~~ Rapport 1 Projet IA ~~

➢ BERGONZI Vinicius ➢ LENING Steve ➢ MASI Alessio ➢ THEUBO Ghislain	**ISI 3A - G5**

Projet choisi :
Projet de développement d'une application

Thème :
Analyse de sentiments pour la signalisation automatique de messages dans un système de messagerie privé d'une entreprise en vue de traquer le harcèlement en milieu professionnel.

Description de l'application :
Créer une application web du style système de messagerie privé d'une entreprise dans laquelle chaque message est analysé automatiquement par un modèle d'Intelligence Artificielle et signalé si le taux de "haine"/"négativité" est supérieur à un certain seuil.

L'analyse de message est faite par une api extérieure qui est en fait un modèle d'intelligence artificielle entraîné avec un jeu de données approprié.

L'objectif étant d'éviter les situations de harcèlement en entreprise, il faut bien s'assurer qu'un message ou une série de messages est effectivement péjoratif à l'égard d'une personne. On ne peut donc pas confier la décision finale au modèle car ce dernier n'est pas parfait.

Si le nombre de messages signalés pour un utilisateur X atteint un certain nombre, transmettre son nom et lesdits messages à un modérateur qui pourra déterminer s'il doit être sanctionné ou pas. Le fait que la décision finale revienne à un modérateur qui est un être humain est dû au fait que notre modèle pourrait ne pas bien interpréter certains messages, il est alors primordial qu'une personne capable d'interpréter une critique constructive, de l'humour ou encore le second degré dans une conversation puisse analyser le message avant de trancher en faveur ou en défaveur de l'individu concerné.

Les limites :
Chercher des datasets adaptés qui permettent véritablement de déterminer si un message est positif ou négatif, car les modèles ont du mal à analyser l'humour, le second degré ou même juste une critique lors d'un débat.

Chercher des modèles qui ont été entraînés avec ces datasets et vérifier qu'ils interprètent "bien" les messages qui peuvent prêter à confusion.

Un outil pour tester l'efficacité de notre modèle :
Google cloud API

Proposition de datasets :

Les datasets utilisés ont un compilé de commentaires toxiques envoyés sur diverses plateformes comme Youtube, Wikipedia ou encore Chat messages. Ces datasets proviennent majoritairement de Kaggle et Hugging Face qui sont des sites proposant des datasets fiables et libres de droit.

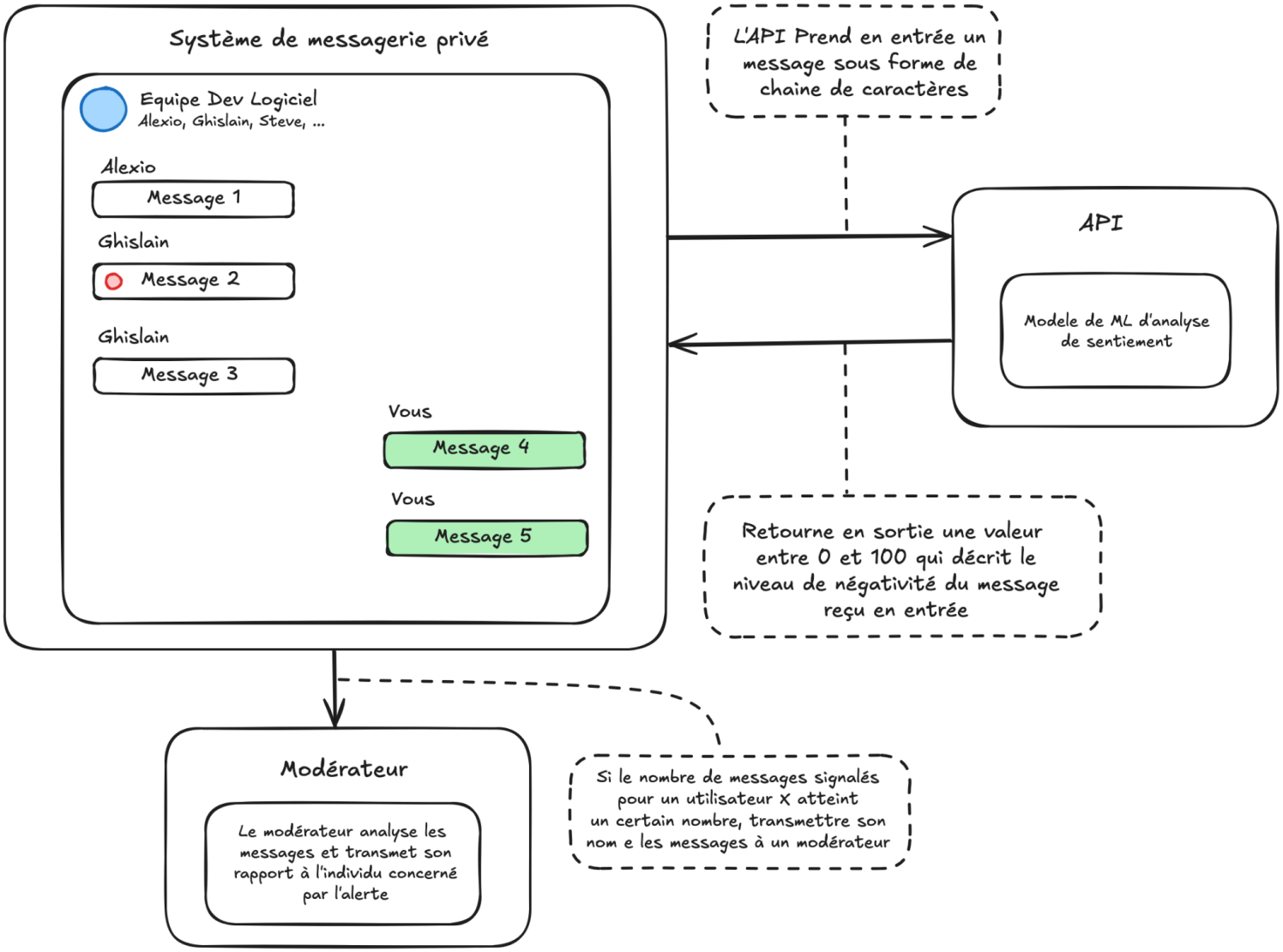
Kaggle

- https://www.kaggle.com/datasets/nursyahrina/chat-sentiment-dataset
- https://www.kaggle.com/datasets/reihanenamdari/youtube-toxicity-data
- https://www.kaggle.com/datasets/get2jawa/toxic-comments-train
- https://www.kaggle.com/c/jigsaw-toxic-comment-classification-challenge/data

Hugging Face

- https://huggingface.co/datasets/ifmain/text-moderation-410K?row=3

Architecture du logiciel :



Méthodes à utiliser

L'application que nous voulons développer entre dans le thème de NLP. Dans ce domaine il y à beaucoup d'options disponibles et notre idée est d'utiliser le transformers pour réaliser cette tâche. L'idée est d'utiliser un modèle BERT comme basis et de faire un fine-tune basé sur notre tâche avec le jeu données que nous avons trouvé. BERT est un modèle très puissant et largement utilisé dans le domaine de NLP et il marche utilisant un architecture encoder-only transformers.

Le projet idéalement sera réalisé à partir d'un modèle BERT pre trained sur la classification du texte, on verra si le modèle sera multilingue ou monolingue, qu'ensuite sera fine tuned pour la tâche de notre intérêt.

Les enjeux environnementaux et sociétaux :

- **Augmentation du stress chez les employés :**
  Etant donné qu'une application de messagerie est d'abord un endroit où on peut échanger librement entre collègues et apporter des critiques constructives pour améliorer la qualité du travail, il est primordial que cet outil ne devienne pas une source de stress pour les employés. Ce qui pourrait arriver si chacun de leurs messages est scruté à la loupe pour détecter le moindre problème.
- **Utilisation détournée de l'application :**
  Les gérants d'une entreprise pourraient par exemple utiliser notre application pour établir une politique de communication trop stricte au sein de leur entreprise, interdisant ainsi l'humour et certains types de messages qui détendent l'atmosphère et améliorent les conditions de vie au travail.