Auteurs :

1- Clemente Coriande, c.clemente@quinten-health.com, Quinten Health
2- Mélissa Rollot, m.rollot@quinten-health.com, Quinten Health

Titre

Une approche de traitement du langage naturel (NLP) pour automatiser l'analyse de témoignages de patients

Introduction

Les témoignages des patients fournissent des informations précieuses pour définir et caractériser la perception de leur maladie.

Cependant, les méthodes traditionnelles d'analyse automatique de texte, basées sur la fréquence de cooccurrence des mots, ne sont pas adaptées aux textes courts.

Nous proposons une approche bout-en-bout, dérivée de travaux précédents [1][2] et basée sur la proximité sémantique, permettant d'identifier et de visualiser les principaux sujets les tendances parmi les témoignages de patients.

Méthode

Tout d'abord, les témoignages sont pré-traités et vectorisés avec le modèle pré-entraîné Sentence-BERT[3], capturant le sens des textes au-delà de la cooccurrence des mots. Pour faciliter l'interprétation, la dimension des vecteurs est réduite à deux en utilisant l'algorithme UMAP[4].

Ensuite, un clustering hiérarchique agglomératif est effectué sur les nouveaux vecteurs. Le dendrogramme de clustering facilite le post-traitement en présélectionnant automatiquement les clusters qui peuvent être fusionnés ou divisés en deux sous-clusters en fonction de leur proximité sémantique.

Chaque cluster est ensuite caractérisé par ses termes les plus fréquents et une analyse de sentiments.

Résultats

Testée sur des témoignages de patients d'une longueur moyenne de 15 mots, cette méthode fournit des sujets plus cohérents que les approches de l'état de l'art. Le post-traitement amélioré du clustering rend la méthodologie plus rapide à exécuter et plus évolutive.

Conclusions

Notre méthode permet d'extraire des sujets cohérents à partir d'un grand volume de textes courts de manière automatisée et rapide. Elle permet de mieux comprendre la perception des patients sur leur maladie et sa prise en charge.

[1]M. Grootendorst, « BERTopic: Neural topic modeling with a class-based TF-IDF procedure », arXiv, arXiv:2203.05794, 2022, march, doi: 10.48550/arXiv.2203.05794.

[2] L.Deplante, P.Hayat et M.Rollot "Une nouvelle approche de traitement automatique des réponses de questionnaires patients", Afrco, 2022, June.

[3] N. Reimers et I. Gurevych, « Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks », ArXiv190810084 Cs, 2019, August.

[4] L. McInnes, J. Healy, et J. Melville, « UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction », ArXiv180203426 Cs Stat, 2020, September.