Rapport

Exploration et Analyse Avancée en Machine Learning :

2023

Réalisé par
Samiha
El Mansouri

Encadré par
M. Aziz
Khamjane

Table des matières

REMERCIEMENTS

Nous tenons à exprimer notre profonde reconnaissance envers l'École Nationale de Sciences Appliquées d'Al Hoceima.

Nos remerciements les plus sincères vont à Monsieur Aziz Khamjane, notre estimé professeur de Machine Learning, pour sa guidance éclairée et son soutien inestimable tout au long de cette expérience. Ses conseils avisés ont illuminé notre parcours académique, contribuant significativement à notre développement professionnel.

Nous saisissons également cette opportunité pour exprimer notre gratitude envers nos familles et nos amis, qui ont été un pilier de soutien inébranlable tout au long de la conception de notre projet professionnel et dans l'élaboration minutieuse de ce rapport.

Un remerciement particulier est adressé à notre relecteur et correcteur, dont les remarques précieuses et les suggestions pertinentes ont grandement amélioré la qualité finale de ce rapport.

Enfin, nos sincères remerciements vont à toute l'équipe de l'École Nationale de Sciences Appliquées, ainsi qu'à chaque personne ayant contribué de près ou de loin à rendre cette expérience exceptionnelle possible. Votre dévouement et votre appui ont été d'une importance cruciale.

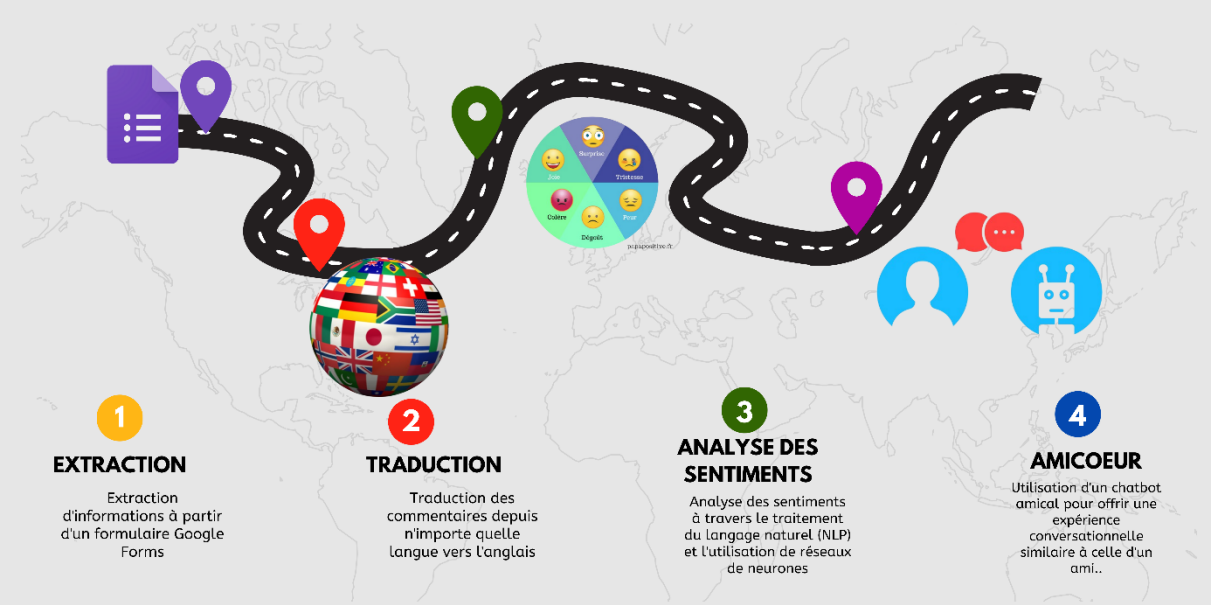
Encore une fois, nous vous exprimons notre gratitude la plus chaleureuse pour avoir fait de cette expérience une opportunité enrichissante et formatrice.

Résumé

Le projet vise à mettre en œuvre une analyse exhaustive des retours et des évaluations des utilisateurs concernant les cours. Cette analyse repose sur l'utilisation de réseaux neuronaux pour interpréter les avis des utilisateurs envers ces cours. Une composante clé est l'intégration d'une fonctionnalité de traitement du langage naturel (NLP) qui permet d'interpréter les commentaires des utilisateurs de manière plus contextuelle et précise.

De plus, pour favoriser la diversité linguistique des utilisateurs, une section de traduction a été implémentée. Cette fonctionnalité autorise les utilisateurs à s'exprimer dans leur langue maternelle parmi un choix de 133 langues, élargissant ainsi la portée et la participation à travers une expérience utilisateur plus inclusive.

Dans le but d'améliorer l'interaction globale, un chatbot a été intégré. Ce dernier offre une plateforme d'assistance et d'interaction pour les utilisateurs, facilitant des échanges plus fluides, personnalisés et efficaces.



Introduction

La genèse de ce projet découle d'une observation perspicace quant à la méthodologie actuelle de collecte des retours d'étudiants, notamment dans le contexte des cours dispensés par M. Tarik Boudaa, enseignant en Java. La pratique courante consiste en l'envoi de formulaires en ligne pour évaluer le niveau de satisfaction de chaque étudiant à l'égard du cours. Toutefois, cette opération, bien qu'indispensable, s'avère chronophage tant du point de vue du temps que de l'effort manuel requis.

La motivation à l'origine de ce projet réside dans la volonté de pallier ces contraintes logistiques en automatisant l'analyse de ces formulaires. L'objectif est d'optimiser l'efficacité du processus d'évaluation, en permettant une analyse rapide et précise des réponses fournies par les étudiants. Une constatation complémentaire a émergé lors de cette réflexion : la diversité linguistique des étudiants, chacun souhaitant s'exprimer dans sa langue maternelle. Ainsi, l'intégration d'une fonction de traduction a été envisagée, afin de garantir une participation sans entraves.

Dans la lignée de la réalisation d'une analyse de sentiment en temps réel, une étape supplémentaire a été entreprise pour résoudre une problématique délicate. Il a été observé que parfois, les étudiants peuvent se retrouver dans des situations difficiles sur le plan émotionnel, mais ne trouvent pas toujours le moyen d'exprimer leur malaise. Pour répondre à ce besoin, une section dédiée à un "ami virtuel" a été incorporée au projet. Cette fonctionnalité vise à créer un espace sûr où les étudiants peuvent partager leurs sentiments sans appréhension, tout en bénéficiant d'un soutien confidentiel pour surmonter leurs difficultés.

Ainsi, ce projet ambitieux se profile comme une réponse intelligente et globale aux défis inhérents à la collecte d'informations précieuses dans un environnement éducatif, tout en prenant en compte les aspects logistiques et émotionnels qui contribuent à l'enrichissement de l'expérience des étudiants.

Méthodologie

1. **Extraction des Données :**
    - **Utilisation de l'API Google Cloud :** La première étape a impliqué l'utilisation de l'API Google Cloud pour extraire les données des formulaires en ligne. Un jeton d'authentification a été généré pour sécuriser l'accès à l'API et garantir une connexion sécurisée.
2. **Conversion des Données au Format Excel :**
    - **Transformation des Données :** Les données brutes extraites ont été traitées et converties dans un format adapté pour une manipulation ultérieure. Les réponses des utilisateurs, comprenant le score en mathématiques, en lecture, et en écriture, le genre, ainsi que le commentaire libre sur le cours de Machine Learning, ont été extraites et structurées.
    - **Utilisation de Bibliothèques Python :** Des bibliothèques Python, telles que pydrive, ont été utilisées pour manipuler et structurer les données, les rendant ainsi compatibles avec le format Excel. pydrive a été spécifiquement employée pour faciliter l'interaction avec l'API Google Drive, permettant ainsi

l'extraction et la manipulation des données extraites depuis les formulaires en ligne, et leur conversion dans le format souhaité pour le fichier Excel.

3. **Création du Fichier Excel**

	C	D	E	F	G	H	I
1	math score	reading score	writing score	gender	ton ressenti vis-à-vis du cours de Machine Learning		
2	68	30	30	female	Plonger dans le monde du Machine Learning a été une expérience captivante et épanouissante pour moi. La complexité des algorithmes et la façon dont ils sont utilisés pour résoudre des problèmes du monde réel m'ont profondément fasciné. Chaque concept abordé dans le cours a renforcé ma compréhension et ma passion pour ce domaine. La possibilité de créer des modèles prédictifs et de voir comment ils peuvent être appliqués dans divers contextes m'a véritablement enchanté. Ce cours a ouvert une porte vers un domaine où la créativité et la résolution de problèmes se rencontrent, suscitant en moi un amour grandissant pour le Machine Learning. Chaque instant passé à explorer ses concepts est une source d'inspiration et de motivation pour approfondir mes connaissances dans cette discipline excitante.		
3							
4							
5							
6							
7							
8							
9							

Utilisation de NLP avec NLTK

La phase d'analyse de sentiment a été cruciale pour comprendre les opinions des utilisateurs exprimées dans les commentaires recueillis. Pour ce faire, j'ai opté pour la bibliothèque NLTK (Natural Language Toolkit) afin de tirer parti de ses fonctionnalités de traitement du langage naturel.

Téléchargement des Ressources NLP : Dans un premier temps, j'ai téléchargé les ressources nécessaires de NLTK, notamment le jeu de données twitter_samples contenant des tweets annotés pour l'entraînement du modèle, ainsi que le fichier de stopwords afin de nettoyer les commentaires des éléments non informatifs.

Prétraitement des Données : Pour analyser les sentiments des commentaires, j'ai effectué un nettoyage en éliminant les caractères spéciaux (laissons les symboles 😊 et le symbole ☹), normalisant le texte en minuscules et supprimant les stopwords (laissons le mot 'not'). De plus, chaque mot contenant 'not' a été concaténé avec le mot suivant. J'ai également supprimé les balises HTML et les liens, réalisé la lemmatisation pour réduire les mots à leur forme de base, supprimé les mots fréquents non informatifs, La tokenization, réalisée avec le module TweetTokenizer, a permis une meilleure séparation des mots, et j'ai étendu cette approche en incluant la tokenization des bigrams pour capturer des combinaisons de mots significatives. Cette incorporation des bigrams a été particulièrement utile pour saisir le contexte des phrases et renforcer la compréhension des relations entre les mots., et pris en compte la gestion des émoticônes. Cette approche a permis de créer une représentation textuelle plus concise

Utilisation du Modèle Naïve Bayes :

Pour entraîner notre modèle d'analyse de sentiment, nous avons choisi l'algorithme Naive Bayes, réputé pour sa rapidité d'entraînement (notamment avec un temps d'exécution maximum de 12 secondes, la plupart des exécutions se situant entre 0,1 et 2 secondes) et son efficacité avérée dans la prédiction. La première étape cruciale était d'identifier le nombre de classes, dans notre cas, les tweets positifs et négatifs. Ensuite, nous avons calculé les probabilités a priori, représentant la probabilité sous-jacente qu'un tweet soit positif ou négatif dans la population cible.

```
P(D_{pos}) = D_{pos}\D
P(D_{neg}) = D_{neg}\D
logprior=log(P(Dneg)\P(Dpos))=log(Dneg\Dpos).
```

Pour chaque mot dans notre vocabulaire, nous avons calculé les probabilités positives et négatives, en tenant compte des fréquences spécifiques de ces mots dans les classes positive et négative. Ces calculs ont été effectués avec la formule détaillée, permettant d'obtenir les probabilités de chaque mot.

La log-vraisemblance de chaque mot a été calculée en utilisant les équations spécifiées, incorporant la somme des logarithmes des probabilités des occurrences de mots dans les classes positive et négative. Ces mesures ont été essentielles pour évaluer la pertinence de chaque mot dans la classification des sentiments

```
 P(Wpos)=freqpos+1\Npos+V
 P(Wneg)=freqneg+1\Nneg+V
loglikelihood = log P(W_{pos})\P(W_{neg})
p=logprior+∑iN(loglikelihoodi)
ratio=pos_words+1\neg_words+1
```

Utilisation SVM

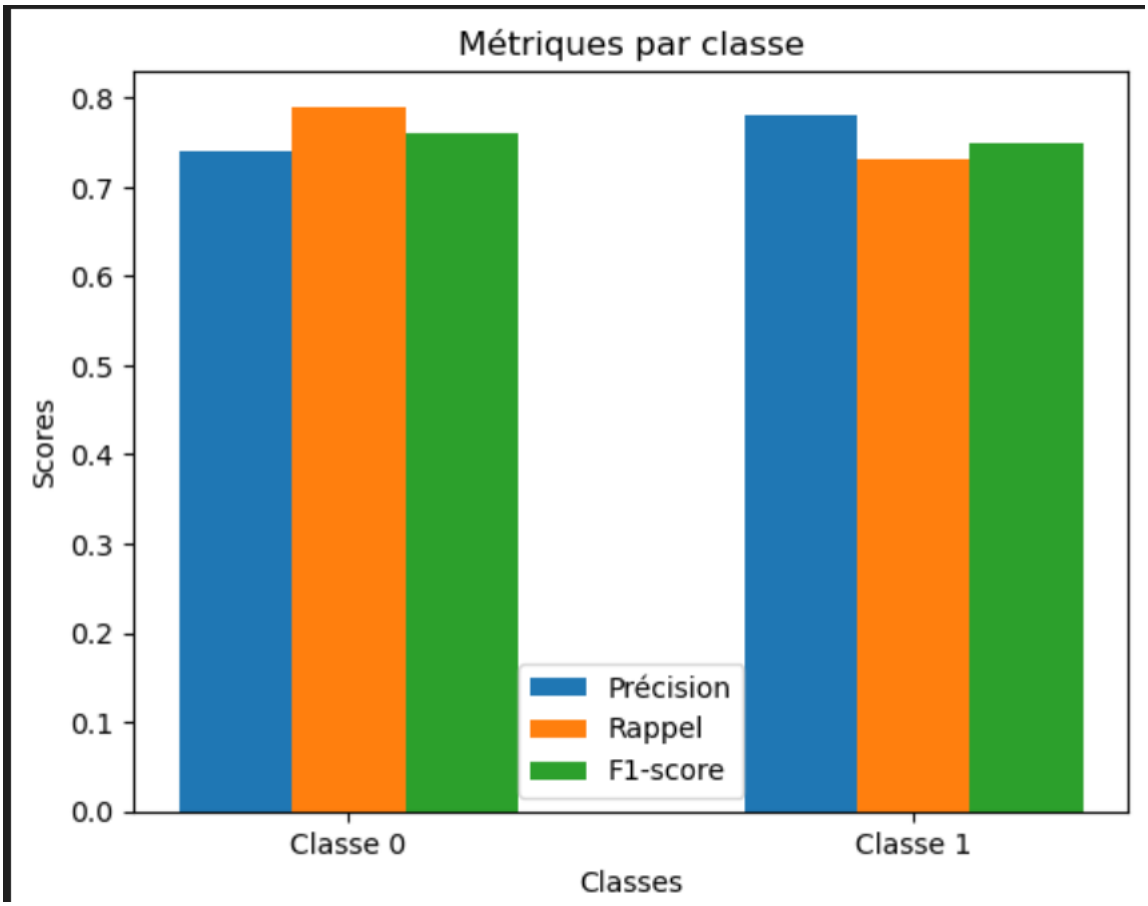
Dans l'objectif d'analyser les sentiments à partir de données textuelles, j'ai employé un modèle de Support Vector Machine (SVM). Ce modèle, implémenté à l'aide de la bibliothèque scikit-learn et la librairie NLTK pour le traitement de texte, s'est révélé particulièrement performant dans la classification des tweets en sentiments positifs et négatifs.

Pour ce faire, j'ai exploité un jeu de données de tweets positifs et négatifs préalablement étiquetés, téléchargé à partir de la collection de données Twitter de NLTK. Ces données ont été divisées en ensembles d'entraînement et de test afin d'évaluer la performance du modèle de manière impartiale.

Avant de fournir les données au modèle SVM, j'ai procédé à une vectorisation des textes à l'aide de la classe `CountVectorizer` de scikit-learn. Cette étape a converti les tweets en représentations numériques, permettant ainsi au modèle SVM de travailler avec ces données. Le noyau linéaire a été choisi pour le SVM, ce qui est souvent efficace pour les tâches de classification de texte.

Après l'entraînement du modèle sur l'ensemble d'entraînement, des prédictions ont été effectuées sur l'ensemble de test. Les performances du modèle ont ensuite été évaluées à l'aide de différentes métriques, dont la précision, le rappel et le score F1. Les résultats ont été affichés, montrant une précision de 76%, indiquant que le modèle a correctement classifié 76% des tweets du jeu de test.

L'analyse du rapport de classification fournit des informations détaillées sur la précision, le rappel et le score F1 pour chaque classe (positif et négatif), ainsi que des mesures agrégées pour évaluer la performance globale du modèle. Ces résultats démontrent l'efficacité du modèle SVM dans la tâche d'analyse de sentiment des tweets.



Utilisation des Réseaux de Neurones pour l'Entraînement :

Dans la phase d'entraînement, j'ai exploité des réseaux de neurones, en utilisant initialement un modèle à deux couches, puis en étendant l'approche à un modèle à L couches. Ces modèles ont été développés en utilisant le fichier de données `exams.csv` comme source pour l'entraînement.

Modèle à Deux Couches : Pour le modèle à deux couches, j'ai défini les dimensions des couches comme suit :

- Couche d'entrée (X) : 3 neurones
- Couche cachée (Hidden Layer) : 1000 neurones, avec une fonction d'activation ReLU pour favoriser l'apprentissage de caractéristiques complexes.
- Couche de sortie (Y) : 1 neurone avec une fonction d'activation sigmoïde pour la régression logistique.

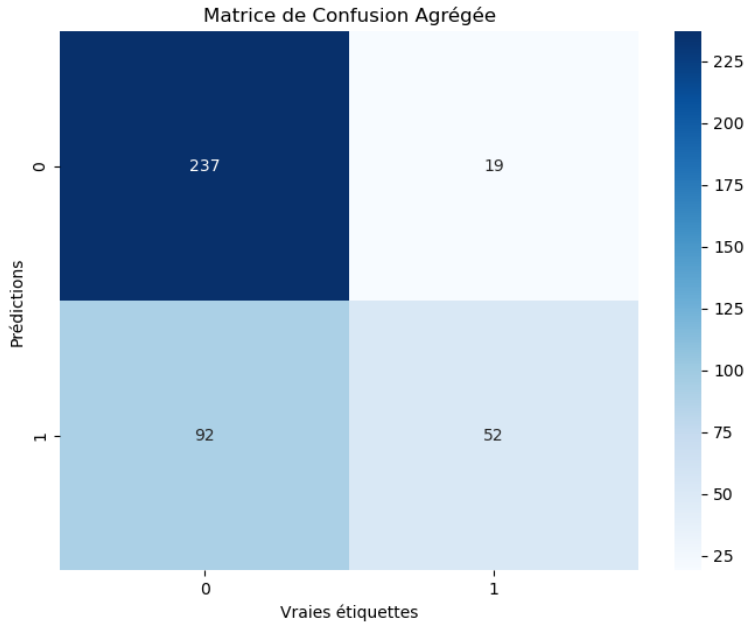
Le calcul du coût a été effectué à la fin du modèle, optimisant les poids grâce à la descente de gradient.

Modèle à L Couches : Dans le cas du modèle à L couches, j'ai spécifié les dimensions des différentes couches comme suit : `[3, 10, 10, 3, 1]`. Cela représentait quatre couches, avec trois neurones dans la première couche, dix neurones dans chaque couche cachée intermédiaire, et un neurone dans la couche de sortie.

Chaque couche cachée utilisait une fonction d'activation ReLU, tandis que la couche de sortie employait une fonction d'activation sigmoïde. Le calcul du coût a été effectué à la fin de l'architecture, et l'ensemble du modèle a été optimisé pour minimiser ce coût.

Résultats et Optimisation : avec un Test F1-score de 0.4837 et une précision de 0.67. Ces scores indiquent une capacité modérée du modèle à classifier correctement les cas, bien que des améliorations soient requises.

Cependant, une limitation notable a été identifiée, impliquant la classification erronée de 92 étudiants qui sont considérés comme ayant réussi le test alors qu'ils ne l'ont pas effectivement fait.



Intégration de la Traduction Automatique pour une Expression Multilingue Libre :

Afin de promouvoir la diversité linguistique et permettre aux utilisateurs de s'exprimer dans leur langue préférée, j'ai mis en place une fonctionnalité de traduction automatique au sein de l'application. Cette fonctionnalité repose sur l'utilisation de Google Translate et offre aux utilisateurs la possibilité de choisir parmi 133 langues différentes. Voici comment cette fonctionnalité est intégrée dans l'interface utilisateur :

Intégration d'un Chatbot OpenAI pour une Interaction Personnalisée :

Afin d'enrichir l'expérience utilisateur et d'encourager une communication plus naturelle, j'ai introduit un chatbot basé sur OpenAI au sein de l'application. Ce chatbot offre une interface conviviale permettant aux utilisateurs de dialoguer de manière conversationnelle, facilitant ainsi l'expression de leurs idées. Voici comment cette fonctionnalité est intégrée dans l'application

Amélioration de l'Expérience Utilisateur à travers une Interface Intuitive avec Streamlit et Animations :

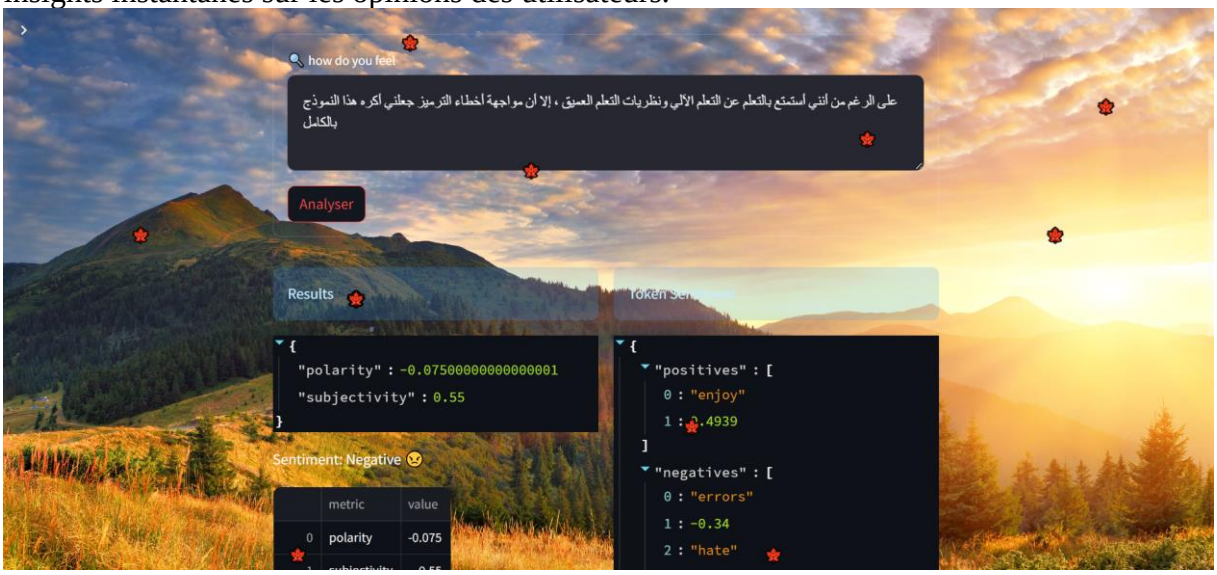
Afin d'optimiser l'expérience utilisateur, j'ai conçu une interface conviviale à l'aide de Streamlit, comprenant cinq sections distinctes pour faciliter la navigation. L'ajout d'animations a été intégré pour rendre l'expérience plus dynamique et attrayante. Voici comment ces éléments ont été intégrés :

1. Page d'Accueil :

La page d'accueil constitue le point de départ du projet, offrant une introduction concise et invitante. Elle présente une brève définition du projet, mettant en avant ses objectifs et son utilité. L'interface de cette section est pensée pour accueillir chaleureusement les visiteurs et les encourager à explorer les différentes facettes du projet.

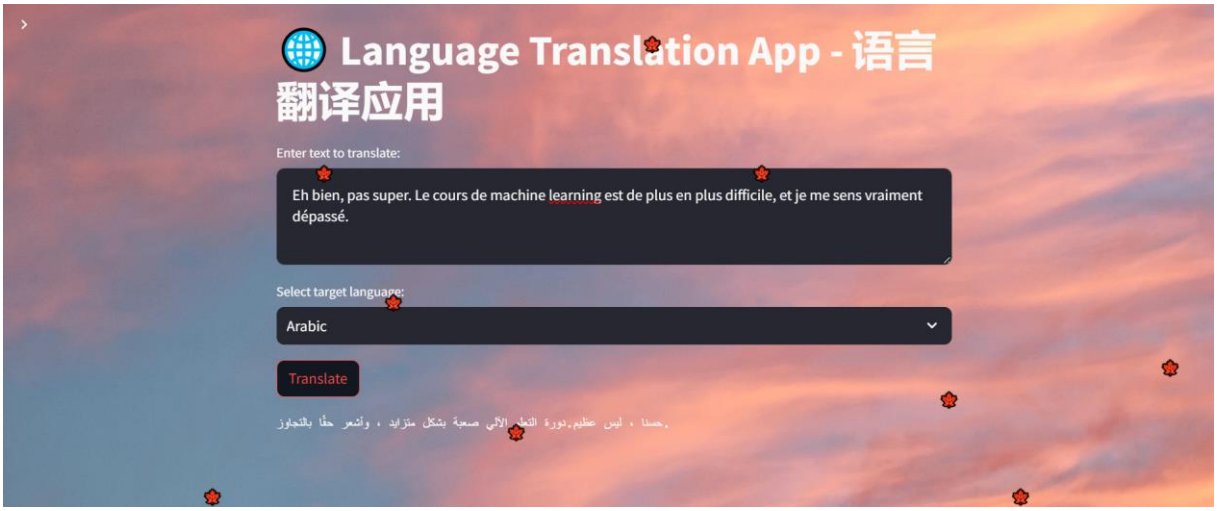
2. Interface d'Analyse des Sentiments :

L'interface d'analyse des sentiments propose une exploration approfondie des commentaires des utilisateurs et de leurs scores associés. Des graphiques interactifs permettent une visualisation claire des résultats, facilitant la compréhension des tendances générales. L'intégration de l'analyse de sentiment en temps réel enrichit cette section en fournissant des insights instantanés sur les opinions des utilisateurs.



3. Interface de Traduction Multilingue :

Cette interface met en avant la fonctionnalité de traduction multilingue, permettant aux utilisateurs de s'exprimer dans leur langue maternelle. Une zone de saisie intuitive associée à un menu déroulant de langues facilite le processus de traduction. L'intégration transparente avec Google Translate garantit une traduction instantanée, améliorant ainsi la compréhension globale des commentaires.



5. Interface de Chatbot Amical :

La dernière interface propose une expérience de conversation avec un chatbot alimenté par OpenAI. Conçu pour être amical et interactif, le chatbot offre aux utilisateurs la possibilité d'exprimer leurs pensées de manière conversationnelle. Des animations dynamiques ajoutent une touche vivante à la conversation, créant une atmosphère conviviale et engageante.

Conclusion

En somme, ce projet offre une solution complète et novatrice pour soutenir les décisions pédagogiques. La combinaison de l'analyse des notes, des commentaires, de l'analyse des sentiments et de l'interaction avec un chatbot crée un outil holistique pour améliorer l'expérience éducative. Des ajustements continus et des retours d'utilisateurs seront essentiels pour affiner davantage ce système afin qu'il réponde aux besoins changeants de la communauté éducative.

Annexes

A.CS229 Lecture Notes Andrew Ng and Tengyu Ma June 11, 2023

B. Code Source du Projet