Tweet Similarity Analysis with Transformer Embeddings

Réaliser par :

Djeghdjegh Sara

M'ziane Houda

Makeri Atika

Introduction:

Ce projet vise à créer un modèle de similarité de tweets basé sur les embeddings de mots en utilisant le modèle Word2Vec pré-entraîné et la distance de Manhattan. L'objectif principal est de déterminer si deux tweets sont similaires, c'est-à-dire s'ils proviennent du même utilisateur ou non.

Méthodologie:

Dans ce projet, nous avons utilisé un modèle Word2Vec pré-entraîné pour obtenir les embeddings de mots des tweets. Ensuite, nous avons calculé la similarité entre les paires de tweets en utilisant la distance de Manhattan entre leurs embeddings. Nous avons échantillonné des paires du même utilisateur et des paires d'utilisateurs différents à partir des données disponibles. Enfin, nous avons utilisé ces paires pour entraîner et évaluer notre modèle.

Remarque : On a diminué les deux dataset car notre PC est incapable de gérer cette immense quantité de données.

Architecture du modèle:

Notre modèle comprend les étapes suivantes :

Prétraitement du texte : Convertir en minuscules, suppression de la ponctuation, des symboles et des caractères spéciaux, jetonisation, suppression des mots vides et lemmatisation.

Calcul des embeddings de texte : Utilisation du modèle Word2Vec pour obtenir les embeddings de mots des tweets.

Calcul de la similarité : Utilisation de la distance de Manhattan pour calculer la similarité entre les paires de tweets.

Prédiction des étiquettes de similarité : Utilisation d'une valeur de seuil pour prédire si deux tweets sont similaires ou non.

Résultats:

Sur la base des données de test, notre modèle a obtenu les performances suivantes :

```
Training Precision: 0.49317738791423
Training Recall: 0.9729833669839438
Training F1 Score: 0.654571326929918

Testing Precision: 0.4917916617456006
Testing Recall: 0.9676969556123635
Testing F1 Score: 0.6521534847298356
```

Ces résultats indiquent que notre modèle est capable de prédire la similarité entre les tweets avec une précision, un rappel et un score F1 raisonnables.

Ce rapport résume le processus suivi pour développer le modèle de similarité de tweets et présente les performances obtenues lors de l'évaluation du modèle sur les données de test.