NATURAL LANGUAGE PROCESSING ET SANTÉ : Use cases et potentiel

Avril 2020

par Tanguy Masgnaux, Business Developer

Contexte

Le potentiel du traitement de la donnée « patient » pour faire progresser la recherche, les soins et l'innovation en santé, est aujourd'hui unanimement reconnu et sa pleine exploitation est devenue un enjeu majeur. Nous assistons, en effet, depuis quelques années à une augmentation exponentielle de la quantité et de la diversité des données disponibles. La donnée « patient » provient de multiples sources. Les plus classiques comprennent les bases de données médico-administratives : le Sniiram et ses 8,9 milliards de feuilles de soins, les images des 80 millions d'actes d'imagerie effectués chaque année, les cohortes et registres, les dossiers médicaux ou encore les essais cliniques. Les sources plus disruptives comportent, quant à elles, les données collectées via les smartphones, les forums en ligne, ou encore les réseaux sociaux.

Parmi les outils permettant de débloquer le potentiel de cette masse de données hétérogène figure le Natural Language Processing (NLP), branche de la Data Science portant sur la manipulation et l'interprétation des données textuelles issues du langage humain. A travers cet article, nous tentons d'en illustrer l'immense potentiel dans le domaine de la santé via trois use cases mettant en jeu trois sources de données médicales très différentes : les retours d'expériences délivrés directement par le patient (dans le cadre de questionnaires de satisfaction par exemple), les réseaux sociaux et enfin les fiches médicales ainsi que les différents rapports écrits par le personnel soignant.

1 - Analyse de l'expérience patient à partir de questionnaires de satisfactions pour optimiser le parcours patient

Contexte : Comprendre et analyser l'expérience des patients depuis la prise en charge jusqu'à la guérison est indispensable dans la quête d'un parcours patient toujours plus efficace et personnalisé. Dans les questionnaires de satisfaction patient, nous retrouvons habituellement deux types de questions : fermées et ouvertes. Les questions fermées sont les plus simples : une seule modalité doit être choisie parmi une liste de réponses proposées (par exemple choisir un chiffre entre 1 et 5). Concernant les questions ouvertes, le patient peut écrire ce qu'il veut et utiliser ses propres mots pour décrire son expérience. Les réponses à la seconde catégorie de questions, au format « texte libre » nécessitent traditionnellement un passage en revue manuel. C'est un processus long, coûteux et peu efficace pour corréler l'information extraite avec d'autres indicateurs de qualité. Une équipe de chercheurs de l'Imperial College à Londres s'est lancé le défi d'utiliser le Natural Language Processing et le Machine Learning pour extraire et catégoriser l'information de manière fiable et continue à partir des réponses au format « texte libre » écrites par les patients. Le but final étant de développer des data visualisations afin d'aider le personnel médical de terrain à améliorer en continu le parcours patient.

Données utilisées : Dataset de 131,946 commentaires au format texte libre issues du questionnaire « Friends and Family Test Survey » (Janvier à Juillet 2017) dans quatre hôpitaux de l'Imperial College Healthcare NHS Trust.

Méthodologie : Les commentaires format texte libre ont été analysés en réponse à deux questions : « qu'est-ce qui a été bien fait ? » et « qu'est-ce qui pourrait être amélioré ? ». 10% des commentaires ont d'abord été catégorisés par thème et sentimentalisés (positif, négatif, neutre) à la main dans l'optique de constituer une base d'entraînement pour le modèle de NLP et de le classifier. Concernant la classification par thème, 6 algorithmes de Machine Learning supervisé ont été testés. C'est finalement le Support Vector Machine (SVM) qui a délivré les meilleures performances : 15 minutes pour catégoriser tous les commentaires versus 4 jours à la main.

Résultat : Un dashboard pour permettre au per-

sonnel médical de visualiser les outputs a été créé sous forme d'un tableau. Les commentaires y sont présentés par thème et sentiment, et ce, quasiment en temps réel.

- Source : "Listen, Learn & Improve: Using language analysis to interpret and act on written patient experience feedback for near real-time patient benefit"
- Auteurs : Erik Mayer, Stephanie Harrison White, Bob Klaber, Joshua Symons, Mustafa Khanbhai, Dave Manton, Kelsey Flott, Jamie Spofforth - Imperial College Healthcare NHS Trust & Imperial College London

2 - Détection des effets indésirables du médicament (EIM) sur les réseaux sociaux

Contexte : Twitter, facebook, forums en ligne, autant de sources qui renferment des informations clés sur le patient et ses ressentis face à la maladie et aux différents traitements. Est-il alors possible de capturer et d'analyser les réactions des patients sur internet ? Des chercheurs Iraniens nous expliquent notamment le potentiel de Twitter (150 millions d'utilisateurs actifs) comme source de données pour la détection précoce des différents effets indésirables du médicament (EIM) à l'aide de méthodes de NLP et de Deep Learning. L'équipe présente une méthode de traitement algorithmique permettant l'extraction, la classification des différents effets indésirables du médicament à partir de tweets et la prédiction du niveau de risque pour un médicament donné.

Données utilisées : Un premier dataset de 6623 commentaires Twitter, combiné avec un data set de commentaires issus du forum en ligne « Ask a Patient ».

Méthodologie : Suite à une phase de pré-processing (nettoyage) des deux datasets, l'équipe de chercheurs a testé trois algorithmes d'extraction de features différents pour reconnaître et classifier les effets indésirables dans les tweets et commentaires : un Convolutional Neural Network (CNN), un Hierarchical Attention Network (HAN) et FastText. Pour surmonter un manque de précision dans la classification, les chercheurs ont combiné le dataset Twitter avec un dataset de commentaires issus du forum en ligne « Ask a Patient ».

Résultat : Cette approche Deep Learning de NLP a permis de prédire le caractère « sans risque » d'un médicament avec une précision allant jusqu'à 93% pour la méthode HAN. Les effets indésirables du médicament sont repérés dans les tweets et commentaires puis regroupés par médicament, catégorie et thème. Cet exemple illustre le potentiel des données « patient » issues des réseaux sociaux et ouvre la porte à de nombreuses applications.

- Source : "Adverse Drug Reaction Detection In Social Media By Deep Learning Methods"
- Auteurs : Zahra Rezaei, Hossein Ebrahimpour-Komleh, Behnaz Eslami, Ramyar Chavoshinejad, Mehdi Totonchi

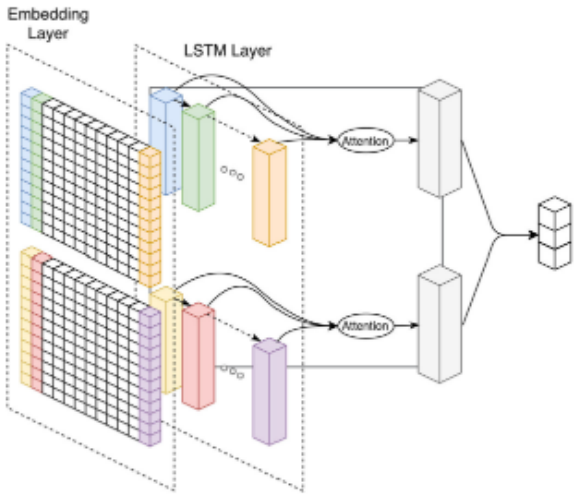
3 - Identification des patients atteints de sténose carotidienne par l'analyse des rapports d'échographie écrits par les médecins

Contexte : En France, les médecins et membres du personnel médical remplissent tous les jours des centaines de dossiers papier, dossiers informatisés, comptes rendus d'hospitalisation, mots d'évolution et autres fiches médicales. Mais en exploite-t-on suffisamment le potentiel ? Comment croiser toutes ces informations de manière optimisée ? En effet, ces fiches médicales bénéficient généralement d'une structure assez générique (même si les réponses sont au format « texte libre »), ce qui rend réalisable et prometteur un ciblage automatisé à grande échelle des patients atteints de telle ou telle pathologie grâce au NLP. Une équipe de chercheurs de l'université de Yale aux Etats-Unis s'est lancée le défi de développer un modèle de NLP pour identifier rétrospectivement les patients ayant des antécédents de sténose carotidienne à partir des rapports d'échographie écrits par les médecins.

Données utilisées : Rapports d'échographie de la Yale School of Medecine de janvier 2016 à Janvier 2017 : 1220 rapports pour la phase d'apprentissage et 307 pour la phase de test (1527 au total).

Méthodologie : Les chercheurs ont d'abord appliqué une approche Machine Learning classique avec la régression logistique pour classifier les paragraphes des rapports. En amont de cette dernière, une phase de features engineering s'est avérée nécessaire avec utilisation des méthodes bag-of-n-grams et TF-IDF. La seconde approche utilisée est celle du Deep Learning, dans l'optique d'augmenter la précision de la classification. Deux algorithmes sont alors testés par l'équipe de chercheurs, un réseau de neurones convolutif (CNN) puis un réseau de neurones récurrent (RNN).

Illustration du RNN- attention model



Source : Yale School of Medecine

L'information est encodée par une chaîne de LSTM (long-short-term memory cells). L'utilisation du mécanisme d'attention permet de concentrer le réseau de neurone sur les mots et phrases du texte qui comportent le plus d'informations.

Résultat : Tous les modèles testés pour prédire la sténose carotidienne chez les patients à partir de leur rapports d'échographie ont délivré une accuracy supérieure à 93%. En particulier, le RNN délivre le meilleur résultat avec une accuracy de 95,4%. Cela démontre le potentiel du NLP sur ce 3ème type de source de données médicales que sont les rapports médicaux. A grande échelle, l'analyse de ces documents pourrait s'avérer être un allié puissant dans le cadre du suivi long-terme des patients et des études cliniques.

- Source : "Identification of patients with carotid stenosis using natural language processing"
- Auteurs : Wu X, Zhao Y, Radev D, Malhotra A.

A propos de Coperneec

"From revolution to performance"

Coperneec est un cabinet de conseil cross-sectoriel spécialiste de la valorisation de la Data. Nous intervenons sur l'ensemble de la chaîne des savoir-faire autour de la Data Science, la Data Analyse et du Data Management.
Nos méthodes et techniques scientifiques éprouvées permettent de résoudre des problématiques dans tous les secteurs de l'industrie.

Notre vocation : extraire la connaissance à partir des données et pérenniser les avancées technologiques qui en découlent. La R&D est au cœur de notre ADN et les expertises de nos consultants (data scientists, data analysts, data engineers) sont en permanence challengées afin d'accompagner au plus près les révolutions technologiques et scientifiques.

Contactez-nous

Aymeric Lisbonne
Partner
alisbonne@coperneec.com
06 88 69 67 75

coperneec

est une marque de

canopee group