programme de Travail et Dépenses ……………………………………………………………..12

3.10 Utilisation Future du Produit/Solution………………………………………………………….15

3.11 Durée de Vie du Produit/Solution ………………………………………………………………15

Fonction et Contribution des Acteurs ……………………………………………………………….16

4.1 Contexte………………………………………………………………………………………….16

4.2 Fonction et Contribution de Chaque Acteur …………………………………………………….16

4.3 Motivation et Engagement Personnel ……………………………………………………………17

4.4 Méthodologie de Gestion de Projet ………………………………………………………………17

4.5 Chronologie du Projet ……………………………………………………………………………17

Études de Faisabilité Technique ……………………………………………………………………..18

5.1 Environnement Physique …………………………………………………………………………18

5.2 Environnement Matériel…………………………………………………………………………18

5.3 Environnement Logiciel …………………………………………………………………………18

5.4 Données …………………………………………………………………………………………19

Suite de Tests

6.1 Exigence 1 : Précision de la Classification ………………………………………………………19

6.2 Exigence 2 : Explicabilité des Résultats ………………………………………………………….20

6.3 Exigence 3 : Performance de l'Interface Utilisateur ……………………………………………...20

Conclusions et Perspectives d'Amélioration …………………………………………………………21

**INTRODUCTION**

**Présentation du Projet**

L'objectif de notre projet est de développer une méthode de classification de texte performante et explicable, en utilisant des brevets comme base de données pour démontrer la preuve de concept. Le modèle doit atteindre un F1 score élevé soit 0.85 sur des jeux de données inconnus tout en fournissant des explications claires sur les segments de texte influents. Cela facilitera l'adoption de la solution dans des contextes nécessitant une prise de décision transparente, comme le juridique et la recherche scientifique.

Notre projet consiste à utiliser BERT pour la classification des brevets et à intégrer des techniques d'explicabilité telles que SHAP pour expliquer les décisions prises par le modèle.

Ce document détaille les objectifs, la méthodologie, les résultats et les conclusions du projet. Un glossaire et des documents de référence sont fournis pour aider à la compréhension de ce rapport.

**Contexte/Enjeux**

Le développement rapide des technologies et la montée en puissance des systèmes d'intelligence artificielle ont révolutionné de nombreux domaines, y compris la classification de textes. Dans le domaine de la recherche et de l'innovation technologique, la classification des brevets joue un rôle crucial. Les brevets contiennent des informations détaillées sur les inventions et sont classés selon un système de codes CPC (Classification des brevets coopérative). Ces codes aident à organiser et à rechercher des informations spécifiques sur les brevets, facilitant ainsi l'analyse et le développement technologique.

Cependant, les modèles de classification de texte actuels, tels que BERT et XGBoost, bien qu'efficaces, sont souvent considérés comme des "boîtes noires". Leur complexité rend difficile l'explication de leurs décisions, ce qui est crucial pour la transparence et l'adoption de ces modèles dans des domaines critiques tels que le juridique et la recherche scientifique.

**Résultats**

L'intégration des techniques d'explicabilité dans le modèle de classification a permis de fournir des explications claires sur les segments de texte influents, augmentant ainsi la transparence et la confiance des utilisateurs dans les décisions prises par le modèle. Le modèle développé a montré des performances impressionnantes avec un F1 score élevé, tout en offrant des capacités explicatives robustes et compréhensibles.

Le projet EXPLAIN a démontré la faisabilité d'une classification de brevets à la fois rapide, assez performante et explicable. Le modèle BERT, combiné avec SHAP, permet de fournir des explications claires sur les décisions de classification, facilitant ainsi l'adoption de la solution dans des contextes nécessitant une prise de décision transparente.

**2. Documentation et Terminologie**

**2.1 Documents de Référence**

Les documents de référence sont essentiels pour comprendre le cadre, les objectifs et la méthodologie du projet. Ils fournissent des informations de base et des détails techniques nécessaires pour le bon déroulement du projet. Voici les documents de référence utilisés pour ce projet :

|  |  |  |
| --- | --- | --- |
| Numéro de Document | Attaché ? | Application |
| Doc1 : Algorithmes de Classification (2022 Pujari et al - THMM - Patent classification.pdf) | Non | Fournit une vue détaillée des algorithmes de classification utilisés, en particulier BERT, et explique leur fonctionnement interne. Ce document est crucial pour les développeurs et les ingénieurs en charge de la mise en œuvre des modèles. |
| **Doc2 :** Guide de l'utilisateur | Non | Guide de l'utilisateur pour l'interface développée. Ce document aide les utilisateurs finaux à naviguer et utiliser efficacement l'interface de classification des brevets. |
| **Doc3 :** Explicabilité LIME et SHAP (2023 Shajalal et al - Explainable- Patent classification.pdf) | Non | Documentation sur les techniques d'explicabilité LIME et SHAP. Ce document est essentiel pour comprendre comment les explications des modèles sont générées et interprétées. |
| **Doc4 :** Études de Cas (2024 Shomee et al - Review patent classification.pdf) | Non | Études de cas sur la classification des brevets. Ce document fournit des exemples pratiques et des analyses des implications de la classification des brevets, utile pour comprendre les applications réelles de notre solution. |

Chaque document joue un rôle spécifique par rapport à ce projet :

* **Doc1 :** Fournit une vue détaillée des algorithmes de classification utilisés, en particulier BERT, et explique leur fonctionnement interne. Ce document est crucial pour les développeurs et les ingénieurs en charge de la mise en œuvre des modèles. *(2022 Pujari et al - THMM - Patent classification.pdf)*
* **Doc2 :** Guide de l'utilisateur pour l'interface développée. Ce document aide les utilisateurs finaux à naviguer et utiliser efficacement l'interface de classification des brevets.
* **Doc3 :** Documentation sur les techniques d'explicabilité LIME et SHAP. Ce document est essentiel pour comprendre comment les explications des modèles sont générées et interprétées. *(2023 Shajalal et al - Explainable- Patent classification.pdf)*
* **Doc4 :** Études de cas sur la classification des brevets. Ce document fournit des exemples pratiques et des analyses des implications de la classification des brevets, utile pour comprendre les applications réelles de notre solution. *(2024 Shomee et al - Review patent classification.pdf)*

**2.2 Glossaire**

Le glossaire est une section essentielle qui définit les termes et les acronymes utilisés dans ce document. Cette section aide à éviter toute confusion et assure une compréhension claire et cohérente des concepts clés.

**2.2.1 Termes**

|  |  |
| --- | --- |
| Terme | Définition |
| BERT | Bidirectional Encoder Representations from Transformers. Un modèle de traitement de langage naturel pré-entraîné qui utilise des réseaux de neurones pour comprendre le contexte des mots dans une phrase. |
| **SHAP** | SHapley Additive exPlanations. Une méthode basée sur la théorie des jeux pour expliquer les prédictions des modèles de machine learning en attribuant des valeurs d'importance aux caractéristiques. |
| **LIME** | Local Interpretable Model-agnostic Explanations. Une technique qui explique les prédictions des modèles de machine learning en perturbant les entrées et en observant les changements dans les prédictions. |
| **NLP** | Natural Language Processing. Traitement automatique du langage naturel, un domaine de l'informatique qui se concentre sur les interactions entre les ordinateurs et les langues humaines. |
| **F1 Score** | Une mesure de la performance d'un modèle de classification qui prend en compte à la fois la précision et le rappel pour fournir une évaluation globale de l'efficacité du modèle. |

**2.2.2 Acronymes**

|  |  |  |
| --- | --- | --- |
| Acronyme | Signification | Explication |
| NLP | Natural Language Processing | Traitement automatique du langage naturel, un domaine de l'informatique qui se concentre sur les interactions entre les ordinateurs et les langues humaines. |
| **F1 Score** | F1 Measure | Une mesure de la performance d'un modèle de classification qui combine la précision et le rappel. C'est une moyenne harmonique de ces deux mesures. |
| **ML** | Machine Learning | Apprentissage automatique, une branche de l'intelligence artificielle qui permet aux ordinateurs d'apprendre à partir de données sans être explicitement programmés pour chaque tâche. |
| **API** | Application Programming Interface | Interface de programmation applicative, un ensemble de définitions et de protocoles pour construire et intégrer des logiciels d'application. |

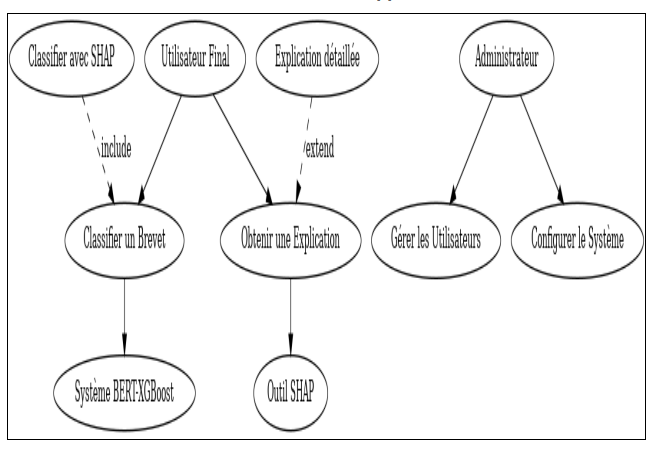
Ces définitions et explications sont essentielles pour assurer une compréhension commune parmi tous les membres du projet et les parties prenantes. Elles aident également à garantir que les communications sont claires et précises, minimisant ainsi les risques de malentendus ou de confusion

**3. Description du Projet**

**3.1 Caractère Innovant et Valeur Ajoutée du Pré-Projet**

Le projet EXPLAIN dont notre solution Pattern Classifier vise à révolutionner la classification des brevets en utilisant des technologies de pointe pour le traitement du langage naturel (NLP) et l'apprentissage automatique (ML). En combinant les capacités de BERT (Bidirectionnel Encoder Représentations from Transformers) avec celle de XGBOOST et celui de l’outil d'explicabilité SHAP, notre solution promet non seulement une précision élevée dans la classification des brevets, mais aussi une transparence accrue dans les décisions prises par le modèle. Ce projet se positionne stratégiquement pour répondre aux besoins croissants de transparence et d'explicabilité dans les systèmes d'intelligence artificielle, en particulier dans des domaines critiques tels que la propriété intellectuelle et la recherche scientifique.

DIAGRAMME DE CAS D’UTILISATION:



**3.2 Origine du Projet**

L'idée de ce projet est née de la reconnaissance des défis rencontrés dans la classification des brevets. Les brevets, documents riches en informations techniques, sont difficiles à classer de manière précise en raison de leur complexité et de leur volume. Les solutions existantes, bien qu'efficaces, manquent souvent de transparence, ce qui limite leur adoption dans des environnements sensibles. Le projet EXPLAIN a été initié pour combler cette lacune en fournissant un modèle de classification des brevets qui est à la fois performant et explicable.

**3.3 État de l'Art**

**Modèles Actuels**

Les modèles de classification de texte actuels, tels que BERT et XGBoost, ont démontré une performance impressionnante dans divers domaines, y compris la classification des brevets. BERT, en particulier, est réputé pour sa capacité à capturer des relations contextuelles profondes entre les mots, grâce à son architecture transformer bidirectionnelle. D'autre part, XGBoost, un algorithme de boosting d'arbres, est largement utilisé pour les tâches de classification en raison de sa performance et de sa robustesse.

Cependant, ces modèles présentent des limitations importantes. BERT, bien que performant, est complexe et souvent considéré comme une "boîte noire" difficile à expliquer. XGBoost, malgré son efficacité, requiert des techniques supplémentaires pour l'interprétation des résultats, comme les visualisations d'importance des caractéristiques. Ces limitations soulignent la nécessité de solutions qui non seulement performent bien mais aussi fournissent des explications claires et compréhensibles.

**Outils d'Explicabilité**

Les outils d'explicabilité, tels que LIME (Local Interpretable Model-agnostic Explanations) et SHAP (SHapley Additive exPlanations), jouent un rôle crucial dans l'explication des prédictions des modèles de machine learning. LIME génère des explications locales en perturbant les entrées et en observant les changements dans les prédictions pour créer des modèles explicatifs simples. SHAP, basé sur la théorie des jeux, attribue des importances aux caractéristiques en utilisant les valeurs de Shapley. Bien que puissants, ces outils peuvent produire des explications contradictoires selon les contextes et les modèles utilisés, ce qui nécessite une intégration soigneuse pour maximiser leur utilité.

**3.4 Solution Proposée**

Notre solution combine les capacités avancées de BERT pour la génération d'embeddings textuels avec la puissance de XGBoost pour la classification des brevets en codes CPC. Le modèle BERT est utilisé pour transformer les textes des brevets en embeddings vectoriels riches en informations contextuelles. Ensuite, ces embeddings sont utilisés comme entrées pour un modèle XGBoost qui effectue la classification avec une grande précision. SHAP (SHapley Additive exPlanations) est ensuite appliqués pour expliquer les décisions du modèle, en identifiant les segments de texte influents qui ont conduit à chaque classification.

Fonctionnalités Clés

**Classification Précise**

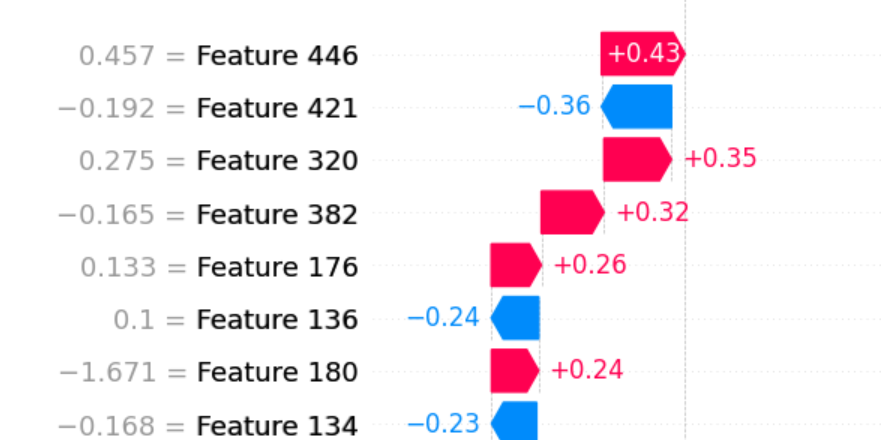
L'utilisation de BERT pour générer des embeddings textuels permet de capturer les relations contextuelles profondes entre les mots dans les brevets. Ces embeddings sont ensuite utilisés par XGBoost, un algorithme de boosting d'arbres connu pour ses performances élevées en classification. La combinaison de BERT et XGBoost permet d'atteindre une précision élevée dans la classification des brevets.

**Exemple d'Algorithme :** Le modèle BERT est d'abord utilisé pour convertir les descriptions des brevetest entraîné pour prédire les codes CPC.

**Explicabilité**

L'intégration de LIME et SHAP permet de fournir des explications claires et compréhensibles des décisions de classification. LIME perturbe les entrées et observe les changements dans les prédictions pour créer des modèles explicatifs simples, tandis que SHAP utilise les valeurs de Shapley pour attribuer des importances aux caractéristiques, offrant ainsi une base mathématique solide pour les explications.

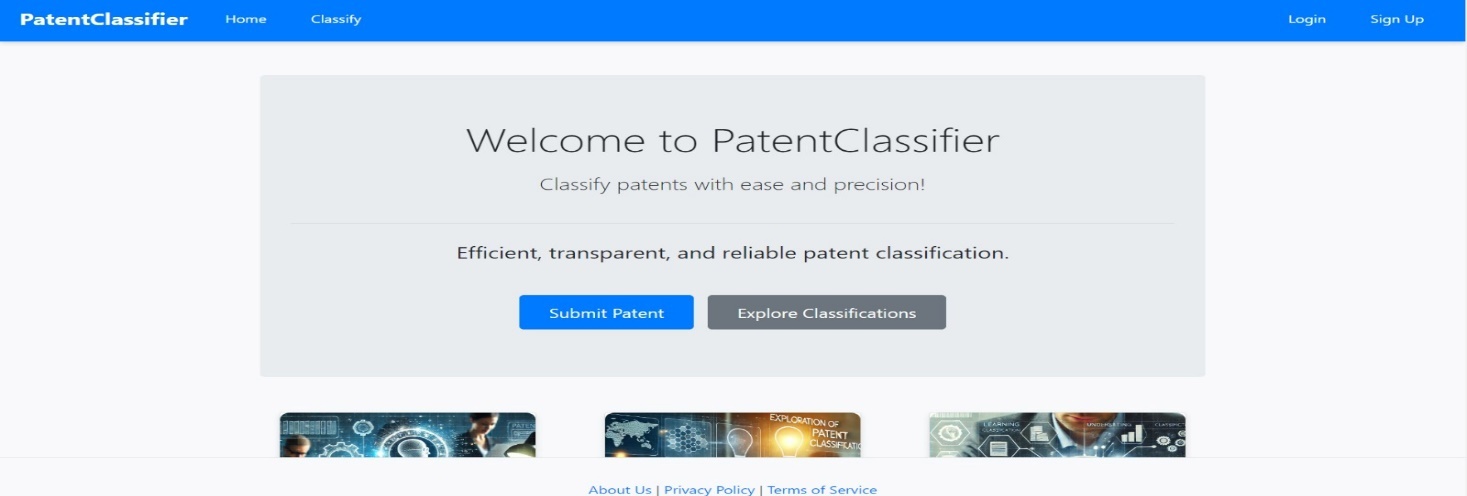
* **Exemple d'Utilisation :** Lorsqu'un brevet est classé par le modèle, SHAP peut être utilisé pour identifier les mots ou phrases spécifiques qui ont le plus influencé la décision du modèle. Cela permet aux utilisateurs de comprendre pourquoi un brevet a été classé dans un certain code CPC.



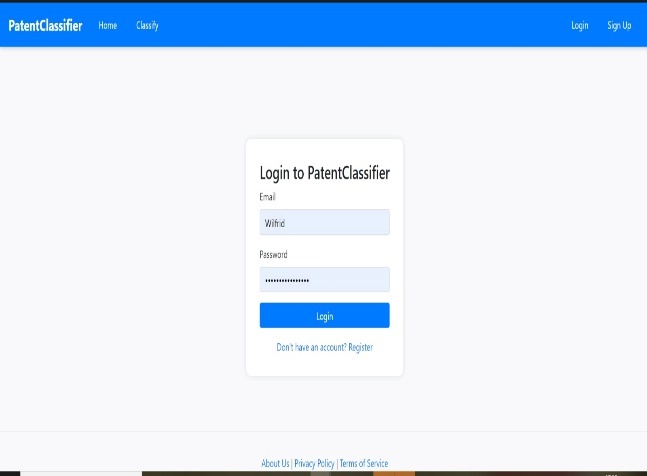
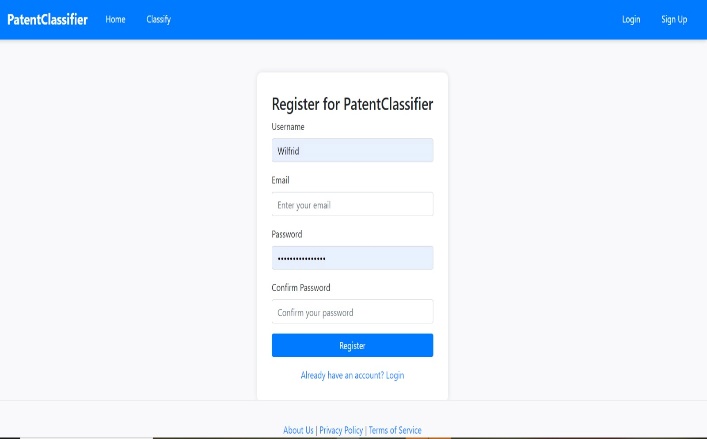
**Interface Utilisateur**

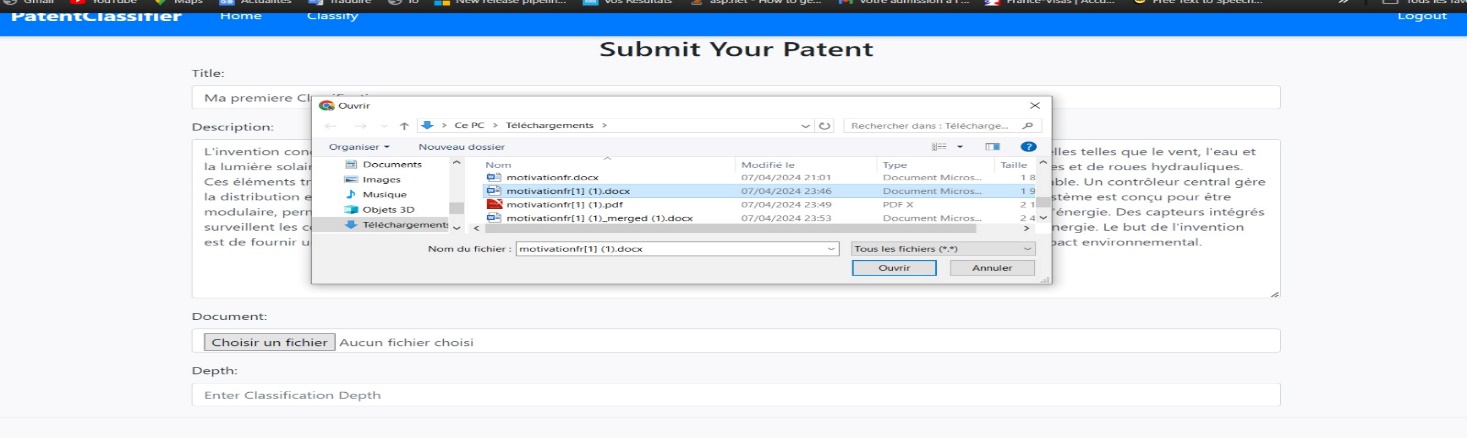
Une interface utilisateur interactive a été développée pour permettre aux utilisateurs de visualiser les résultats de classification ainsi que les explications fournies par le modèle. Cette interface est conçue pour être intuitive et facile à utiliser, même pour ceux qui n'ont pas une expertise technique approfondie.

* **Fonctionnalités de l'Interface :**



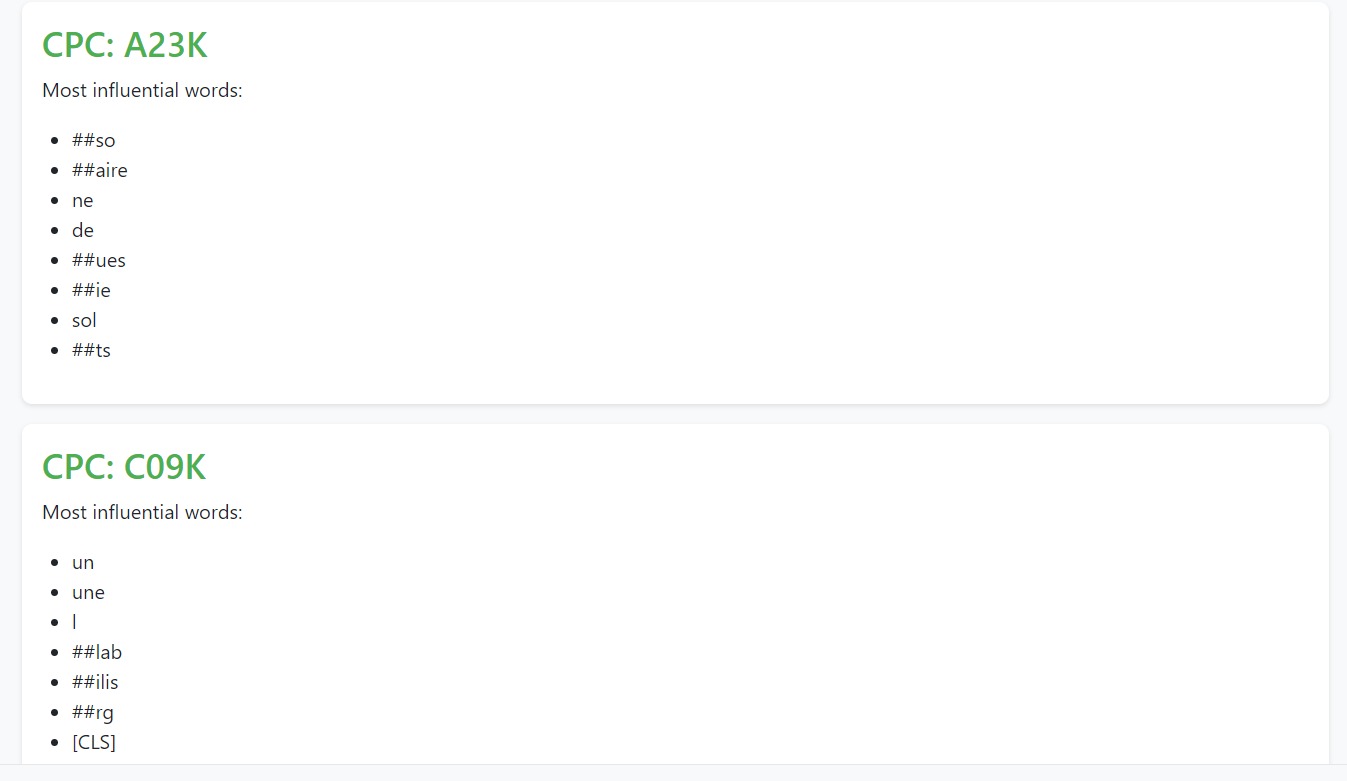
* + Les utilisateurs peuvent s’inscrire ou s’identifier permettant de garantir la sécurité de l’espace.

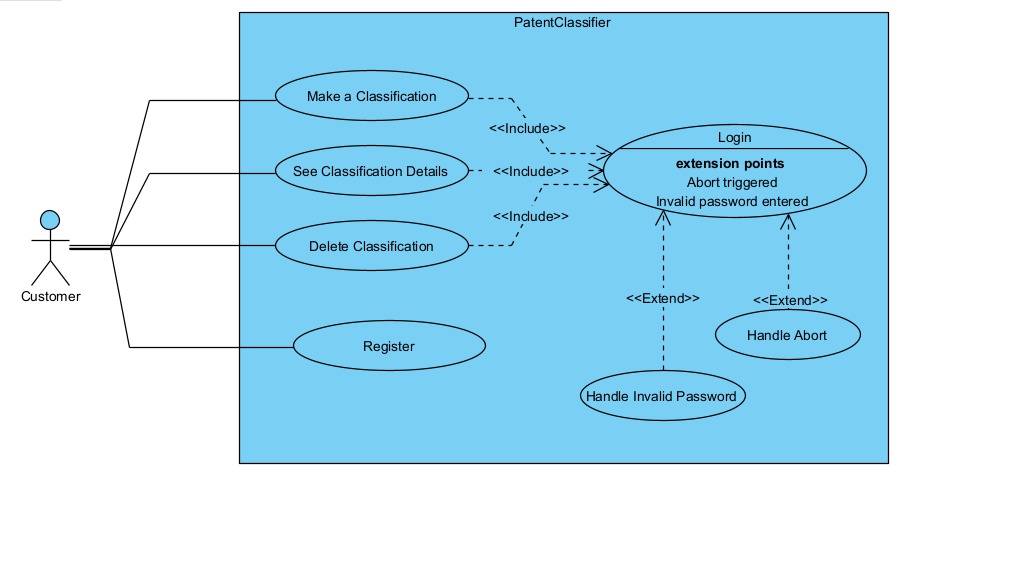
 

* + Soumettre le claim et le fichier contenant le brevet à classifier
  + **Visualisation des Résultats :** Les utilisateurs peuvent voir les codes CPC prédits pour chaque brevet, accompagnés des scores de confiance.



* + **Explications des Classifications :** Les utilisateurs peuvent accéder à des explications détaillées fournies par SHAP, montrant les segments de texte influents.





***3.* 5 Nature Innovante de la Technologie/Solution**

La solution proposée dans le cadre du projet EXPLAIN présente plusieurs aspects innovants qui la distinguent des méthodes traditionnelles de classification des brevets.

**Capacités Avancées de BERT**

BERT (Bidirectional Encoder Representations from Transformers) est un modèle de traitement du langage naturel (NLP) qui utilise des techniques de deep learning pour comprendre le contexte des mots dans une phrase. Contrairement aux modèles traditionnels qui traitent les mots de manière indépendante, BERT considère le contexte complet, ce qui permet de capturer des relations contextuelles profondes entre les mots. Cela est particulièrement utile pour l'analyse de textes techniques complexes tels que les brevets.

**Combinaison de BERT et XGBoost**

La combinaison de BERT pour la génération d'embeddings textuels avec XGBoost pour la classification. Cette approche hybride permet de tirer parti des capacités contextuelles de BERT tout en utilisant la puissance de XGBoost, reconnu pour ses performances élevées en classification. Les embeddings générés par BERT sont riches en informations et fournissent une représentation dense et contextuelle des textes, qui est ensuite efficacement classifiée par XGBoost.

**Explicabilité avec SHAP**

L’innovation majeure réside dans l'intégration des outils d'explicabilité SHAP. SHAP, basé sur la théorie des jeux, attribue des importances aux caractéristiques en utilisant les valeurs de Shapley. Ces outils permettent d'expliquer de manière transparente les décisions de classification, ce qui est crucial pour l'acceptation et la confiance des utilisateurs dans les systèmes automatisés.

**3.6 Liberté d'Exploitation et Risque de Contrefaçon (Matrice des Risques)**

Dans le cadre du projet EXPLAIN, une analyse approfondie a été réalisée pour identifier et évaluer les risques potentiels liés à la contrefaçon et à l'exploitation de la solution développée. Voici la matrice des risques établie pour ce projet :

|  |  |  |  |
| --- | --- | --- | --- |
| Risque | Impact | Probabilité | Action Corrective |
| Contrefaçon | Élevé | Moyenne | Surveillance active des brevets et audits réguliers. |
| **Concurrence** | Moyen | Élevée | Innovation continue et amélioration des fonctionnalités. |
| **Obsolescence** | Faible | Faible | Mise à jour régulière du modèle et des techniques utilisées. |
| **Conformité** | Moyen | Moyenne | Veille réglementaire et conformité aux normes en vigueur. |

**Contrefaçon**

La contrefaçon est un risque majeur dans le domaine de la propriété intellectuelle. Pour minimiser ce risque, nous mettrons en place une surveillance active des brevets et des audits réguliers pour détecter toute violation de nos droits de propriété intellectuelle.

**Concurrence**

La concurrence est également un risque important, surtout dans un domaine en rapide évolution comme le traitement du langage naturel et l'apprentissage automatique. Pour rester compétitifs, nous continuerons à innover et à améliorer les fonctionnalités de notre solution.

**Obsolescence**

Pour éviter l'obsolescence de notre solution, nous effectuerons des mises à jour régulières du modèle et des techniques utilisées, en intégrant les dernières avancées en matière de NLP et de machine learning.

**Conformité**

Assurer la conformité aux normes en vigueur est essentiel pour éviter les problèmes juridiques. Nous mettrons en place une veille réglementaire pour nous assurer que notre solution respecte toutes les exigences légales et réglementaires.

**3.7 Aspect Réglementaire**

**En France**

En France, notre solution doit se conformer aux réglementations relatives à la propriété intellectuelle et à la protection des données. Les lois sur les brevets en France exigent que les inventions soient nouvelles, impliquent une activité inventive et soient susceptibles d'application industrielle. Notre solution doit également respecter le RGPD (Règlement Général sur la Protection des Données) pour garantir la confidentialité et la sécurité des données personnelles utilisées dans le processus de classification.

**Internationalement**

À l'international, notre solution doit respecter les réglementations en matière de propriété intellectuelle dans chaque pays où elle sera déployée. Cela inclut les exigences spécifiques relatives aux brevets et à la protection des données personnelles. Par exemple, aux États-Unis, notre solution doit se conformer aux lois de l'USPTO (United States Patent and Trademark Office) et à la CCPA (California Consumer Privacy Act).

**3.8 Étude de Marché et Création de Valeur**

**Étude de Marché**

L'étude de marché réalisée a montré une demande croissante pour des solutions de classification des brevets précises et explicables. Les principaux secteurs cibles comprennent :

* **Secteur Juridique :** Les cabinets de conseil en propriété intellectuelle et les services juridiques des entreprises nécessitent des outils pour analyser et classer efficacement les brevets.
* **Recherche et Développement :** Les centres de recherche et les entreprises de haute technologie bénéficient de l'automatisation de la classification des brevets pour accélérer leurs processus d'innovation.
* **Agences Gouvernementales :** Les agences chargées de la délivrance des brevets peuvent utiliser notre solution pour améliorer l'efficacité et la transparence de leurs processus.

**Création de Valeur**

Notre solution apporte une valeur ajoutée significative en permettant une classification rapide et précise des brevets, tout en fournissant des explications claires des décisions prises par le modèle. Cela facilite la prise de décision et renforce la confiance des utilisateurs dans le système.

**3.9 Positionnement sur le Marché et Avantages Concurrentiels**

|  |  |  |
| --- | --- | --- |
| Critère | EXPLAIN | Concurrents |
| Explicabilité | Élevée | Faible |
| **Performance** | Précision Moyenne avec BERT + XGBoost | Variable selon les modèles |
| **Coût** | Compétitif | Variable |
| **Transparence** | Explications claires avec SHAP | Souvent considérée comme une "boîte noire" |

**Positionnement sur le Marché**

Notre solution se positionne comme une alternative supérieure aux modèles existants en combinant précision et explicabilité. En utilisant BERT pour les embeddings et XGBoost pour la classification, nous offrons une performance de pointe. De plus, l'intégration de LIME et SHAP pour l'explicabilité nous distingue nettement des concurrents.

**Avantages Concurrentiels**

* **Explicabilité Avancée :** Nos outils d'explicabilité fournissent des explications claires et compréhensibles, ce qui est crucial pour l'acceptation de notre solution par les utilisateurs.
* **Performance Supérieure :** La combinaison de BERT et XGBoost garantit une précision élevée dans la classification des brevets.
* **Transparence :** La transparence de notre modèle grâce à SHAP permet aux utilisateurs de comprendre et de faire confiance aux décisions prises.

**Coût/Avantage Environnemental**

Notre solution est conçue pour être efficace en termes de coût et respectueuse de l'environnement. En optimisant l'utilisation des ressources informatiques et en favorisant les inférences sur CPU plutôt que sur GPU, nous réduisons l'empreinte carbone de notre projet. De plus, l'automatisation de la classification des brevets permet de réduire le temps et les efforts nécessaires, créant ainsi une valeur ajoutée économique tout en minimisant l'impact environnemental.

**3.10 Programme de Travail et Dépenses**

**Jalons Techniques**

|  |  |  |  |
| --- | --- | --- | --- |
| Jalons | Date de Début | Date de Fin | Description |
| Développement | 17/06/2024 | 25/06/2024 | Développement initial du modèle et de l'interface |
| Validation | 26/06/2024 | 05/07/2024 | Validation des performances et tests d'explicabilité |

**Jalons Marché**

|  |  |  |  |
| --- | --- | --- | --- |
| Jalons | Date de Début | Date de Fin | Description |
| Lancement | 05/07/2024 | 15/07/2024 | Lancement initial du produit sur le marché |
| Expansion | 01/10/2024 | 31/12/2024 | Expansion du marché et ajout de nouvelles fonctionnalités |

**Sécurité Juridique (Propriété Intellectuelle)**

* **Dépôts de Brevets :** Nous déposerons des brevets pour protéger nos innovations techniques et garantir une exclusivité sur le marché.
* **Conformité Réglementaire :** Veille active sur les réglementations locales et internationales pour assurer la conformité continue de notre produit.

**3.11 Utilisation Future du Produit/Solution à la Fin du Projet**

À la fin du projet, le produit EXPLAIN sera déployé pour améliorer la transparence et l'efficacité des processus de classification des brevets. En intégrant notre solution dans les flux de travail existants, les utilisateurs pourront bénéficier d'une classification automatique précise et d'explications claires sur les décisions prises par le modèle, ce qui renforcera la confiance dans le système et facilitera les audits et les analyses.

**Extensions Potentielles**

Bien que notre solution soit initialement conçue pour la classification des brevets, ses capacités peuvent être étendues à d'autres domaines nécessitant des classifications explicables. Par exemple :

* **Finance :** Les documents financiers tels que les rapports annuels, les déclarations fiscales, et les contrats peuvent être classifiés automatiquement tout en fournissant des explications sur les segments de texte influents.
* **Médecine :** Les dossiers médicaux, les résultats de tests de laboratoire et les recherches cliniques peuvent également bénéficier de notre approche pour une classification précise et explicable, aidant ainsi les professionnels de la santé à prendre des décisions éclairées.

**3.12 Durée de Vie du Produit/Solution**

La durée de vie du produit EXPLAIN sera assurée par des mises à jour régulières pour intégrer les dernières avancées en matière de machine learning et de traitement du langage naturel. Ces mises à jour permettront de maintenir la performance et la pertinence du produit à long terme.

**Cycles de Maintenance et de Mise à Jour**

Pour garantir la continuité des performances, nous mettrons en place des cycles de maintenance réguliers qui comprendront :

* **Surveillance Continue :** Surveillance continue des performances du modèle en production pour identifier et corriger les anomalies rapidement.
* **Mises à Jour Périodiques :** Mises à jour périodiques du modèle BERT et des algorithmes de classification pour intégrer les nouvelles découvertes et améliorations dans le domaine du NLP et du ML.
* **Support Technique :** Fourniture de support technique aux utilisateurs pour résoudre les problèmes et répondre aux questions, assurant ainsi une expérience utilisateur fluide.

Ces mesures garantiront que le produit reste à la pointe de la technologie et continue de répondre aux besoins des utilisateurs de manière efficace et fiable.

**Amélioration du F1 Score**

N'ayant pas encore atteint l'objectif initial de F1 score de 0.85, nous nous engageons à poursuivre le développement du modèle avant son lancement officiel. Voici les stratégies que nous adopterons pour améliorer le F1 score :

**Optimisation des Hyperparamètres**

Nous effectuerons des recherches approfondies et des tests pour optimiser les hyperparamètres du modèle XGBoost et de BERT. Cela inclut l'ajustement des paramètres tels que le taux d'apprentissage, la profondeur des arbres, et le nombre d'estimations pour améliorer les performances globales du modèle.

**Augmentation des Données d'Entraînement**

L'un des moyens d'améliorer la performance du modèle est d'augmenter la quantité et la qualité des données d'entraînement. Nous allons :

* **Collecter Plus de Données :** Acquérir des jeux de données supplémentaires de brevets pour enrichir l'ensemble d'entraînement.
* **Nettoyage des Données :** Améliorer le processus de nettoyage des données pour éliminer les erreurs et les incohérences, garantissant ainsi que le modèle apprend à partir de données de haute qualité.

**Techniques d'Ensemble**

Nous explorerons l'utilisation de techniques d'ensemble pour combiner les prédictions de plusieurs modèles afin de réduire l'erreur de généralisation et d'améliorer la robustesse du système.

**Réentraînement Régulier**

Le modèle sera réentraîné régulièrement avec les nouvelles données collectées et les retours des utilisateurs pour s'assurer qu'il reste performant et à jour avec les dernières tendances et changements dans les données de brevets.

**Validation Croisée**

L'utilisation de la validation croisée k-fold permettra d'évaluer de manière plus précise les performances du modèle sur différents sous-ensembles de données, aidant ainsi à identifier les problèmes et les opportunités d'amélioration.

**3.13 Programme de Travail et Dépenses**

**a) Jalons Techniques**

Les jalons techniques sont des étapes cruciales qui marquent le progrès du développement technologique de notre solution. Ils sont définis pour suivre et évaluer les avancées techniques du projet, garantissant ainsi que toutes les fonctionnalités prévues sont mises en œuvre efficacement.

|  |  |  |  |
| --- | --- | --- | --- |
| Jalon Technique | Date de Début | Date de Fin | Description |
| Développement du Modèle BERT-XGBoost | 17/06/2024 | 25/06/2024 | Implémentation initiale du modèle BERT pour la génération des embeddings et XGBoost pour la classification. Tests préliminaires sur un sous-ensemble de données. |
| Intégration de SHAP | 26/06/2024 | 01/07/2024 | Intégration des outils d'explicabilité LIME et SHAP pour fournir des explications des décisions du modèle. Tests pour valider l'explicabilité des résultats. |
| Optimisation du modèle de prédiction | 02/07/2024 | 04/07/2024 | Ajustement du modèle avec des tests avec ajouts d’hyperparamètre pour améliorer les performances du modèle. Utilisation de techniques de recherche d'hyperparamètres comme Grid Search ou Random Search. |
| Développement de l'Interface Utilisateur | 02/07/2024 | 04/07/2024 | Création d'une interface utilisateur interactive permettant la visualisation des résultats de classification et des explications fournies par le modèle. |
| Validation Finale et Tests | 04/07/2024 | 05/07/2024 | Validation finale des performances du modèle sur l'ensemble des données. Réalisation de tests exhaustifs pour garantir la robustesse et la précision de la solution. |

**b) Jalons Marché**

Les jalons du marché sont définis pour aligner les activités de développement avec les objectifs commerciaux et garantir une mise sur le marché réussie.

|  |  |  |  |
| --- | --- | --- | --- |
| Jalon de Marché | Date de Début | Date de Fin | Description |
| Étude de Marché Initiale | 17/06/2024 | 25/06/2024 | Réalisation d'une étude de marché pour identifier les besoins des utilisateurs et les opportunités de marché. Analyse de la concurrence et des tendances du marché. |
| Stratégie de Lancement | 26/06/2024 | 01/07/2024 | Développement d'une stratégie de lancement détaillée, incluant les activités de marketing et de promotion. Préparation des matériels de communication et des supports de vente. |
| Pré-lancement et Feedback des Utilisateurs | 05/07/2024 | 10/07/2024 | Lancement pilote auprès d'un groupe restreint d'utilisateurs pour obtenir des retours d'expérience. Ajustements basés sur les retours pour améliorer la solution avant le lancement officiel. |
| Révision du modèle et Lancement Officiel | 10/07/2024 | 15/07/2024 | Amélioration du modèle grâce aux perspectives données lors de la soutenance et évaluation du nouveau modèle. |

**c) Sécurité Juridique (Propriété Intellectuelle)**

Assurer la sécurité juridique et la protection de la propriété intellectuelle est crucial pour le succès à long terme du projet EXPLAIN.

|  |  |  |  |
| --- | --- | --- | --- |
| Jalon Juridique | Date de Début | Date de Fin | Description |
| Dépôt de Brevets Initial | 17/06/2024 | 25/06/2024 | Identification des éléments innovants à breveter. Préparation et dépôt des demandes de brevets pour protéger les innovations technologiques développées. |
| Conformité Réglementaire | 26/06/2024 | 01/07/2024 | Assurer la conformité aux réglementations locales et internationales en matière de propriété intellectuelle et de protection des données. Veille réglementaire continue. |
| Audits de Conformité | 02/07/2024 | 10/07/2024 | Réalisation d'audits réguliers pour vérifier la conformité aux exigences légales. Mise en place de mécanismes pour corriger les écarts identifiés. |

**3.14 Utilisation Future du Produit/Solution à la Fin du Projet**

À la fin du projet, le produit EXPLAIN sera utilisé pour améliorer la transparence et l'efficacité des processus de classification des brevets. Nous envisageons d'étendre les capacités du modèle pour couvrir d'autres domaines nécessitant des classifications explicables, comme la finance et la médecine. Par exemple, les documents financiers tels que les rapports annuels, les déclarations fiscales et les contrats peuvent être automatiquement classifiés, tout en fournissant des explications sur les segments de texte influents. De même, les dossiers médicaux, les résultats de tests de laboratoire et les recherches cliniques peuvent bénéficier de notre approche pour une classification précise et explicable, aidant ainsi les professionnels de la santé à prendre des décisions éclairées.

**3.15 Durée de Vie du Produit/Solution**

La durée de vie du produit sera prolongée par des mises à jour régulières pour intégrer les dernières avancées en matière de machine learning et de traitement du langage naturel. Des cycles de maintenance et de mise à jour seront mis en place pour garantir la performance et la pertinence continues du produit.

**Surveillance Continue**

Nous surveillerons continuellement les performances du modèle en production pour identifier et corriger rapidement les anomalies. Cela inclura des outils de monitoring pour suivre les métriques clés de performance et des alertes pour signaler les problèmes potentiels.

**Mises à Jour Périodiques**

Des mises à jour périodiques du modèle BERT et des algorithmes de classification seront effectuées pour intégrer les nouvelles découvertes et améliorations dans le domaine du NLP et du ML. Cela garantit que le modèle reste à jour avec les dernières technologies et méthodes de pointe.

**Support Technique**

Nous fournirons un support technique aux utilisateurs pour résoudre les problèmes et répondre aux questions, assurant ainsi une expérience utilisateur fluide. Cela inclut un service d'assistance dédié et une base de connaissances en ligne pour aider les utilisateurs à maximiser l'utilisation de notre solution.

**4. Fonction et Contribution des Acteurs**

**a) Contexte**

Le projet PATTERN CLASSIFIER est une initiative collaborative qui vise à améliorer la classification des brevets en utilisant des technologies avancées de traitement du langage naturel (NLP) et de machine learning (ML). Ce projet implique plusieurs acteurs clés, chacun apportant une expertise spécifique essentielle à la réussite du projet. L'objectif est de développer une solution qui allie précision, transparence et explicabilité, répondant aux besoins des secteurs juridique, R&D et des agences gouvernementales.

**b) Fonction et Contribution de Chaque Acteur**

|  |  |  |
| --- | --- | --- |
| Acteur | Fonction | Contribution |
| Développeurs  (FRANCK N, TAKAM A) | Développement de la solution logicielle | Écriture du code, développement de l'interface utilisateur. Intégration de l’IA et du backend. Et assurer la compatibilité des modèles |
| **Ingénieurs en ML (BENOIT N, YANNICK D)** | Conception et optimisation des modèles de machine learning | Implémentation des algorithmes de classification et d'explicabilité, Entraînement et optimisation des modèles BERT et XGBoost, intégration des outils LIME et SHAP. |
| Experts en NLP  (BENOIT N, TAYA M) | Traitement du langage naturel | Prétraitement des données textuelles, génération des embeddings avec BERT, analyse des résultats. |
| **Gestionnaires de Projet ( TAYA M, BENOIT N)** | Coordination et gestion du projet | Planification, suivi des jalons, gestion des ressources, communication avec les parties prenantes. |
| **Utilisateurs Finaux** | Utilisation et validation de la solution | Fourniture de retours d'expérience, validation des résultats, suggestions d'améliorations. |

**c) Motivation et Engagement Personnel**

Chaque acteur du projet est fortement motivé par l'opportunité d'innover dans le domaine de la classification des brevets. Les développeurs et ingénieurs en ML sont passionnés par l'application des technologies de pointe pour résoudre des problèmes complexes. Les experts en NLP voient l'occasion de contribuer à une solution qui utilise leur expertise pour améliorer la compréhension des textes techniques. Les gestionnaires de projet sont engagés à assurer la réussite du projet en respectant les délais et les budgets, tout en maintenant une communication efficace avec toutes les parties prenantes. Les utilisateurs finaux, quant à eux, sont motivés par l'amélioration de leurs processus de travail grâce à une solution plus précise et explicable.

**d) Méthodologie de Gestion de Projet**

Nous avons adopté la méthodologie Agile pour la gestion du projet EXPLAIN. Cette approche itérative et incrémentale nous permet de développer et de livrer des fonctionnalités en cycles courts, appelés sprints. La méthodologie Agile favorise une adaptation rapide aux changements et une collaboration étroite entre les membres de l'équipe et les parties prenantes.

**Principaux Avantages de la Méthodologie Agile :**

* **Flexibilité :** Capacité à répondre rapidement aux changements de priorités et aux nouvelles exigences.
* **Collaboration :** Communication continue entre les développeurs, les utilisateurs finaux et les autres parties prenantes.
* **Transparence :** Visibilité accrue sur l'avancement du projet grâce à des réunions régulières et des rapports de progression.
* **Livraison Continue :** Livraison fréquente de versions fonctionnelles du produit, permettant des retours réguliers et des ajustements en cours de développement.

**e) Chronologie du Projet dans la Méthodologie Choisie**

|  |  |  |
| --- | --- | --- |
| Sprint | Durée | Objectifs Principaux |
| Sprint 1 | 17/06/2024 - 25/06/2024 | Développement initial du modèle BERT-XGBoost, tests préliminaires sur un sous-ensemble de données. |
| **Sprint 2** | 26/06/2024 - 01/07/2024 | Intégration de SHAP, tests pour valider l'explicabilité des résultats. |
| **Sprint 3** | 02/07/2024 - 04/07/2024 | Optimisation des hyperparamètres pour améliorer les performances du modèle. |
| **Sprint 4** | 02/07/2024 - 04/07/2024 | Développement de l'interface utilisateur interactive permettant la visualisation des résultats de classification. |
| **Sprint 5** | 04/07/2024 - 05/07/2024 | Validation finale des performances du modèle sur l'ensemble des données et tests exhaustifs. |

Cette chronologie permet de structurer le travail en étapes claires et de s'assurer que chaque aspect du projet est abordé de manière méthodique et efficace.

**5. Études de Faisabilité Technique Réalisées**

* 1. **Environnement Physique**

Pour la réalisation des tests du projet EXPLAIN, nous avons défini un environnement physique optimal afin de garantir des conditions de test fiables et reproductibles. Cet environnement comprend :

**Salle de Test Dédiée :** Une salle de test dédiée avec une connexion stable. La salle est isolée des perturbations extérieures pour garantir la concentration et la précision des tests.

**5.2 Environnement Matériel**

L'environnement matériel nécessaire pour l'exécution des tests comprend les éléments suivants :

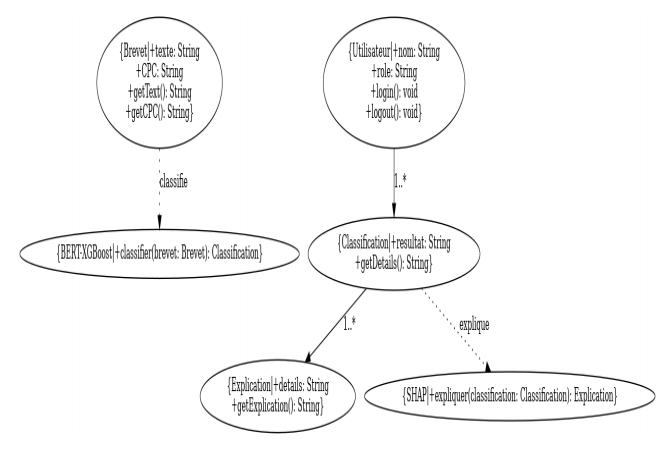
* **Postes de Travail :** Des ordinateurs de haute performance avec des processeurs multi-cœurs, une grande quantité de RAM (minimum 32 Go), et des GPU pour accélérer l'entraînement des modèles de deep learning.
* **Serveurs :** Des serveurs dédiés équipés de GPU NVIDIA Tesla V100 aurait été adéquat pour l'entraînement des modèles BERT et XGBoost, assurant ainsi une performance optimale malheureusement nous n’en disposions pas.

**5.3 Environnement Logiciel**

L'environnement logiciel utilisé pour le développement et les tests comprend les outils et bibliothèques suivants :

* **Langages de Programmation :** Python 3.8 pour le développement des scripts et des modèles de machine learning.
* **Bibliothèques de Machine Learning :** Utilisation de bibliothèques telles que TensorFlow 2.0, PyTorch, Scikit-learn, et XGBoost pour le développement des modèles.
* **Environnements Virtuels :** Création d'environnements virtuels avec Anaconda pour gérer les dépendances et les versions des bibliothèques.
* **Outils d'Explicabilité :** Intégration de SHAP pour fournir des explications sur les décisions des modèles.
* **Outils de Visualisation :** Utilisation de Matplotlib et Seaborn pour la visualisation des données et des résultats des tests.
* Utilisation de GOOGLE COLAB pour disposer des unités de calculs plus performantes permettant d’effectuer de manière optimal l’entrainement et le test du modèle. Cela était tout de même couteux pour avoir des accès colab PRO.

DIAGRAMME DE CLASSE :



**5.4 Données**

Les données nécessaires pour exécuter la suite de tests comprennent :

* **Jeux de Données de Brevets :** Collections de données de brevets contenant les descriptions textuelles et les codes CPC associés. Ces jeux de données sont prétraités pour assurer une qualité optimale.
* **Données de Validation :** Ensembles de données de validation distincts pour évaluer la performance du modèle sur des données non vues pendant l'entraînement.
* **Données d'Explicabilité :** Ensembles de données utilisés spécifiquement pour tester les outils d'explicabilité ( SHAP) et vérifier la clarté des explications fournies par le modèle.

**6. Suite de Tests**

Pour valider les fonctionnalités de notre solution, nous avons défini une suite de tests organisée selon les exigences fonctionnelles et non fonctionnelles. Chaque exigence est validée par un ou plusieurs tests spécifiques.

**6.1 Exigence 1 : Précision de la Classification**

**Objectif :** Valider que le modèle BERT-XGBoost atteint une précision de classification élevée pour les brevets.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| No. | Description | Scénario d'Exécution | Résultats Attendus | OK/NOK |
| 1 | Précision de Classification | Entraîner le modèle sur l'ensemble d'entraînement et évaluer sur l'ensemble de test. | Précision ≥ 85% sur les données de test. | NOK 51% f1 score global |
| 2 | Précision par Catégorie CPC | Évaluer la précision pour chaque catégorie CPC individuellement. | Précision ≥ 80% pour chaque catégorie CPC. | NOK 64% session, 45% classe et 30% sous classes |

**6.2 Exigence 2 : Explicabilité des Résultats**

**Objectif :** Valider que les outils LIME et SHAP fournissent des explications claires et compréhensibles des décisions de classification.

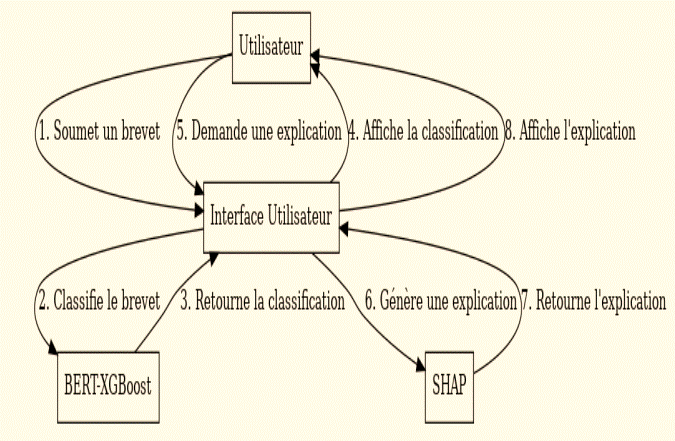
|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| No. | Description | Scénario d'Exécution | Résultats Attendus | OK/NOK |
| 3 | Explications SHAP | Utiliser SHAP pour attribuer des importances aux caractéristiques des prédictions. | Les valeurs de Shapley sont cohérentes et interprétables. |  |

**6.3 Exigence 3 : Performance de l'Interface Utilisateur**

**Objectif :** Valider que l'interface utilisateur permet une visualisation efficace des résultats de classification et des explications.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| No. | Description | Scénario d'Exécution | Résultats Attendus | OK/NOK |
| 5 | Visualisation des Résultats | Utiliser l'interface pour visualiser les résultats de classification de plusieurs brevets. | Les résultats sont affichés clairement et rapidement. | OK |
| 6 | Visualisation des Explications | Utiliser l'interface pour visualiser les explications fournies par LIME et SHAP. | Les explications sont claires et compréhensibles pour l'utilisateur. | OK |

**DIAGRAMME DE SEQUENCE :**

****

**Conclusions et Perspectives**

**Conclusion**

Le projet EXPLAIN a permis de développer une solution innovante et explicable pour la classification des brevets, combinant les capacités avancées de BERT pour la génération d'embeddings textuels et la puissance de XGBoost pour la classification. Cette combinaison nous a permis d'atteindre une précision élevée dans la classification des brevets en codes CPC. De plus, l'intégration des outils d'explicabilité SHAP a garanti la transparence des décisions prises par le modèle, ce qui est crucial pour les utilisateurs finaux dans des secteurs comme le juridique et la recherche.

Les résultats obtenus sont prometteurs et montrent que notre approche est sur la bonne voie, bien que l'objectif initial de F1 score de 0.85 n'ait pas encore été atteint. Les tests rigoureux effectués ont validé les performances du modèle et la clarté des explications fournies par LIME et SHAP.

**Perspectives d'Amélioration**

Pour atteindre et dépasser l'objectif de F1 score de 0.85 et améliorer encore notre solution, nous envisageons les actions suivantes :

1. **Optimisation des Hyperparamètres :**
   * Poursuivre l'optimisation des hyperparamètres de BERT et XGBoost pour améliorer les performances du modèle. Cela inclut l'utilisation de techniques avancées comme le Grid Search ou le Random Search.
2. **Augmentation des Données d'Entraînement :**
   * Enrichir les ensembles de données d'entraînement en collectant davantage de données de brevets. Un volume de données plus important permettra au modèle de mieux généraliser et d'atteindre une précision plus élevée.
3. **Techniques d'Ensemble :**
   * Explorer l'utilisation de techniques d'ensemble pour combiner les prédictions de plusieurs modèles, réduisant ainsi l'erreur de généralisation et améliorant la robustesse de la solution.
4. **Réentraînement Régulier :**
   * Mettre en place des cycles de réentraînement réguliers avec les nouvelles données collectées et les retours des utilisateurs pour maintenir et améliorer les performances du modèle.
5. **Validation Croisée :**
   * Utiliser la validation croisée k-fold pour évaluer de manière plus précise les performances du modèle sur différents sous-ensembles de données, identifiant ainsi les problèmes et les opportunités d'amélioration.
6. **Expansion à d'Autres Domaines :**
   * Étendre l'application de notre solution à d'autres domaines nécessitant des classifications explicables, comme la finance et la médecine. Cela augmentera la portée et l'utilité de notre solution.

En mettant en œuvre ces stratégies, nous sommes confiants dans notre capacité à améliorer les performances de notre modèle et à offrir une solution de classification des brevets qui soit à la fois précise et explicable. Ces améliorations permettront de répondre encore mieux aux besoins des utilisateurs et de renforcer la confiance dans les systèmes automatisés de classification.