~~ Rapport 2 - Projet IA ~~

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

Projet choisi :

Projet de développement d'une application d'analyse de sentiments pour la signalisation automatique de message dans un système de messagerie privée.

Description détaillée du jeu de données choisi

Parmis les differents jeux de donnees que nous avions mis dans le premier rapport et apres un parcours de ceux-ci, nous avons decide d'utiliser celui de Hugging Face accessible via le lien suivant : https://huggingface.co/datasets/ifmain/text-moderation-410K

Nous avons choisi d'utiliser ce dataset parce qu'il ne donne pas juste des valeurs binaires du style, un message est négatif ou positif, il donne plutôt un degré de négativité du message. Ce dataset contient uniquement des commentaires positifs et négatifs, c'est-a-dire pas de commentaires neutres. Il est donc question d'utiliser notre modèle d'intelligence artificielle entraîné sur ce jeu de données pour prédire la polarité (caractère positif ou négatif) d'un nouveau message.

Ce jeu de données a aussi été choisi parmi les autres alternatives en raison de la qualité des données qu'elle contient. Nous pensons qu'il est plus adapté au type d'application que nous souhaitons créer (messages plus courts, utilisation de l'argot, comportement humain). En plus de cela, le jeu de données contient de nombreux échantillons de données (410 Ko) avec plusieurs formats et structures, ce qui est préférable lors de la construction (entraînement et test) d'un modèle.

Pour chaque échantillon de données (ligne du jeu de données), il est fourni le commentaire/message proprement dit, et les classifications binaires pour chaque type de catégorie toxique, ainsi que les scores de 0 à 1 pour chacune d'entre elles. Voici les différentes métriques dont nous disposons pour chaque ligne du dataset :

- harassment
- harassment_threatening
- hate
- hate_threatening
- self_harm
- self_harm_instructions
- self_harm_intent
- sexual
- sexual_minors
- violence

- violence_graphic

```
“You were right to be embarrassed, casual.”

{
  "categories": {
    "harassment": false,
    "harassment_threatening": false,
    "hate": false,
    "hate_threatening": false,
    "self_harm": false,
    "self_harm_instructions": false,
    "self_harm_intent": false,
    "sexual": false,
    "sexual_minors": false,
    "violence": false,
    "violence_graphic": false
  },
  "category_scores": {
    "harassment": 0.020891988649964333,
    "harassment_threatening": 0.0000451470295956824,
    "hate": 0.00011699183232849464,
    "hate_threatening": 8.157169162359423e-9,
    "self_harm": 4.6329483893714496e-7,
    "self_harm_instructions": 5.776672296065044e-8,
    "self_harm_intent": 2.0666092837018368e-7,
    "sexual": 0.0007244459702633321,
    "sexual_minors": 0.00000897045811143471,
    "violence": 0.0005294209695421159,
    "violence_graphic": 0.00001184129814646439 6
  },
  "flagged": false
}
```

Description des premiers traitements mis en place, ou des premiers algorithmes codé
Pour le moment, on s’est concentré sur un premier traitement des données qu’on utilisera. Nous avons divisé le jeu de données de façon classique, c´est-à-dire 80% pour l'entraînement, 10% pour la validation et 10% pour le test final. Ensuite, on part du modèle pré-entraîné de **DeBERTaV3**, qui est un modèle multilingue développé par Microsoft.

Après, on utilisera Python pour programmer et pytorch comme librarie de support. Nous n’avons pas encore commencé à coder cette partie, mais notre idée pour réaliser le fine-tune est de faire un entrainement avec l’early stopping. C'est-à-dire, à chaque itération (époque) on va enregistrer le modèle si la validation loss s'améliore. Si après x itérations

cette valeur ne s'améliore pas, l’entraînement va s'arrêter et on retourne le dernier modèle enregistré.