haute école neuchâtel berne jura arc ingénierie www.he-arc.ch

Hes·so Haute Ecole Spécialisée de Suisse occidentale

News Clustering

David DARMANGER

Travail de Bachelor 2023

Filière Informatique et Systèmes de Communication - Orientation Ingénierie des données

Professeur: Jonathan DREYER

Expert: Jérôme MORET

Description

Ce projet se concentre sur l'exploration et l'évaluation de l'utilisation de Sentence-BERT (S-BERT), une variante optimisée du modèle de langage BERT pour la comparaison de similarités sémantiques entre textes, appliquée à un large ensemble de données d'articles financiers.

L'objectif principal du projet est d'explorer l'efficacité de S-BERT pour encoder des textes en se basant sur une notion de similarité, avec l'ambition de regrouper ces articles par similarité de contenu, puis d'ouvrir la voie à des analyses avancées. De plus, le projet vise à évaluer si l'utilisation de S-BERT offrait une valeur ajoutée tangible par rapport aux autres techniques d'encodage de texte.

Les objectifs spécifiques sont:

- Encodage de texte : utiliser un réseau neuronal siamois pour l'encodage basé sur une notion de distance.
- Regroupement d'articles : utiliser les embeddings générés pour classer les articles par similarité.
- Cas pratique : implémenter une visualisation des regroupements de news ou la prédiction de sujets d'articles.

Déroulement

Différentes phases du projet :

- Analyse des données de news et étude des réseaux neuronaux siamois, NLP, BERT et Sentence-Bert.
- Évaluation de l'état de l'art
- Prise en main de techniques NLP avec les Transformers et sélection de l'approche optimale.
- Conception d'un modèle siamois pour comparer les news, incluant l'encodage textuel et prédiction de similarités.
- Tests de performance du modèle, prédictions de sujets de news, et visualisation des sujets tendance.

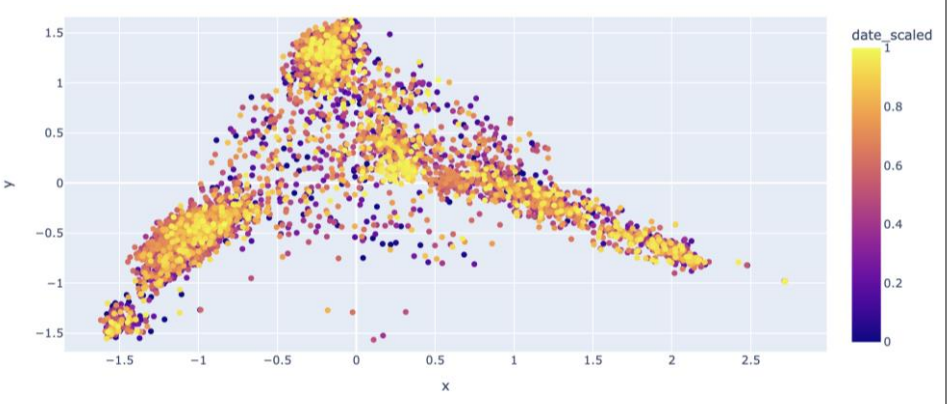
Résultats

Le modèle S-BERT, spécialement entraîné pour incorporer la notion de similarité, a montré une compétence plus élevée que d'autres modèles pour extraire des informations pertinentes dans le but de créer des encodages avancés. De plus, la prédiction de sujets s'est révélée efficace, même face à un grand nombre de classes.

model	accuracy	recall	precision	f1	sujet par news	nombre de news	nombre de classes
fine-tuned	76%	89%	94%	91%	3	141965	109
without fine-tuning	54%	78%	91%	83%	3	141965	109
random frequency-based	4%	51%	53%	50%	3	141965	109
tf-idf	66%	83%	89%	85%	3	141965	109

Résultats de l'évaluation par classification

Les encodages d'articles peuvent être représentés avec précision dans un espace bidimensionnel, ouvrant ainsi la porte à des analyses approfondies et à une variété de cas d'utilisation. Ces cas incluent la détection de tendances ainsi que la visualisation de groupes d'articles.



Visualisation des embeddings

Discussion : Conclusions et perspectives

L'application de S-BERT à un ensemble de données d'articles financiers a démontré son potentiel en termes d'encodage de texte et de prédiction de sujets. Ce modèle offre un avantage concret par rapport aux autres techniques d'encodage de texte et ouvre la voie à des analyses et des utilisations plus avancées telles que la détection de plagiat ou d'articles dupliqués, l'analyse des tendances actualités, recherche d'information et système de recommandation ou encore la surveillance de la réputation d'entreprise.