DataSpace

Plan prévisionnel

Dataset retenu

Le dataset utilisé pour ce projet comprend 50 000 questions extraites du site Stack Overflow via l'outil StackEchange Data Explore. Stack Overflow est une plateforme populaire pour les développeurs où ils peuvent poser des questions et obtenir des réponses sur divers sujets de programmation. Les questions sont généralement bien structurées, incluant des titres (Title), des descriptions détaillées (Body), et parfois des étiquettes (Tags) qui indiquent les sujets abordés.

Modèle envisagé

Pour améliorer les performances initialement faibles obtenues avec le classifieur de base (logic classifier), le modèle SciBERT-CNN a été envisagé.

Des études montrent que les modèles basés sur BERT, et particulièrement ses variantes adaptées à des corpus spécifiques comme SciBERT, surpassent généralement les autres modèles de traitement du langage naturel en termes de précision et de compréhension contextuelle.

Les CNN ont prouvé leur efficacité dans l'extraction de caractéristiques pertinentes des séquences textuelles, ce qui les rend complémentaires aux capacités de contextualisation de SciBERT.

L'objectif principal de l'algorithme SciBERT-CNN est de permettre de capter les relations sémantiques et de résoudre les déséquilibres de classe inhérents.

Dans ce contexte, il est utilisé pour :

Identifier les catégories des questions (par exemple, langages de programmation, frameworks spécifiques, concepts algorithmiques).

Aider à la recherche et à la récupération d'informations en classant les questions dans des catégories pertinentes.

Améliorer les systèmes de recommandation et les moteurs de recherche au sein de plateformes similaires à Stack Overflow.

Références bibliographiques

https://arxiv.org/pdf/2404.13078

https://www.sciencedirect.com/science/article/pii/S187705092300234X

https://www.kaggle.com/code/giovanimachado/hate-speech-bert-cnn-and-bert-mlp-in-tensorflow

Explication de votre démarche de test du nouvel algorithme (votre preuve de concept)

- **Méthode Baseline**

La méthode baseline utilisée est un logic classifier. Ce modèle a été choisi pour sa simplicité et sa rapidité de mise en œuvre. Nous avons entraîné ce modèle en utilisant une division des données en ensembles d'entraînement (80%) et de test (20%), et avons évalué ses performances en utilisant des métriques telles que l'accuracy et le F1-score. Les résultats obtenus montrent une accuracy de 55% et un F1-score de 0.66, ce qui indique des performances limitées, surtout sur les classes minoritaires.

- **Méthode Proposée**

Pour améliorer ces performances, nous proposons d'utiliser scibert-cnn, un modèle basé sur BERT, spécialement pré-entraîné sur des corpus scientifiques, et combiné avec des réseaux de neurones convolutifs (CNN) pour capturer des relations locales dans les données textuelles.

Prétraitement des données : Les questions de Stack Overflow seront tokenisées et encodées en utilisant la tokenizer de SciBERT.

Entraînement du modèle : Le modèle sera entraîné avec une division des données similaire à celle utilisée pour le logistic classifier. Nous ajusterons les hyperparamètres (learning rate, batch size) et utiliserons des techniques de régularisation telles que le dropout.

Évaluation et comparaison : Les performances de scibert-cnn seront évaluées en utilisant les mêmes métriques que pour la baseline. Nous nous attendons à voir une amélioration significative de l'accuracy et du F1-score, surtout pour les classes minoritaires.