quinten health

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

Coriande Clemente
Quinten Health

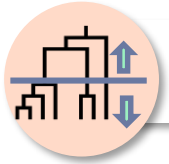
Journées de Biostatistique 2022
17-18 nov. 2022 Rennes (France)

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

Développement d'une nouvelle approche et application aux opiacés forts

Contexte du projet

Appliquer la méthodologie à de nouvelles données



Améliorer la méthodologie et faciliter l’application de la méthodologie

Extraire des résultats métiers

Conclusion

Contexte du projet

L’extraction automatique d’information à partir de témoignages patients

Une génération d’informations médicales plus rapide et plus fiable

Enjeux de l’exploitation de données de témoignages patients

- Améliorer la compréhension des prises en charge
- Mieux comprendre une maladie
- Recenser les besoins de patients
- Comprendre comment est perçu un médicament
- Identifier les effets secondaire d’un traitement
- Anticiper une épidémie
- …

Intérêt de l’automatisation de l’extraction

- Extraire de nouvelles informations (importante volumétrie de données)
- Gain de temps
- Gain en fiabilité

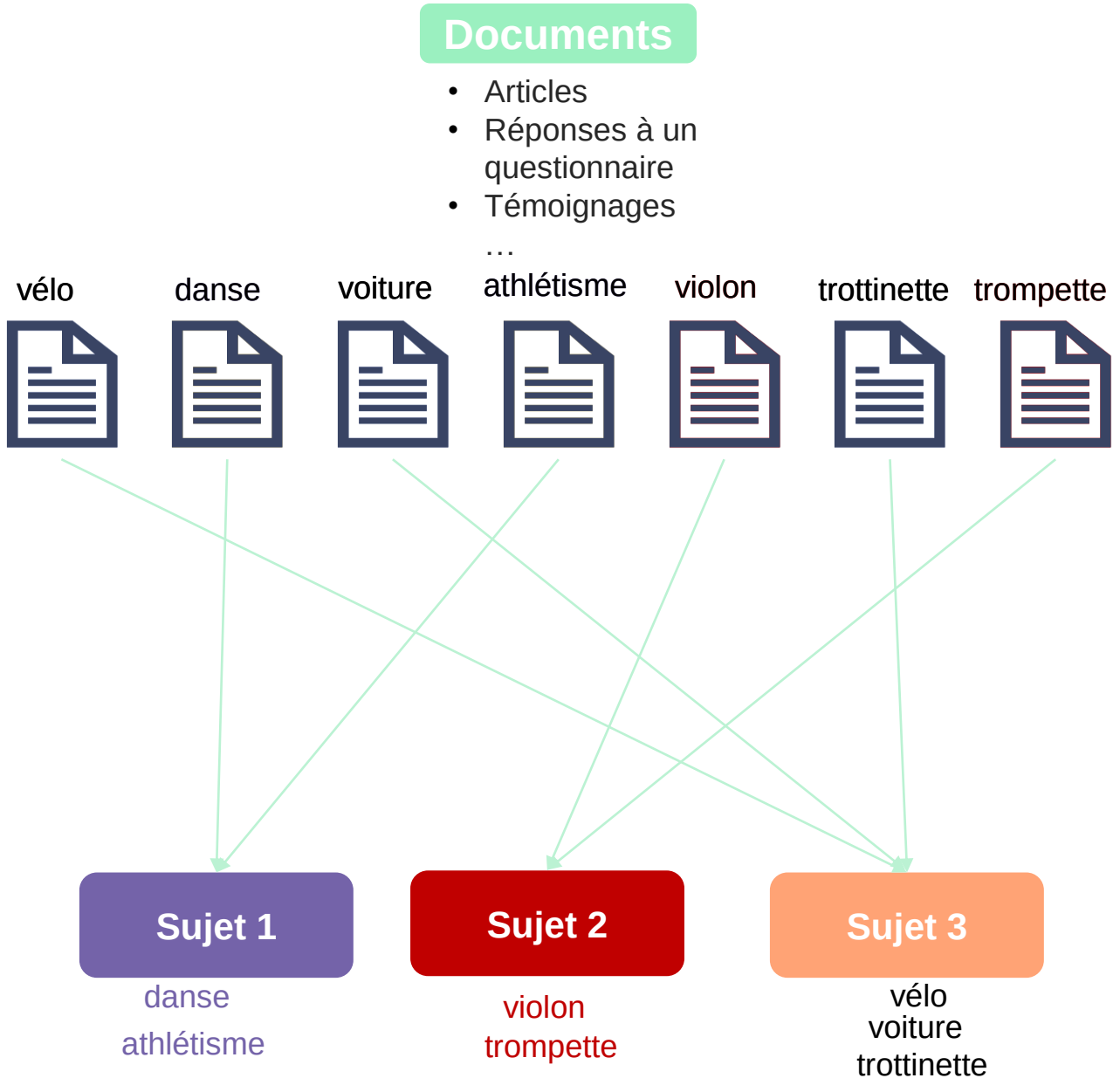
Une méthode d'extraction automatique d'information: Le topic modeling

Objectif : Extraction automatique des thèmes sous-jacents de documents textuels de façon non supervisée

Remarque :
Les topics models ne nomment pas les thèmes

Méthodes classiques :
- Latent Dirichlet Allocation (LDA),
- Non-negative matrix factorization (NMF)

Basées sur la co-occurrence des mots



Contexte du projet initial

Une méthodologie de topic modeling adaptée aux textes courts

Contexte

Projet PX :

- **Extraction automatique** des thèmes de réponses patients sur l'observance de leur traitement
- **Topic modeling :** méthodes classiques non adaptées aux textes courts
- **Nouvelle méthodologie** adaptée aux textes courts développée pour le projet initial
- **Présentation** d'un poster à l'AFCRO*

quinten health

*L.Deplante, P.Hayat et M.Rollot "Une nouvelle approche de traitement automatique des réponses de questionnaires patients", Afrco, juin 2022 https://fr.calameo.com/read/0067593919a3a22a3f318

La méthodologie initiale : topic modeling adapté aux textes courts

Chaque étape pensée pour un topic modeling de textes courts

Données nettoyées

« My treatment… »

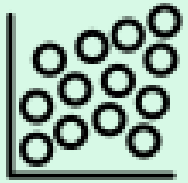
Transformation des phrases

Vectorisation avec Sentence Bert*

[0.09, -0.87 … 0.32]

Concentration du signal

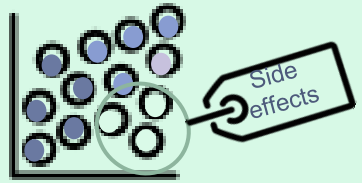
Réduction de dimensions avec UMAP 2D**



[0.19, -0.80]

Regroupement des textes

Modélisation avec un Clustering Hiérarchique***



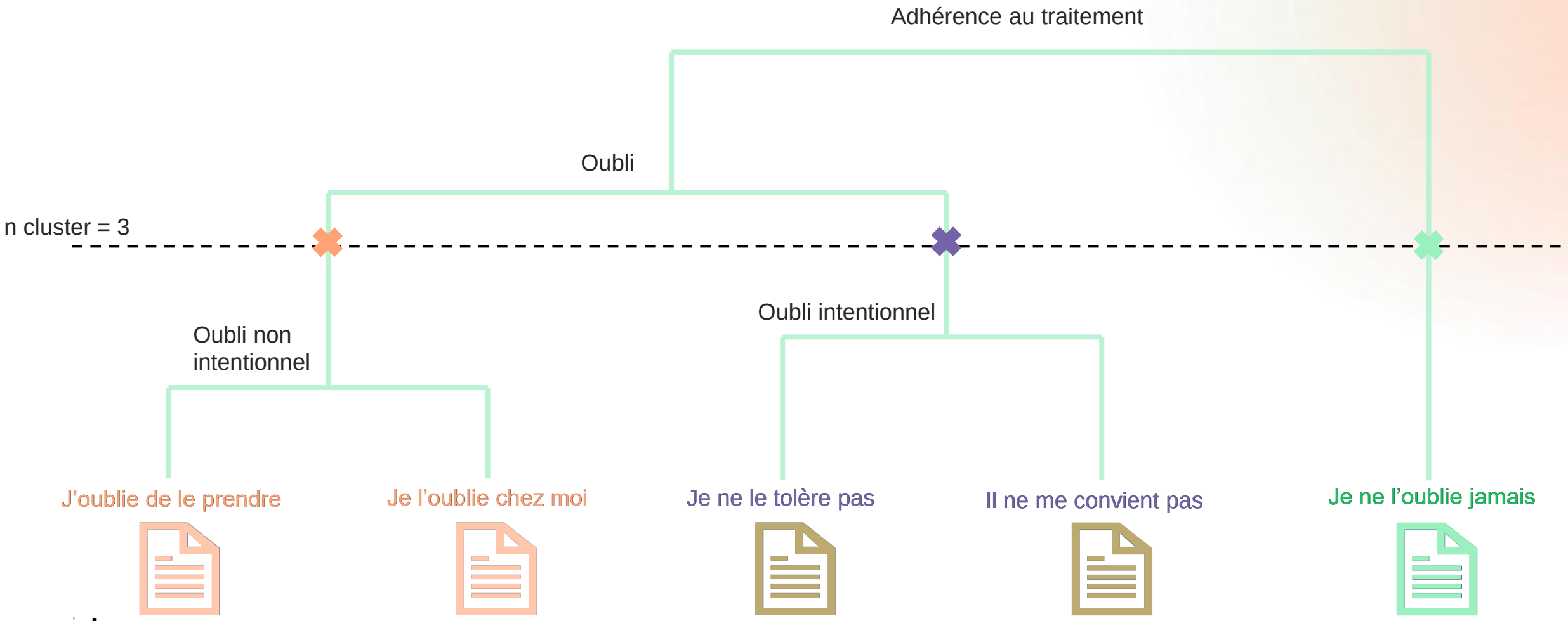
Cluster: side effects

* sbert.net, 2022, Nils Reimers
** L. McInnes, J. Healy, et J. Melville, « UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction », ArXiv180203426 Cs Stat, 2020, September.
*** Fahad, A., Alshatri, N., Tari, Z., Alamri, A., Khalil, I., Zomaya, A. Y., ... & Bouras, A., A survey of clustering algorithms for big data: Taxonomy and empirical analysis. IEEE transactions on emerging topics in computing, 2(3), 267-279, 2014

Zoom sur le clustering hiérarchique

Clustering basé sur le dendrogramme

Regroupement des textes



quinten health

La méthodologie initiale : topic modeling adapté aux textes courts

Adaptation, validation et amélioration de la méthodologie de Topic Modeling initiale pour la valorisation de témoignages de patients

Limites de la méthodologie initiale

- Utilisée uniquement pour le projet
  → Non validée sur d'autres données
- Est-elle optimale?
- Une partie de la méthodologie est manuelle

Objectifs du projet R&D

- **Appliquer et valider la méthodologie sur de nouvelles données**
- **Améliorer la méthodologie**
- **Faciliter l'application de la méthodologie et la rendre plus généralisable**
- **Extraire des résultats métiers**

Appliquer la méthodologie à de nouvelles données

Choisir des données et adapter la méthodologie

Traitement d’intérêt : les antalgiques de palier 3

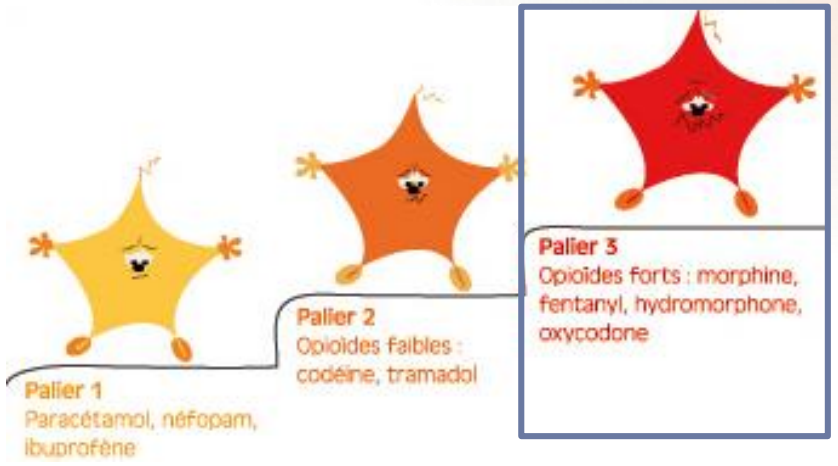
Besoin d’une meilleure compréhension de l’expérience patient dans le cadre de la prise d’opiacés forts

Vocabulaire:
- Opiacés : dérivés d'origine naturelle du pavot
- Opioïdes : composés semi-synthétiques ou synthétiques (par abus de langage, inclue les opiacés)

Quand sont-ils prescrits ?
- Pallier 3 des antalgiques
- Douleurs intenses ou non soulagées par les autres antalgiques
- Douleurs chroniques

Dangers potentiels connus :
- Effets indésirables (troubles digestifs, confusion, sédation…)
- Addictions

Conséquence : Difficulté de prescription
- Rapport bénéfice/risque évalué avec précision

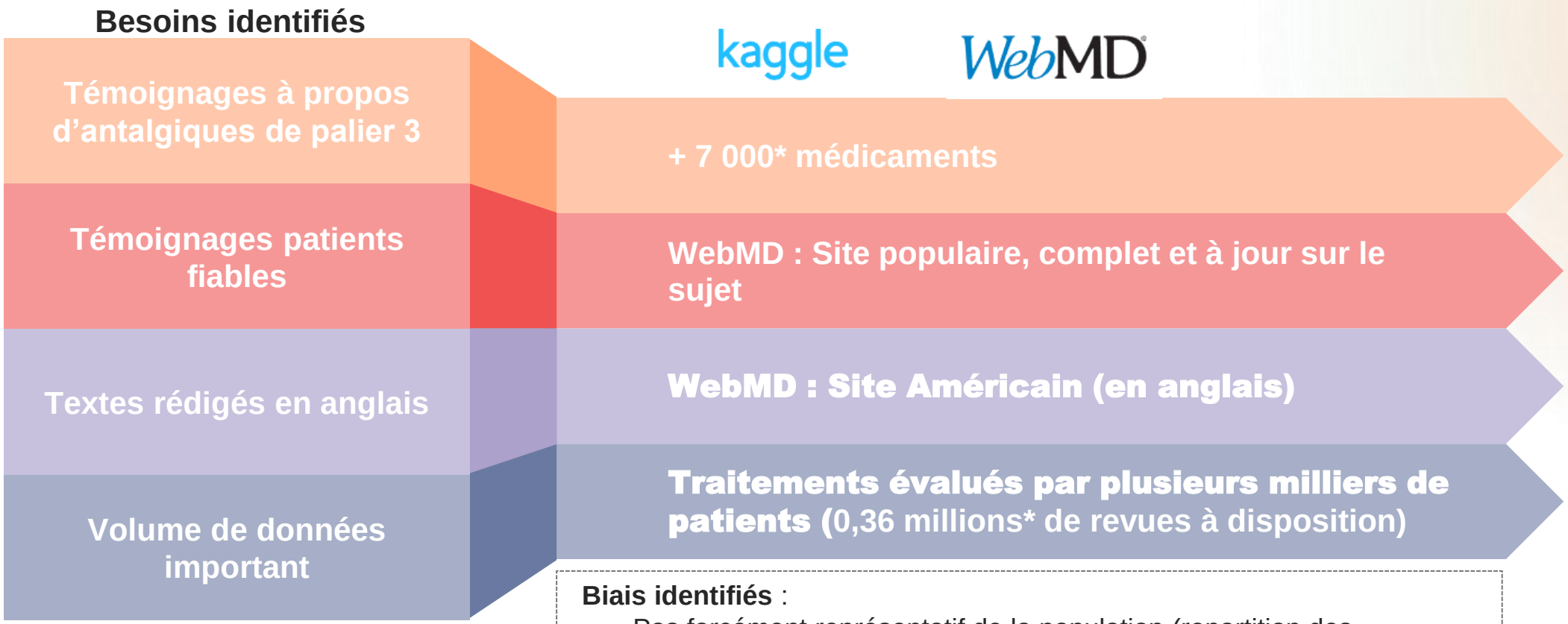


Pyramide des antalgiques

Problématique : Comment identifier les sujets de témoignages de patients dans le cadre de leur prise d’opioïdes « forts » ?

Choix de la source de données : un enjeu majeur du projet

WebMD : Une source de données qui répond aux besoins identifiés



Biais identifiés :

- Pas forcément représentatif de la population (repartition des utilisateurs en terme de sexe, age…)
- Eventuelle censure au niveau des commentaires (ex : difficile de poster à propos de certains effets secondaires)

quinten health

*chiffres basés sur les données mises à disposition par kaggle : https://www.kaggle.com/datasets/rohanharode07/webmd-drug-reviews-dataset

Collection de commentaires patients sur le site WebMD

Objectif : Données de qualité (forme, fond, postés par des patients…)

Données nettoyées

WebMD

362 806

Extraction des commentaires d'intérêt du kaggle WebMD

- Liste de traitements
- Commentaires

Nombre moyen de mots par commentaire : 64

1 573

Reconstruction

- Découpage par phrase
- Règles de selection des références au traitement

Nombre moyen de mots par commentaire : 15

4 474

Nettoyage et aggrégation

- Nettoyage "rapide"

4 474

Exemple (factice)

« Bonjour à tous. Après des années de mal de dos, mon médecin m'a prescrit des opioïdes forts **#Opioides.** Depuis, je peux enfin reprendre le travail. »

« Après des années de mal de dos, mon médecin m'a prescrit des opioïdes forts **#Opioides.**»

« Depuis, je n'ai aucun effet secondaire. »

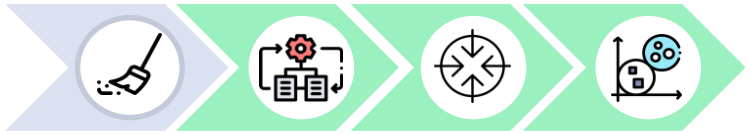
« Après des années de mal de dos, mon médecin m'a prescrit des opioïdes forts. »

Depuis, je peux enfin reprendre le travail. »

quinten health

Premiers résultats

Application de la méthodologie initiale aux données WebMD



Résultats
Identification de sujets de témoignage

Limites
Labélisation difficile :

- Présence de valeurs aberrantes dans les clusters
- Tâche manuelle fastidieuse

Certains clusters sont similaires
Parfois plusieurs idées dans un même cluster

Améliorer la méthodologie et faciliter son application

Rechercher d’éventuelles optimisations
Automatiser la partie manuelle

Multiples itérations afin de répondre aux axes d'amélioration identifiés

A ce stade, les itérations n'ont pas permis de lever les limites de la méthode initiale

Données nettoyées

« My treatment… »

Transformation des phrases

Vectorisation avec Sentence Bert

[0.09, -0.87 … 0.32]

Concentration du signal

Réduction de dimensions avec UMAP 2D

[0.19, -0.80]

Regroupement des textes

Modélisation avec un Clustering Hiérarchique

Classe : side effects

Enjeux	**Jouer sur la retranscription des mots, les liens entre eux, la sémantique**	**Jouer sur la quantité d'information conservée et le nombre de dimensions**	**Jouer sur la cohérence des clusters et leur nombre**
Tests	Moyenne des vecteurs de Biobert →Sens de la phrase mal retranscrit	Analyse en Composantes Principales (ACP) → Pas de nouveaux clusters, beaucoup de dimensions	K-medoids →Hiérarchique plus performant (silhouette)
		Interversion clustering hiérarchique et UMAP → Pas de nouveaux clusters, moins visuel donc difficile à interpréter	
	Application d'une méthode classique : LDA → Phrases courtes non adaptées		

quinten health

Zoom sur une itération : changement du calcul des vecteurs

Objectif : mieux capter les termes médicaux → obtenir des clusters plus fiables

Changement de vectorisation

Données nettoyées

« My treatment… »

Transformation des phrases

Vectorisation avec BioBert*

[0.09, -0.87 … 0.32]

Concentration du signal

Réduction de dimensions avec UMAP 2D

[0.19, -0.80]

Regroupement des textes

Modélisation avec un Clustering Hiérarchique

Classe : side effects

Biobert :

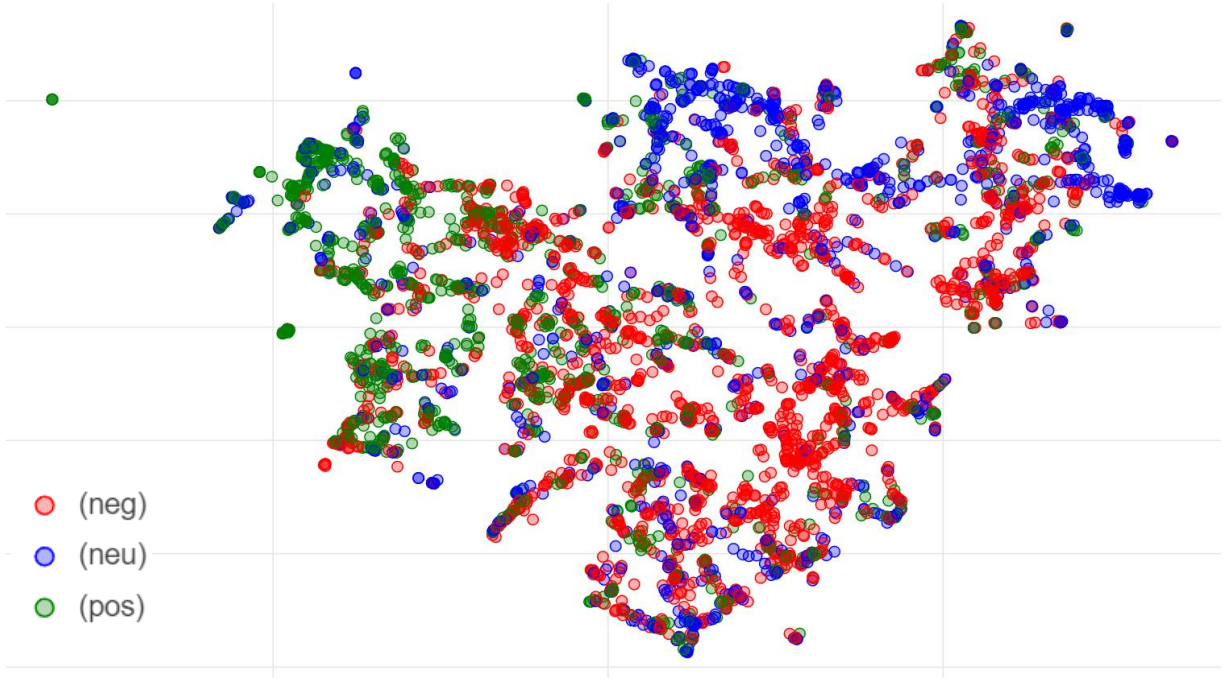
- Algorithme pré-entrainé sur des textes médicaux
- Vectorise les mots et non les phrases ➔ Besoin d'appliquer une moyenne pour obtenir le vecteur de la phrase

*J. Lee et al, BioBERT: a pre-trained biomedical language representation model for biomedical text mining

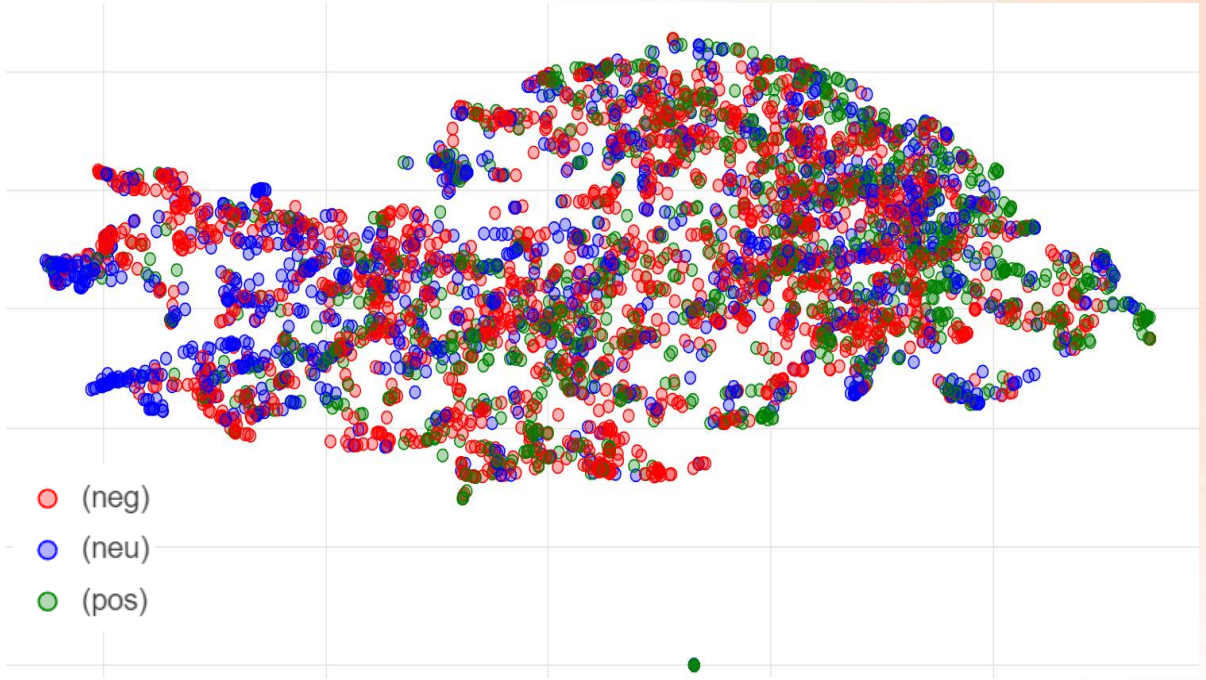
Zoom sur une itération : changement du calcul des vecteurs

Objectif : mieux capter les termes médicaux → obtenir des clusters plus fiables

Transformation des phrases



Analyse de sentiment sur les vecteurs de sentence Bert



Analyse de sentiment sur les vecteurs de Biobert

Conclusions

- Sens de la phrase moins pris en compte
- Pas de meilleurs clusters
- Certains clusters toujours difficiles à interpréter

quinten health

La méthodologie améliorée finale : une réponse aux limites de la méthodologie initiale

Objectif : définir une méthode reproductible et fiable, regrouper les commentaires du même thème sous un même label

Données nettoyées

« My treatment… »

Transformation des phrases

Vectorisation avec Sentence Bert

[0.09, -0.87 … 0.32]

Concentration du signal

Réduction de dimensions avec UMAP 2D

[0.19, -0.80]

Regroupement des textes

1- Clustering

Clustering hiérarchique basé sur un maximum du score de silhouette

2- Suppression des valeurs aberrantes

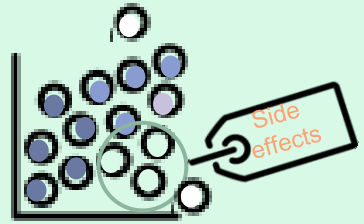
LocalOutlierFactor

3- Post processing semi-automatisé

Choix manuel : aperçu des commentaires, tf-idf, analyse de sentiments

4- Labélisation

Choix manuel : basé sur tf-idf
Regroupement par thèmes



Classe : side effects

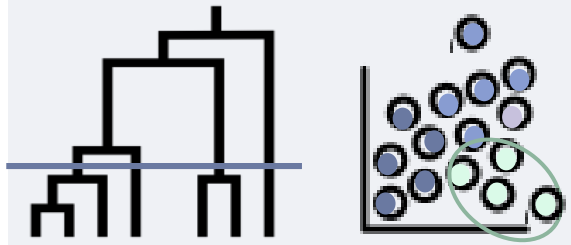
Méthodologie améliorée finale : détail des étapes du clustering

Objectif : définir une méthode reproductible et fiable, regrouper les commentaires du même thème sous un même label

Regroupement des textes

1- Clustering

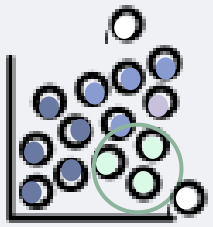
Clustering hiérarchique basé sur un maximum du score de silhouette



Labels faussés par des valeurs aberrantes

2- Suppression des valeurs aberrantes

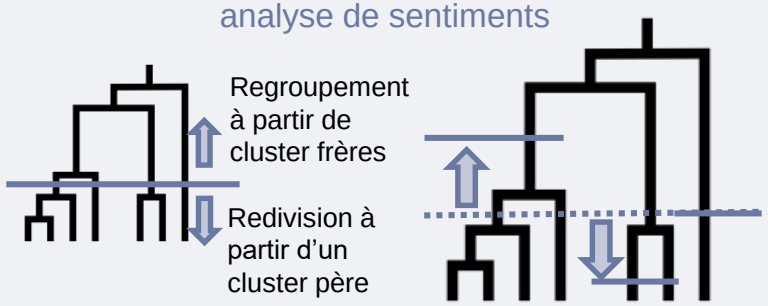
LocalOutlierFactor*

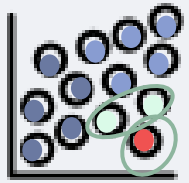


- Plusieurs idées dans un cluster
- Même idée dans plusieurs clusters

3- Regroupements et redivisions

Choix manuel : aperçu des commentaires, tf-idf, analyse de sentiments

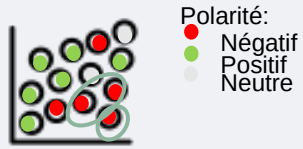




Identification de thèmes

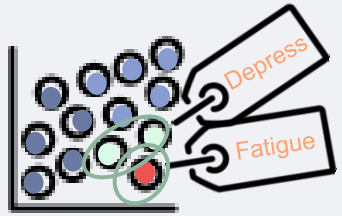
4- Labélisation

Analyse de sentiments avec SentimentIntensityAnalyzer*



Nuage de mots des TF-IDF

Labélisation finale manuelle



*https://www.nltk.org/api/nltk.sentiment.html

Extraire des résultats métiers

Identifier automatiquement des sujets de témoignages patients à propos de leur traitement

Résultats de la méthode améliorée

4474 commentaires issus de WebMD

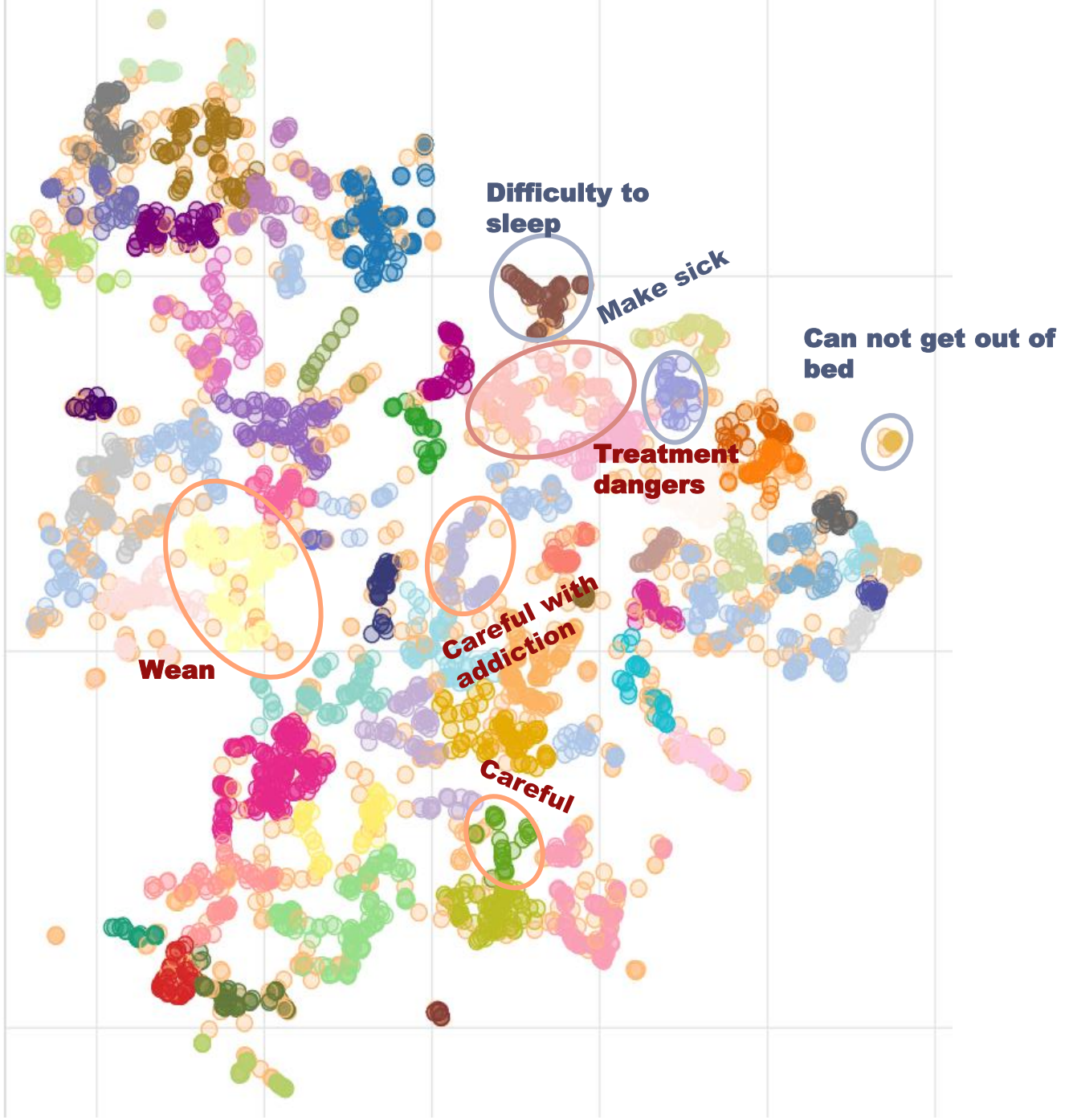
15% de valeurs aberrantes
12% « messy » clusters

60 clusters + messy + outliers

9 thèmes principaux + messy + outliers

Commentaires par cluster :
- Moyenne : 72
- Min : 5 (god sent treatment)
- Max : 165 (effective pain and chronic pain treatment)

Commentaires par thème général:
- Moyenne : 407
- Min : 67 (system)
- Max : 908 (treatment effectivness)



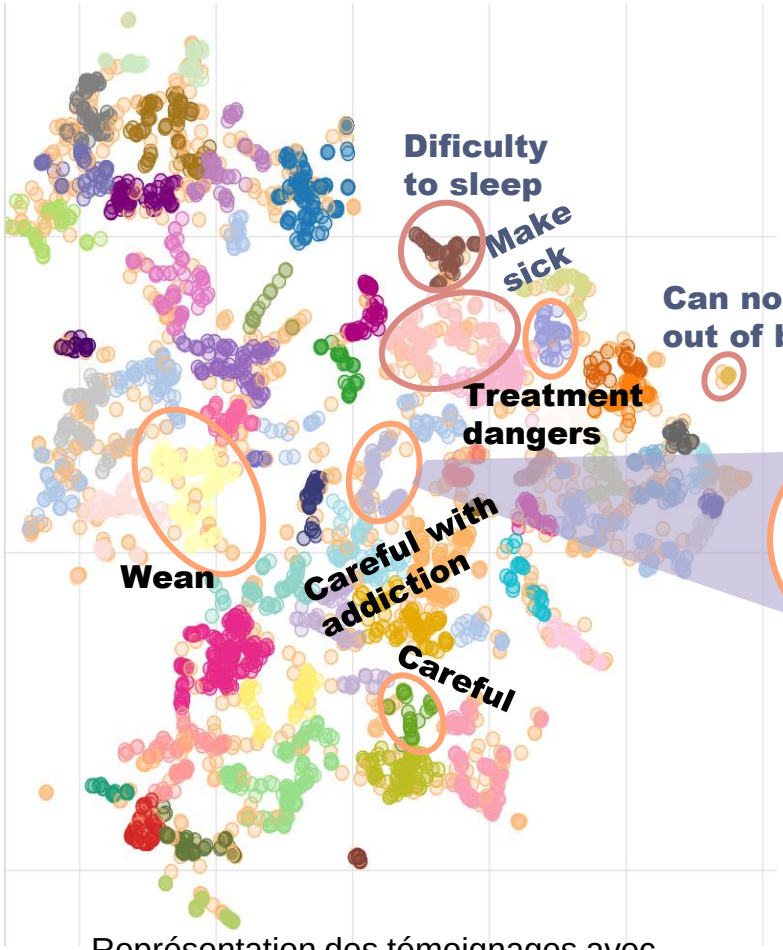
quinten health

Prévalence des différents thèmes identifiés hors outliers

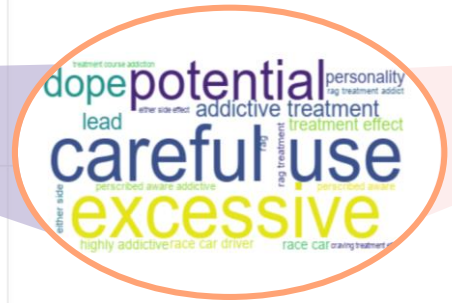
Thèmes principaux: efficacité de traitement, éventuels dangers et effets secondaires

3739 commentaires*

*15% d'outliers exclus

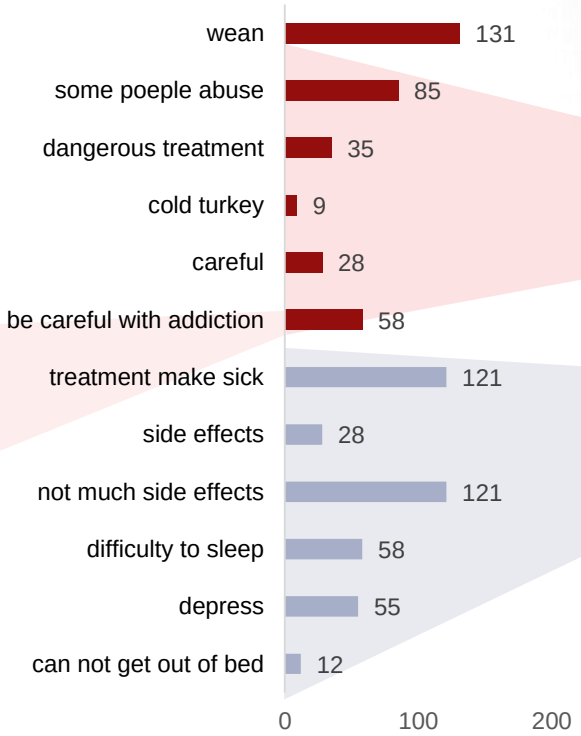


Représentation des témoignages avec UMAP 2D

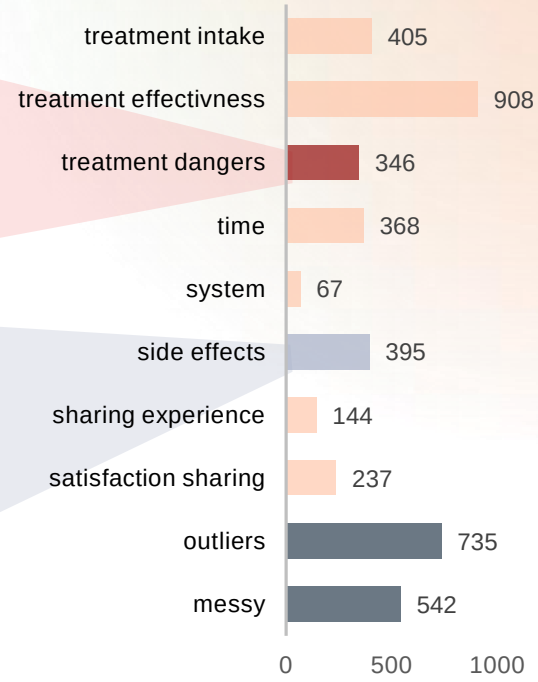


Careful with addiction

« Careful with addiction » nuage de mots des n-grams du cluster sélectionnées avec TF-IDF



Histogramme des clusters liés aux effets secondaires et dangers du traitement



Histogramme des thèmes généraux

quinten health

Conclusions

Conclusion

Objectif initial : Extraction automatique et fiable de connaissances sur les opiacés forts à partir de témoignages patients

Méthodologies testées :

- **Méthodes classiques** (LDA, NMF) → non adaptées
- Vectorisation avec **Biobert**→ ne tient pas compte du sens de la phrase
- Méthodologie finale basée sur le **dendrogramme**

Résultats :

- Méthodologie finale **reproductible** et **itérative**
- Sujets de témoignages cohérents identifiés (efficacité du traitement, effets secondaires, peur de certains dangers tels que l'addiction…)
- Poster à l'**ISPOR***

Discussions

- Evaluation des modèles difficile : labélisation manuelle nécessaire
- 12% de messy clusters : travail sur des embeddings customisés
- Confrontation des résultats à des experts pour les valider cliniquement
- Biais de la source de donnée

quinten health

*Hayat P, Clemente C, Martenot V, Rollot M, A Natural Language Processing (NLP) Approach to Automate Patients' Testimonials Analysis https://www.ispor.org/heor-resources/presentations-database/presentation/euro2022-3566/122424

Références

- L.Deplante, P.Hayat et M.Rollot “Une nouvelle approche de traitement automatique des réponses de questionnaires patients”, Afrco, juin 2022 https://fr.calameo.com/read/0067593919a3a22a3f318
- Fahad, A., Alshatri, N., Tari, Z., Alamri, A., Khalil, I., Zomaya, A. Y., ... & Bouras, A., A survey of clustering algorithms for big data: Taxonomy and empirical analysis. IEEE transactions on emerging topics in computing, 2(3), 267-279, 2014
- Hayat P, Clemente C, Martenot V, Rollot M, A Natural Language Processing (NLP) Approach to Automate Patients’ Testimonials Analysis
- J. Lee et al, BioBERT: a pre-trained biomedical language representation model for biomedical text mining
- Les Médicaments des douleurs intenses, Vidal, fev. 2022, https://www.vidal.fr/maladies/douleurs-fievres/prise-charge-douleur/douleurs-intenses.html
- L. McInnes, J. Healy, et J. Melville, « UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction », ArXiv180203426 Cs Stat, Sept 2020
- Nils Reimers, sbert.net, 2022