ÉNONCÉ DU MINI-PROJET 3IDL : Détection d'Émotions dans les Textes avec Deep Learning

Contexte du Projet

Vous êtes ingénieur en Data Science chez une entreprise de médias sociaux qui souhaite améliorer la compréhension des émotions des utilisateurs. Votre mission est de développer un système avancé de détection d'émotions à partir de textes en utilisant le dataset GoEmotions.

Objectifs Principaux

1. **Prétraitement** des données textuelles et préparation du dataset GoEmotions
2. **Conception et implémentation** d'architectures de deep learning pour la classification multi-label
3. **Évaluation comparative** des différentes architectures
4. **Analyse d'explicabilité** des prédictions du modèle

Dataset GoEmotions

Le dataset contient :

- 58,000 commentaires Reddit
- 27 catégories d'émotions + 1 catégorie neutre
- Annotation multi-label (plusieurs émotions possibles par texte)
- Répartition train/validation/test

Tâches Détaillées

Partie 1 : Analyse et Prétraitement des Données

- Analyse exploratoire de la distribution des émotions
- Nettoyage et tokenisation des textes
- Gestion du déséquilibre des classes
- Préparation des embeddings (TF-IDF, Glove, FastText, BERT)

Partie 2 : Architectures à Implémenter

1. **Modèle de Base** : LSTM simple
2. **Modèle Intermédiaire** : BiLSTM avec mécanisme d’attention
3. **Modèle Avancé** : Architecture hybride CNN-BiLSTM avec mécanisme d'attention
4. **Modèle Transformer** : BERT-base

Notez bien :

- Le modèle doit respecter la nature des labels du jeu de données GoEmotions. Autrement dit, chaque texte doit être traité en **multi-label**, et lors de l'évaluation, une prédiction n'est considérée correcte que si **l'ensemble des labels associés au texte est exactement conforme** aux labels réels.
- À la fin de l'entraînement, le modèle doit être sauvegardé (par exemple au format **.pickle**) afin de pouvoir être facilement chargé et testé ultérieurement.

Partie 3 : Protocole d'Évaluation

- **Métriques** : Precision, Recall, F1-score (micro/macro), Hamming Loss, AUC-ROC
- **Validation croisée** : Split 80-10-10
- **Benchmark** : Comparaison avec modèles baseline

Partie 4 : Étude d'Ablation

Pour l'architecture hybride, évaluer l'impact de :

- Mécanisme d'attention
- Couches CNN vs LSTM
- Différents types d'embedding
- Techniques de régularisation

Partie 5 : Analyse d'Explicabilité

- Utilisation de LIME/SHAP pour l'interprétation
- Visualisation des poids d'attention
- Analyse des erreurs de classification

Livrables Attendus

1. **Rapport technique** (10-15 pages) (modèle latex de snarticle)
2. **Code source** commenté
3. **Présentation** des résultats (15 min)
4. **Démonstration** interactive du modèle

Durée : 5 semaines

Option : Toute implémentation intégrée dans une interface (JS, Streamlit, etc.) sera prise en compte dans l'évaluation finale.