PCR172

ISPOR 2023, November

A Natural Language Processing (NLP) approach to automate patients' testimonials analysis

Paul Hayat[1], Coriande Clemente[1], Vincent Martenot[2] , Mélissa Rollot[1]
[1]Quinten Health – Paris – France, [2]Quinten – Paris - France

INTRODUCTION

- Patients' testimonies (e.g. posts on forums or responses to questionnaires) provide valuable insights to define and characterize patient-reported outcomes (PRO), quality of life and patients' perspective on their disease.
- Traditional NLP methods used for the automatic extraction of topics from textual data are based on the frequency of co-occurrence of words in documents and are therefore not adapted to the analysis of patient testimonies which tend to be rather short texts and where co-occurrences are rare.
- We present an innovative methodology, more appropriate to the analysis of shorter texts such as testimonies, which allows in the presented use case to identify the items raised by 4474 patients' testimonies of kaggle data from WebMD [1] on their use of strong opiates.

OBJECTIVES

Building upon an efficient NLP topic modeling method based on semantic proximity we introduced recently [2][3], the objective is to improve results coherence and interpretability using a fit-for-purpose post processing step.

METHODOLOGY

The methodology is inspired by a recently introduced topic extraction analytical pipeline based on the comparison of semantics of testimonies, which has been adapted with new steps (❖) to generalize the method, improve its performances and the results interpretation.

1- Data cleaning: Testimonies are cleaned to remove special characters and non-informative elements, and to isolate important information (see Figure 1).

2- Data pre-processing:

- Sentences transformation: Vector representations of the testimonies (embeddings) capturing the meaning of the testimonies are derived with Sentence-BERT [4], a language pre-trained model based on BERT [5].
- Concentrate the signal: To concentrate the signal, remove the noise and ease interpretation, embeddings dimensionality is reduced to two using the UMAP algorithm [6].

3- Modeling:

- Gather similar testimonies: Semantically related texts are grouped together into an optimal number of clusters (based on silhouette scores) using a hierarchical clustering (HCA) [7].

❖ Anomaly deletion: In order to obtain more robust clusters and improve their consistency, outliers of each cluster are identified with Local Outlier Factor [8] (LOF) and excluded.

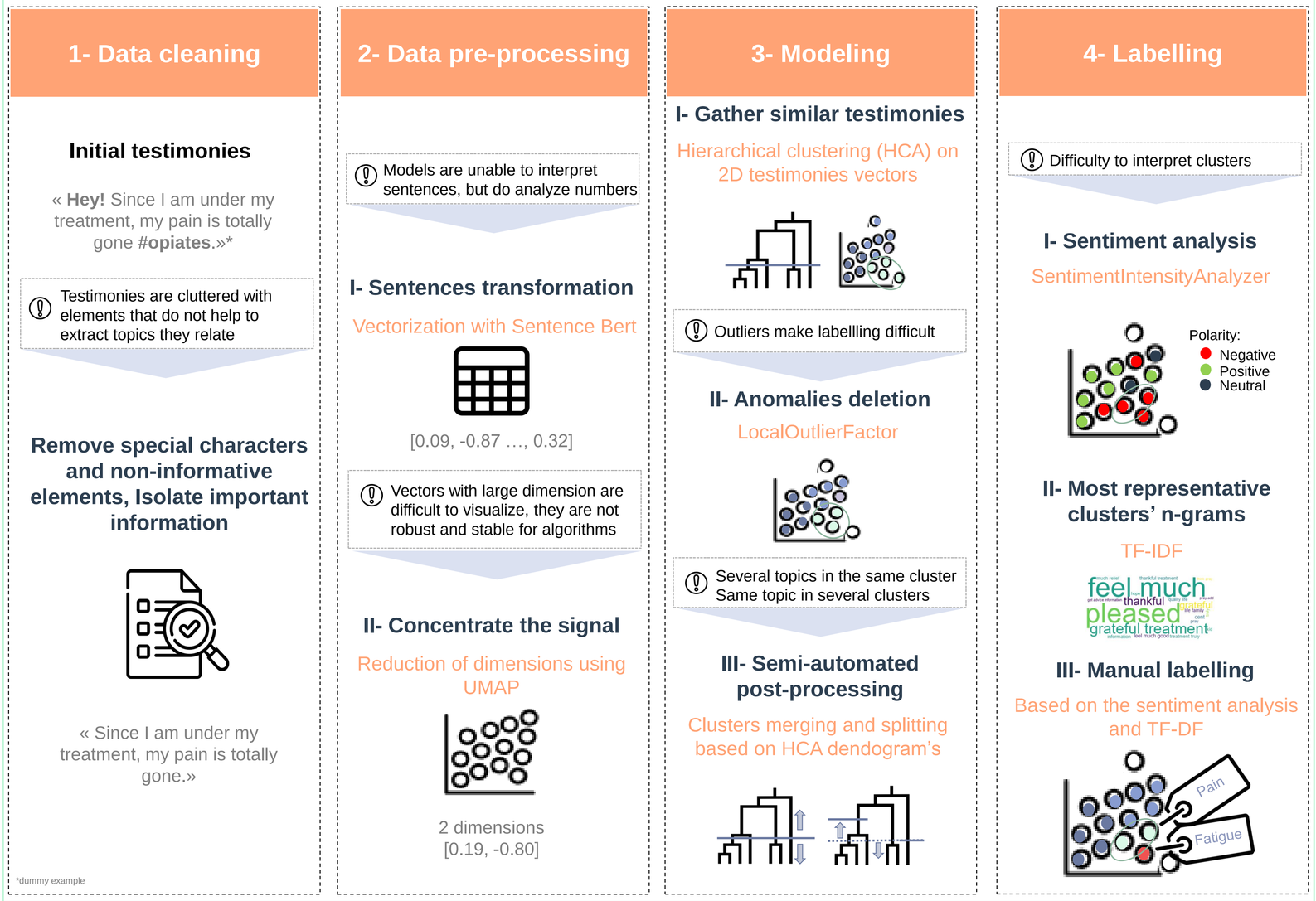
❖ Semi-automated post-processing: HCA dendrogram's facilitates post-processing interventions by automatically suggesting the clusters that can be merged together or split into two subclusters.

4- Clusters' interpretation and labelling :

❖ Sentiment analysis: To accompany the cluster's interpretation and to account for the polarity of the testimonies that constitute them, a sentiment analysis [9] is performed.

- Most representative clusters' words: Clusters understanding is assisted by the TF-IDF top scores selection for word unigrams, bigrams and trigrams inside each cluster. This provides the discriminative words and phrases for each cluster.
- Labelling: Clusters are finally labelled manually based on the most prevalent words and the sentiment of the testimonies they group.

Figure 1: Overview of the methodology

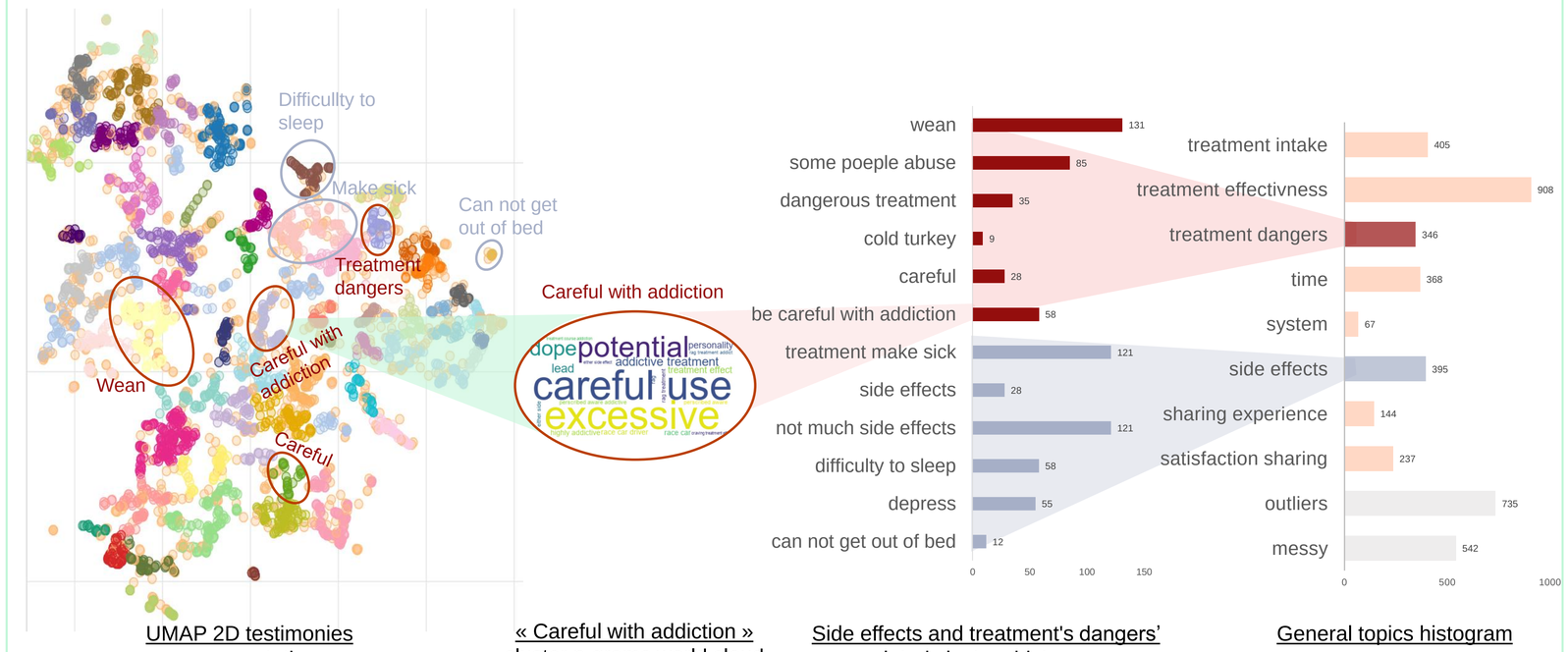


an end-to-end analytical pipeline to extract value from patient testimonials: extraction of topics discussed by patients.

RESULTS

- Tested on 4474 patients' testimonies, the method provides 60 coherent and interpretable topics clusters, which cover 9 different general themes
- Among the identified clusters, the most prevalent topics were related to treatment efficacy and side effects. Other topics also reflect the fears of some patients regarding potential addictions to these treatments. Some clusters contain testimonies that are too varied or rare to be grouped together: these clusters have been grouped into a "messy" theme.
- Compared to previous works, the improvement of the clustering post-processing step makes the analysis pipeline much faster to execute, especially on the costly part of interpreting the results, without altering the performance.

Figure 2: Results visualization



UMAP 2D testimonies representation

« Careful with addiction » cluster n-grams world cloud based on TF-IDF

Side effects and treatment's dangers' related clusters histogram

General topics histogram

UMAP provides a 2D representation of the evidence. Each point is a testimony, and each color represents the cluster to which it belongs.

The cluster's word cloud provides a visual representation of its content.

A zoom is made here on clusters about side effects and others about opiates' dangers.

In order to facilitate the reading of the results, the clusters have been grouped by general themes.

DISCUSSION

- The 15% of testimonies in a messy theme show that the methodology is still perfectible, although the results also depend on the richness of the data.
- To improve the fineness of the clusters, the method can be improved by having more adapted embeddings for each type of data (e.g. with a model trained on sentences from the same field as the testimonies)
- Medical expertise is required for the interpretation of the results, so that the groupings, distinctions and labelling of clusters are as relevant as possible.

CONCLUSION

The proposed method makes possible the extraction of coherent topics from a large volume of short texts in an automated and efficient way. Applied to patients' testimonies, such analysis provides strong insights on patients' perception about a wide range of healthcare topics (side effects, treatment, symptoms...), paving the way for better PRO definitions and patient-centric evaluation, and striving better adherence to treatments.

[1] WebMD Drug Reviews Dataset, https://www.kaggle.com/datasets/rohanharode07/webmd-drug-reviews-dataset, 2020
[2] 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.
[3] L.Deplante, P.Hayat et M.Rollot "Une nouvelle approche de traitement automatique des réponses de questionnaires patients", Afrco, 2022, June.
[4] N. Reimers et I. Gurevych, « Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks », ArXiv190810084 Cs, 2019, August.
[5] J. Devlin, M. Chang et. Al. «BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding», arXiv:1810.04805, 2018, October
[6] L. McInnes, J. Healy, et J. Melville, « UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction », ArXiv180203426 Cs Stat, 2020, September.
[7] 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
[8] scikit-learn developers (BSD License), Local Outlier Factor, 2007-2022
[9] NLTK project, SentimentIntensityAnalyzer, 2022, March

The authors declare no conflict of interest.

quinten health