Plan prévisionnel

Dataset retenu

Le dataset utilisé est le Mental Health Text Classification Dataset, disponible sur la plateforme Kaggle. Il est constitué de textes rédigés par des utilisateurs partageant leurs expériences personnelles, leurs émotions et leurs difficultés psychologiques.

Chaque texte est associé à une catégorie correspondant à un état ou trouble de santé mentale. Les données présentent une forte hétérogénéité en termes de longueur, de style d'écriture et de vocabulaire utilisé. Le dataset reflète ainsi un cas d'usage réaliste, proche des données rencontrées dans des applications de santé numérique ou de modération de contenu.

Une analyse exploratoire sera menée afin d'étudier la distribution des classes, la longueur des textes et la présence éventuelle de déséquilibres de classes, bien que ces problèmes soient déjà pour l'essentiel déjà pris en compte par le dataset fourni.

Modèle envisagé

Notre démarche est novatrice, pour suivre la grande tendance des embeddings universels de nouvelle génération, nous envisageons de tester le modèle BGE-M3 qui sera utilisé comme extracteur de features, d'embeddings, en effet la performance en NLP vient surtout des embeddings, qui cherchent a expliciter plus que du modèle final.

Ces modèles ont été conçus pour intégrer explicitement l'intention de la tâche dans la représentation du texte, ce qui constitue une avancée majeure par rapport aux approches antérieures.

Références bibliographiques

Références bibliographiques (2–3)
Devlin et al. (2019) – BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding
Référence de base pour situer les approches classiques.

Reimers & Gurevych (2019) – Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks
Approche largement utilisée sur ce type de dataset.

Wang et al. (2023) – BGE: General Embeddings from Large Language Models
Modèles d'embeddings récents, instruction-tuned, utilisés comme référence principale dans ce projet.

Explication de la démarche de test du nouvel algorithme

Notre démarche sera …

D'abord, une rapide analyse exploratoire des données, le dataset étant déjà « très abouti » cette étape sera rapide. Mais ce sera aussi l'occasion de se pencher sur ce qui a déjà été fait en terme de features engineering sur ce dataset.

Ensuite nous testerons en parallèle une baseline classique utilisant SBERT pour extraire les embeddings associé à une Régression Logistique et LightGBM comme classificateurs.

Pour valider le bien fondé d'un nouvelle extracteur d'embeddings, nous reproduirions exactement le même enchainement de traitements (pipeline) avec le nouveau modèle testé.

Donc nous réaliserons une extraction des caractéristiques (features), ici le sens du texte, avec le modèle BGE-M3 en remplacement de SBERT, le reste du pipeline étant préservé. D'abord une régression logistique puis un LightGBM comme classificateurs.

Les autres étapes auront pour objectif, une évaluation comparative des deux solutions comprenant ; la comparaison des performances quantitatives, l'analyse des confusions entre classes proches, l'étude qualitative des erreurs et une analyse critique permettront d'évaluer les apports réels du modèle BGE-M3 plus actuel.

Ainsi, ce projet vise à démontrer l'intérêt des modèles NLP récents dans un contexte applicatif sensible, en mettant en évidence les apports concrets des embeddings instruction-tuned et des LLM modernes face aux approches classiques.
Il s'inscrit dans une démarche expérimentale rigoureuse, comparative et conforme aux pratiques actuelles de la R&D en NLP.