Visualisation pour l'interprétation et l'explicabilité des prédictions issues de modèles d'apprentissage profond en TAL

Alexis Delaforge

▶ **To cite this version:**

Alexis Delaforge. Visualisation pour l'interprétation et l'explicabilité des prédictions issues de modèles d'apprentissage profond en TAL. Informatique et langage [cs.CL]. Université de Montpellier, 2022. Français. NNT : 2022UMONS033 . tel-04096353

HAL Id: tel-04096353
https://theses.hal.science/tel-04096353

Submitted on 12 May 2023

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

THÈSE POUR OBTENIR LE GRADE DE DOCTEUR DE L'UNIVERSITÉ DE MONTPELLIER

En Informatique

École doctorale : Information, Structures, Systèmes

Unité de recherche Laboratoire d'Informatique, de Robotique et de Microélectronique de Montpellier (LIRMM), France

Visualisation pour l'interprétation et l'explicabilité des prédictions issues de modèles d'apprentissage profond en traitement automatique de la langue

Présentée par Alexis DELAFORGE

Le 7 Novembre 2022

Sous la direction de Sandra BRINGAY et Caroline MOLLEVI

Devant le jury composé de

Philippe LENCA, Professeur, IMT Atlantique	Rapporteur
Alexandru C. TELEA, Professeur, Utrecht University	Rapporteur
Alexis JOLY, Directeur de recherche, Université de Montpellier	Président
Romain BOURQUI, Maître de Conférences, Université de Bordeaux	Examinateur
Fleur MOUGIN, Professeur des universités, Université de Bordeaux	Examinatrice
Sandra BRINGAY, Professeur des universités, Université Paul-Valéry Montpellier 3	Directrice
Caroline MOLLEVI, Chargée de recherche HDR, Université de Montpellier	Directrice
Arnaud SALLABERRY, Maître de Conférences HDR, Université Paul-Valéry Montpellier 3	Encadrant
Maximilien SERVAJEAN, Maître de Conférences, Université Paul-Valéry Montpellier 3	Encadrant
Jérôme AZÉ, Professeur des universités, Université de Montpellier	Invité

ICM Institut régional du Cancer Montpellier | Val d'Aurelle

UNIVERSITÉ DE MONTPELLIER

La Région Occitanie Pyrénées · Méditerranée

Résumé

Les réseaux de neurones profonds ont montré, cette dernière décennie, un important accroissement des performances pour de nombreuses tâches prédictives. Néanmoins, leur grand nombre de paramètres en font des boîtes noires, que les utilisateurs ont du mal à s'approprier. Récemment, la communauté s'est intéressée à l'interprétation de leur fonctionnement. Dans cette thèse, nous nous focalisons sur les techniques de visualisation de données qui sont un levier important pour améliorer l'interprétabilité des réseaux de neurones.

Tout d'abord, nous proposons une nouvelle approche pour expliquer les prédictions des réseaux de neurones dans une tâche de classification dichotomique de textes. Elle s'adresse à des spécialistes ou utilisateurs réguliers de réseaux de neurones. Cette visualisation repose sur la construction de localités dans l'espace de représentation des textes et la visualisation de la frontière de décision avec des distances fidèles à celles présentes dans l'espace de représentation au sein d'une localité.

Nous proposons également une seconde approche pour présenter les prédictions des réseaux de neurones, pour une tâche de classification multiclasse de textes. Elle s'adresse à des non spécialistes des réseaux de neurones. Cette étude se focalise sur un cas d'étude, celui de l'exploration des mentions d'interventions non-médicamenteuses liées au cancer dans les médias sociaux.

De manière générale, ces travaux ont ouvert des perspectives prometteuses sur la production d'explications exploitant la frontière de décision et sur l'exploration des données issues des médias sociaux.

Mots-clés – Visualisation de données, apprentissage profond, traitement automatique de la langue, classification de textes, interprétabilité, explicabilité, médias sociaux, cancer, interventions non-médicamenteuses

Abstract

Deep neural networks have shown, over the last decade, a significant increase in performance for many predictive tasks. Nevertheless, their complexity (large number of trainable parameters) make them appear as black boxes to the users, with the latter having difficulties in the understanding of their predictions. Recently, the interest of the scientific community in interpreting their predictions is significantly increased. In this thesis, we focus on data visualization techniques that leverage importantly the improvement of their interpretability.

First, we propose a new visualization approach to explain the predictions of neural networks in a binary text classification setting, that is addressed to regular as well as to users who are neural networks experts. This visualization is based on the construction of localities in the text representation space. The visualization of the decision boundary is then made using distances similar to those present the representation space.

We also propose a second approach for presenting the predictions of neural networks, for a multiclass text classification task. It is addressed to users who are not neural network experts. This module focuses primarily on a case study : the investigation of mentions, in the social media, of non-pharmaceutical interventions used that are linked to cancer.

In general, this work has opened promising perspectives on the production of explanations based on the decision boundary and on the exploration of data from social media.

Keywords – Data visualization, deep learning, natural language processing, text classification, interpretability, explicability, social media, cancer, non-pharmaceutical interventions

TABLE DES MATIÈRES

GLOSSAIRE

ACP du français Analyse en Composantes Principales. En anglais Principal Component Analysis (PCA). 28, 29, 37–39, 44, 56, 74, 85

AE du français Auto-Encodeur ou de l'anglais Auto-Encoder. 22, 23, 40, 47, 59, 77, 78, 132

CNN de l'anglais Convolutional Neural Network. 15, 16, 19, 20, 23, 26, 27, 32, 34, 37, 41, 42, 49

GRU de l'anglais Gated Recurrent Unit. 18, 19, 23, 42

IA du français Intelligence Artificielle. En anglais, Artificial Intelligence (AI). 2, 24, 26

INM du français Intervention Non-Médicamenteuse. En anglais, Non-Pharmaceutical Intervention (NPI). 2, 4–7, 93, 94, 96–100, 102–106, 108, 110, 112, 116, 117, 120–123, 127, 133

LRP de l'anglais Layer-Wise Relevance Propagation. 30

LSTM de l'anglais Long-Short Term Memory. 17–20, 23, 25, 27, 33, 40–44

MCC de l'anglais Matthews Correlation Coefficient, aussi coefficient de corrélation de Matthews en français. 65, 67, 71, 74, 75, 78

MLP de l'anglais Multi-Layer Perceptron. 14, 15, 20, 23, 34, 37, 41

POS de l'anglais Part-Of-Speech tagging, aussi étiquetage morpho-syntaxique en français. 44, 45

RNN de l'anglais Recurrent Neural Network. 16, 17, 19–23, 25, 30, 31, 35–37, 42–45, 47, 55, 58, 59, 76, 78, 86

SVM de l'anglais Support Vector Machine, aussi machine à vecteurs de support. 14, 39, 49, 85

t-SNE de l'anglais t-Distributed Stochastic Neighbor Embedding. 37–39, 41, 49, 50, 56, 74, 81, 82, 85

TAL du français Traitement Automatique des Langues. En anglais, Natural Language Processing (NLP). 2, 3, 5, 11, 19, 20, 23, 24, 27, 30, 32, 34, 35, 37, 42, 45, 47, 51, 55, 56, 75, 76, 126, 127

TF-IDF de l'anglais Term Frequency-Inverse Document Frequency. 12, 13, 15, 26, 28, 31

UMAP de l'anglais Uniform Manifold Approximation and Projection. 5, 31, 37–40, 49, 50, 56, 59–62, 64, 74, 78, 83–85, 87, 88, 128

CHAPITRE 1

INTRODUCTION

Sommaire

Dans cette introduction, nous présentons le contexte de l'apprentissage automatique pour le Traitement Automatique de la Langue (TAL) et évoquons notamment les problématiques soulevées par les techniques d'apprentissage profond. Ensuite, nous définissons le domaine métier autour duquel s'articule cette thèse, à savoir, les mentions d'Interventions Non-Médicamenteuses (INM) sur les médias sociaux. Dans cette partie, nous présentons plus précisément les INM utilisées en complement de soins traditionnel pour une maladie : le cancer. La troisième partie de cette introduction résume nos trois contributions. Enfin, nous présentons le plan de la thèse, associée aux publications issues de nos travaux.

La thèse présentée dans ce manuscrit est financée par l'Institut du Cancer de Montpellier [1] (ICM) [Grant INCa_Inserm_DGOS_12553] et par la Région Occitanie [Program "Allocation Doctorale 2019"].

1.1 Apprentissage profond en traitement automatique de la langue

L'avènement de l'Intelligence Artificielle (IA), et plus spécifiquement des modèles d'apprentissage profond, a été accompagné de résultats impressionnants, notamment dans le domaine du TAL [Tor+20]. Dans cette section, nous décrivons brièvement les concepts mis en jeux et les problématiques que ceux-ci soulèvent, avant de présenter comment les techniques de visualisation de données peuvent répondre à ces problématiques.

L'IA est un terme populaire qui désigne des programmes et des algorithmes créés pour copier une forme d'intelligence réservée aux humains. Elle englobe toutes les architectures, systèmes ou algorithmes qui permettent à des machines de réaliser des tâches qui n'étaient précédemment accessibles qu'aux humains et qui nécessitent des raisonnements complexes. L'apprentissage automatique est une des méthodes de l'IA. Son objectif est d'extraire des connaissances d'un jeu de données automatiquement, de manière à accomplir une tâche. Le terme "automatique" fait donc référence au fait que l'humain n'intervient que dans la conception de l'algorithme et de la procédure d'apprentissage. Une fois cet algorithme et sa procédure d'apprentissage mise au point, l'algorithme apprend tout seul à l'aide des données. Ces données sont le plus souvent annotées, *i.e.* elles contiennent des éléments encodant le résultat de la tâche que l'algorithme apprend à réaliser. L'apprentissage profond regroupe les méthodes d'apprentissage automatique qui font appel à des réseaux de neurones. Ces réseaux sont des algorithmes très complexes avec des millions de paramètres à apprendre. Ils sont composés d'un très grand nombre d'opérations mathématiques et de fonctions non linéaires (on parle de fonctions d'activation) exécutées les unes à la suite des autres. L'apprentissage profond soulève un fort intérêt de la communauté scientifique depuis quelques années [Tor+20]. Néanmoins, derrière les performances des réseaux de neurones se cachent de nombreuses problématiques, comme l'interprétabilité [Gui+18b ; Zha+21 ; Hua+20], l'éthique [Mit19 ; Hag20] et la sécurité [He+20 ; Hua+20]. Dans

1. https://www.icm.unicancer.fr/fr/linstitut-du-cancer-de-montpellier/la-recherche/siric-montpellier-cancer

cette thèse, nous nous concentrons sur les problématiques d'interprétabilité. L'interprétabilité en apprentissage profond concerne toutes les méthodes ou les architectures qui permettent de pallier le caractère boîte noire des modèles. Cette caractéristique est héritée des nombreux paramètres qui les composent et donc de l'impossibilité pour les êtres humains de tous les analyser. L'interprétabilité permet de mieux comprendre comment fonctionne ces modèles. Concrètement, augmenter l'interprétabilité d'un modèle se traduit par la mise en place de techniques poursuivant deux objectifs distincts : (1) augmenter la transparence du modèle et (2) expliquer les prédictions. La transparence correspond au fait que les processus inhérents au modèle sont facilement compréhensibles. Un arbre de décision, par exemple, est transparent car les processus le composant sont simples et facilement compréhensibles et reproductibles par un humain. Un réseau de neurones avec de nombreuses couches est très peu transparent car les modules le composant ne sont pas compréhensibles par tous et personne ne peut reproduire exactement le processus pour une prédiction. L'explicabilité des prédictions correspond à l'identification des raisons pour lesquelles un modèle a pris une décision, à l'aide d'indicateurs résultant ou non du processus du modèle. Augmenter l'interprétabilité augmente également la confiance des utilisateurs dans ces outils.

Le traitement automatique de la langue (TAL) est une discipline qui a pour objectif de modéliser, à l'aide de techniques d'apprentissage automatique par exemple, le langage, qu'il soit écrit ou parlé. On y retrouve des applications dans tous les domaines, toutes les langues et pour tous types de tâches. On peut par exemple citer les agents conversationnels, qui, du fait du caractère ludique de la tâche et disruptif des résultats, ont été popularisés notamment par Turing [Tur50] et son test ("test de Turing"), aujourd'hui très populaire, dont il énonce le protocole en 1950. Le traitement automatique de la langue regroupe aussi, non exhaustivement, les tâches de traduction de textes [Wu+16], de classification de textes [Kow+19 ; Li+20 ; Min+21], de construction de résumés [El-+20 ; BR20] et de génération de textes [IQ20 ; Yu+22]. La complexité du domaine du TAL, associée à la complexité des méthodes d'apprentissage profond, nécessite des méthodes spécifiques pour tendre vers une plus grande interprétabilité. En effet, lorsque les réseaux de neurones traitent des images (dont la couleur d'un ou plusieurs pixels est interprétable) ou des données numériques brutes (dont le sens de chaque valeur est interprétable), ils bénéficient d'une certaine interprétabilité. À l'inverse, représenter un concept associé à un mot, à l'aide de valeurs numériques, de manière à ce que celui-ci soit compréhensible de tous, n'est pas possible. Avec cet exemple, on comprend qu'il est essentiel d'adapter les méthodes d'interprétabilité au type particulier des données textuelles, mais aussi à la tâche à accomplir. Les techniques de TAL cherchent à représenter les mots, tokens, phrases ou textes dans des espaces de représentation difficilement traduisibles par des concepts du langage naturel. Une agrégation des informations sur ces espaces de représentation, sur les procédures d'apprentissage des réseaux de neurones et sur leur architecture semble être une première piste pour une meilleure interprétabilité. Les techniques de visualisation de données offrent alors de nombreuses opportunités pour répondre à ces problèmes [Hoh+19], que ce soit pour représenter les architectures des réseaux de neurones, leurs résultats, les modules les composant ou leur procédure d'apprentissage.

1.2 Interventions non-médicamenteuses, du cancer et des médias sociaux

Dans cette section, nous nous intéressons aux INM définies de la manière suivante :

> [Une INM] est une intervention non invasive et non pharmacologique sur la santé humaine fondée sur la science. Elle vise à prévenir, soigner ou guérir un problème de santé. Elle se matérialise sous la forme d'un produit, d'une méthode, d'un programme ou d'un service dont le contenu doit être connu de l'usager. Elle est reliée à des mécanismes biologiques et/ou des processus psychologiques identifiés. Elle fait l'objet d'études d'efficacité. Elle a un impact observable sur des indicateurs de santé, de qualité de vie, comportementaux et socio-économiques. Sa mise en œuvre nécessite des compétences relationnelles, communicationnelles et éthiques.
>
> (Plateforme CEPS, 2017, plateforme-ceps.fr)

Des exemples d'INM sont : les produits psychoactifs comme le cannnabis et ses dérivés (produits), les régimes spéciaux comme les régimes cétogènes ou le jeûne (programme), les pratiques sportives ou culturelles (programme). Les INM doivent être reliées à des mécanismes biologiques et/ou des processus psychologiques identifiés mais aussi avoir un impact observable sur des indicateurs de santé, de qualité de vie, comportementaux et socio-économiques. Dès lors, il faut être capable d'étudier les nouvelles INM ou les potentielles INM mais aussi de recueillir le discours, les avis et les ressentis des utilisateurs lorsqu'ils les utilisent. Les médias sociaux sont, depuis de nombreuses années, un moyen d'expression très plébiscité. En juillet 2022, sur les dix sites les plus visités aux États-Unis figurent quatre médias sociaux [2], à savoir : facebook.com, twitter.com, reddit.com et instagram.com. Explorer les publications sur les médias sociaux mentionnant les INM est une piste prometteuse pour les étudier et obtenir des informations sur les impacts qu'elles peuvent avoir sur la santé et la qualité de vie des populations [DM04 ; FM13].

Dans cette thèse, nous nous focalisons sur le discours des patients et de leurs proches quand ils parlent d'INM sur les médias sociaux dans le cadre du cancer. Il existe de nombreux types de cancer. On estime à environ 380 000 le nombre de personnes touchées par le cancer, par an, en France (INCa [3]). Ainsi, les mentions de cette maladie sur les médias sociaux sont nombreuses. Pour récolter un volume de données plus conséquent, nous nous intéresserons aux mentions d'INM en anglais car cette langue est plus utilisée que le français.

1.3 Contributions

Nous présentons, dans cette thèse, trois contributions. La première, dans le chapitre 2, est un état de l'art des techniques de visualisation de données et d'explicabilité des prédictions appliquées à la classification de texte. La seconde contribution, dans le chapitre 3, est une nouvelle méthode pour l'explication des prédictions des réseaux de

2. https://www.similarweb.com/fr/top-websites/united-states/
3. https://www.oncorif.fr/panorama-des-cancers-en-france-edition-2021/

neurones dans une tâche de classification dichotomique de textes. Dans cette contribution, nous introduisons une méthode innovante de visualisation de la frontière de décision des réseaux de neurones. La troisième contribution, dans le chapitre 4, est une nouvelle méthode d'exploration des données issues des médias sociaux lorsque celles-ci sont classifiées selon plusieurs labels à l'aide de réseaux de neurones. Nous nous intéressons aux traces que laissent des patients atteint d'un cancer ou leurs proches quand ils mentionnent des INM dans les médias sociaux. Nous présentons plus en détails, dans la suite, ces trois contributions.

1.3.1 État de l'art

L'état de l'art, du chapitre 2, porte sur les techniques de visualisation de données appliquées à l'explicabilité des prédictions des réseaux de neurones pour le TAL. Pour cela, nous présentons, dans la section 2.2, les architectures basiques de réseaux de neurones utilisées en TAL, en soulignant l'intérêt de la communauté et la chronologie de leur apparition à l'aide de frises chronologiques. Nous proposons également un ensemble de figures originales et unifiées illustrant ces différentes architectures de manière à pouvoir les comparer. Dans la section 2.3, nous définissons les problématiques d'interprétabilité, de transparence et d'explicabilité en listant les différentes méthodes d'explication des prédictions. Ces méthodes servent à l'interprétabilité des réseaux de neurones et sont, pour la plupart, non spécifiques au domaine du TAL. Une figure présente comment se situent ces méthodes les unes par rapport aux autres avec leur chronologie d'apparition. Enfin, dans la section 2.4, nous présentons les applications des techniques de visualisation de données dans le cadre de l'apprentissage profond. Nous présentons également les méthodes de visualisation de la frontière de décision des classifieurs et les méthodes représentant les espaces de représentation en grande dimension car notre contribution suivante traite de ces aspects dans le chapitre 3. Une publication sur cet état de l'art est en cours de rédaction.

1.3.2 EBBE-Text : expliquer les réseaux de neurones

EBBE-Text, présenté dans le chapitre 3, est une nouvelle approche pour expliquer les prédictions des réseaux de neurones dans une tâche de classification dichotomique de textes. Elle s'adresse à des spécialistes ou utilisateurs réguliers de réseaux de neurones. Les principales contributions de notre approche sont doubles. Elles concernent : (1) la construction de localités dans l'espace de représentation des textes à l'aide de la méthode UMAP et (2) la visualisation de la frontière de décision avec des distances fidèles à celles présentes dans l'espace de représentation au sein d'une localité. Une localité est ici définie comme une zone de l'espace de représentation dans laquelle se trouvent des données proches les unes des autres. Ces deux principales contributions, respectivement présentées dans les sections 3.4 et 3.5.2.1, sont associées à d'autres visualisations, métriques et scores pour expliquer les prédictions issues de la tâche de classification de textes et ainsi renforcer l'interprétabilité des réseaux de neurones. Ce travail a donné lieu à trois publications [Del+21a ; Del+21b ; Del+].

1.3.3 Cancer-Annot & Cancer-Vis : explorer les données des médias sociaux

La procédure complète d'exploration des données concernant les mentions d'INM sur les médias sociaux est présentée dans le chapitre 4. Elle s'adresse à des professionels de santé non spécialistes des réseaux de neurones. Cette procédure comporte trois parties. La première concerne l'extraction et l'annotation des données issues des médias sociaux. La deuxième partie concerne l'entraînement des réseaux de neurones à l'aide de deux jeux de données, le premier annoté manuellement et le deuxième par des règles. La troisième partie concerne l'exploration des données issues des médias sociaux classifiées par les réseaux de neurones entraînés. De cette procédure ont émergé deux outils de visualisation, présentés respectivement dans les sections 4.4 et 4.6. Cette procédure d'exploration des données n'est pas spécifique au cas des INM dans le cadre du cancer et pourrait être généralisée à d'autres thématiques métier. Toutefois, ce cas d'étude a orienté certains des choix que nous avons faits pour le développement des deux outils. Finalement, nous montrons, dans ce chapitre via deux cas d'études quels ont été les apports de notre méthode pour répondre aux questions que peuvent soulever l'utilisation des INM dans le cadre du cancer.

1.4 Organisation de cette thèse et publications

Cette thèse s'organise en cinq chapitres (Fig 1.1). Cette introduction est suivie du chapitre 2 qui présente l'état de l'art de la visualisation et de l'interprétation des réseaux de neurones dans le cadre de tâches de classification de textes. Le chapitre 3 porte sur l'explicabilité des prédictions à l'aide de l'outil EBBE-Text. Le chapitre 4 présente comment explorer les données issues des médias sociaux à l'aide des outils Cancer-Annot et Cancer-Vis. Le dernier chapitre clôture cette thèse en énonçant les conclusions issues de nos travaux et en présentant les perspectives de ces derniers (chapitre 5).

Nous souhaitons dans le second chapitre (chapitre 2), mettre en avant l'effervescence des recherches autour de l'apprentissage profond [Tor+20] et des méthodes poursuivant une plus grande interprétabilité pour chaque type d'architecture. Nous présentons également dans ce chapitre comment les techniques de visualisation de données sont un levier indispensable à l'interprétabilité des réseaux.

Le chapitre 3 concerne les travaux sur l'interprétabilité des réseaux de neurones dans une tâche de classification dichotomique de textes. Nous présentons une méthode innovante d'interprétation des réseaux de neurones dans une tâche de classification qui utilise la méthode de visualisation de la frontière de décision. Un outil de visualisation est présenté, EBBE-Text. Il intègre une visualisation de la frontière de décision dans différentes parties de l'espace de représentation d'un corpus de textes sur une tâche de classification de textes. Ce chapitre inclut une évaluation par des utilisateurs débutants, une évaluation par des experts ainsi que trois cas d'étude. Nous montrons comment les méthodes de visualisation de données permettent de répondre efficacement aux besoins d'interprétabilité des réseaux de neurones.

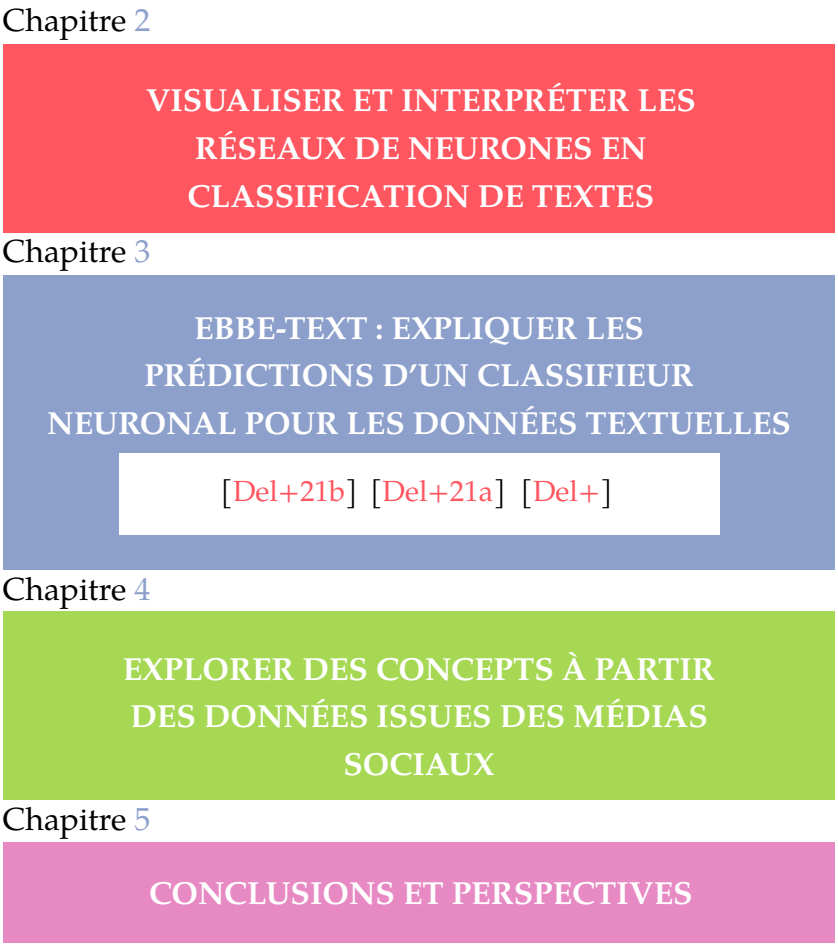


Figure 1.1 – *Plan du manuscrit de thèse. Les travaux concernant les explications des prédictions d'un classifieur neuronal pour les données textuelles ont donné lieu à trois publications. Une publication sur l'état de l'art (chapitre 2) et une autre sur le chapitre 4 sont en cours de rédaction.*

Le chapitre 4 concerne l'analyse des prédictions issues d'un réseau de neurones pour plusieurs labels. Nous nous sommes concentrés sur les données issues des médias sociaux et l'univers des mentions d'un concept (*i.e.* comment les intervenants mentionnent un concept, quelles sont les informations à tirer de ces mentions, leurs nombres, leurs provenances, leurs objectifs). Notre cas applicatif se focalise sur les mentions d'INM (concept) sur les médias sociaux, dans le cadre du cancer. Nous souhaitons, dans ce chapitre, montrer comment il est possible de répondre, à l'aide de prédictions issues d'un réseau de neurones et à l'aide d'un outil de visualisation, aux questions d'un utilisateur expert (patient ou professionnel de santé) non spécialiste en classification sur un concept donné mentionné sur les médias sociaux. Notre démarche consiste à identifier, dans des textes, les concepts évoqués (les INM dans notre cas) et les labels d'intérêt, puis à entraîner un réseau de neurones pour chaque label. Ces réseaux doivent être ensuite capables de produire des prédictions sur les labels identifiés avec de bonnes performances. Pour entraîner ces réseaux, nous avons construit un jeu de données à partir d'une démarche d'annotation des données et une autre à partir de règles produites en fonction de la provenance des données. Dans ce chapitre, nous montrons également, au travers de deux cas d'étude, comment les mé-

thodes de visualisation de données permettent de répondre efficacement aux besoins de compréhension de l'univers des mentions d'un concept dans les médias sociaux.

Le dernier chapitre conclue cette thèse en résumant les contributions accompagnées des publications produites et en présentant les perspectives de nos travaux (chapitre 5). Ces dernières concernent la visualisation d'espace de représentation dans l'objectif d'une meilleure interprétabilité et les améliorations possibles pour une meilleure exploration des données sur les médias sociaux.

CHAPITRE 2

VISUALISER ET INTERPRÉTER LES RÉSEAUX DE NEURONES EN CLASSIFICATION DE TEXTES

Sommaire

2.1 Introduction

Dans ce chapitre, nous présentons la tâche de classification de textes et les solutions basées sur les réseaux de neurones dans la section 2.2. Nous souhaitons montrer comment les textes sont abstraits de manière à être utilisés dans des classifieurs neuronaux et quels sont les architectures les plus populaires.

Ensuite, nous abordons les problématiques de transparence, d'interprétabilité et d'explicabilité des réseaux de neurones en présentant les méthodes faisant croître la confiance des utilisateurs dans ces réseaux dans la section 2.3. Nous présentons les concepts qui entrent en jeu et comment ceux-ci s'articulent les uns avec les autres, tout en se concentrant sur l'explicabilité des réseaux de neurones et de leurs prédictions.

Enfin, nous présentons, dans la section 2.4, comment les techniques en visualisation de données sont un levier indispensable à chaque étape de l'utilisation des réseaux de neurones, de leur conception, en passant par leur entraînement, jusqu'à l'analyse des prédictions, de manière à traiter les problématiques de transparence, d'interprétabilité et d'explicabilité. Plus globalement, nous présentons les méthodes utilisées pour interpréter les réseaux de neurones pour une tâche de classification, du choix du réseau de neurones à la compréhension des prédictions, tout ceci, avec l'appui des méthodes de visualisation de données.

Dans ce chapitre, il est important de noter que l'ensemble des figures sont des productions originales qui uniformisent la manière de présenter les différentes architectures des réseaux de neurones et de représentation des textes. Nous avons également produit différentes frises chronologiques qui soulignent l'effervescence autour de ces sujets.

2.2 Classification de textes

Les textes sont la principale composante des échanges dans les réseaux sociaux. Ces données textuelles peuvent être traitées automatiquement pour différentes tâches. On parle alors de traitement automatique des langues. Ces différentes tâches regroupent de manière non exhaustive : la classification [Kow+19 ; Li+20 ; Min+21], la traduction [Wu+16], la reconnaissance d'entités [YB18], la construction de résumés [El-+20 ; BR20], l'analyse de sentiments [RR15], la détection d'infox (ou "fake news" en anglais) [OQW20], la génération de textes [IQ20 ; Yu+22] ou encore la production de réponses à des questions [Die+18]. Pour accomplir ces tâches, il existe une multitude de méthodes. Dans cet état de l'art, nous nous concentrons sur les méthodes d'apprentissage automatique et plus spécifiquement d'**apprentissage automatique profond**, en présentant brièvement les autres méthodes. Enfin, la tâche sur laquelle nous nous concentrons est la tâche de **classification de textes**, qui est l'une des tâches fondatrices du traitement automatique des langues (TAL). La plupart des méthodes d'apprentissage ont besoin d'obtenir des représentations formelles des textes. Nous commençons par nous intéresser aux méthodes d'extraction de caractéristiques des textes. Ensuite, nous nous intéressons aux différents classifieurs utilisés pour la tâche de classification de textes.

2.2.1 Extraction des caractéristiques des textes

L'extraction des caractéristiques des textes consiste à construire des représentations des textes utilisés en entrée des processus d'apprentissage automatique. Certaines méthodes s'appuient sur la présence, le nombre ou la fréquence de mots quand d'autres s'appuient sur la représentation des mots pour construire celles des textes. Ces représentations sont illustrées dans la figure 2.1. Nous allons dans un premier temps aborder les méthodes qui n'utilisent pas de représentation des mots (section 2.2.1.1) pour ensuite nous intéresser à la représentation des mots à l'aide de vecteurs (section 2.2.1.2). Ces méthodes utilisent donc les mots et/ou caractères (*i.e.* lettres) et parfois leur ordre pour construire, pour chaque texte, un vecteur de représentation. Ces mots et/ou caractères sont aussi appelés tokens.

2.2.1.1 Occurrences des termes

Les méthodes pour représenter un texte, sans se servir de la représentation des mots, sont toutes dérivées du comptage des occurrences des termes, mots et caractères (ou tokens) qui le composent. Il s'agit de créer, pour un texte, une matrice éparse, aussi longue que le vocabulaire des tokens du corpus. Cette matrice contient le nombre d'occurrences des tokens contenu dans le texte et des zéros pour les mots qu'il ne contient pas. Pour affiner cette représentation en conservant les proximités entre les mots, on peut également compter les N-grammes qui s'y trouvent. Un N-gramme est une succession de N tokens. Au fur et à mesure que le nombre de tokens dans les N-grammes et le nombre de textes augmentent, ces matrices deviennent extrêmement grandes. Travailler sur la fréquence des tokens ou des N-grammes peut être aussi une solution. Par exemple, le TF-IDF [SM86] construit pour chaque texte d une matrice où, pour chaque token w, la fréquence d'apparition de ces derniers (*i.e.* $tf_{w,d}$) est multipliée par une mesure de son caractère spécifique à ce texte (*i.e.* $\text{idf}_w = \log \frac{|D|}{|d_j : t_w \in d_j|}$ où $|D|$ est le nombre de textes dans le corpus et $|d_j : t_w \in d_j|$ le nombre de textes dans le corpus contenant le token w pour lequel on mesure le TF-IDF). Finalement le TF-IDF d'un token w dans un texte d est égal à $tf_{w,d}.idf_w$. L'objectif est d'attribuer un poids plus important aux tokens les moins fréquents dans le corpus et les plus fréquents dans un texte. L'inconvénient de ces méthodes est qu'elles ignorent ou tirent peu parti de l'ordonnancement des tokens dans le texte. Or, l'ordre des tokens est une composante essentielle à la compréhension d'un texte.

2.2.1.2 Plongement lexical

L'avènement des réseaux de neurones ("neural network" (NN) en anglais) a permis l'émergence des méthodes construisant des vecteurs de représentation des mots (ou tokens). Ces méthodes sont appelées des méthodes de plongement lexical ("word embedding" en anglais) et comptent comme méthodes fondatrices Word2Vec [Mik+13a] et Glove [PSM14]. Elles partent de l'hypothèse que l'on peut construire la représentation d'un token à l'aide de son contexte. Une fois ces représentations des tokens connues, il est possible de tirer parti du sens des tokens spécifiquement mais également du sens de la phrase, puisqu'un token ou un N-gramme apparaissant plusieurs fois n'est plus simplement sommé et l'ordre des tokens est donc conservé. Ces méthodes

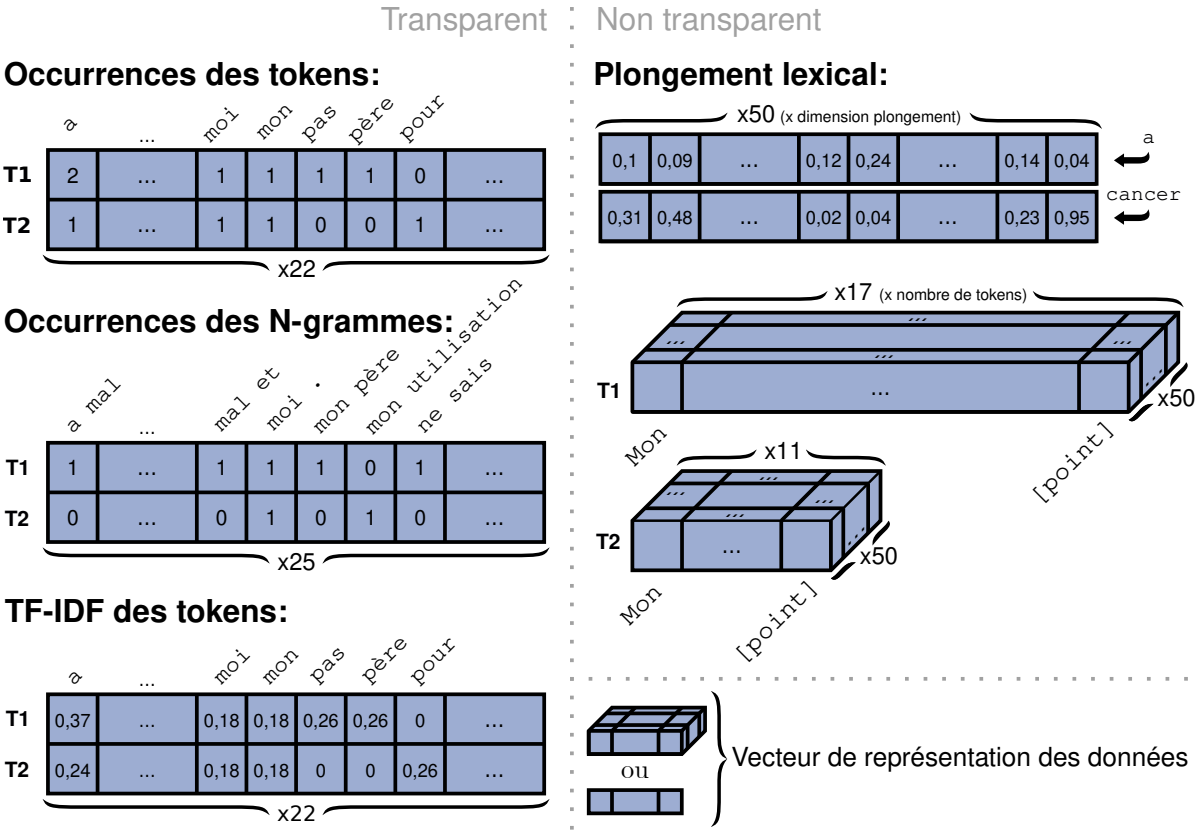


FIGURE 2.1 – *Principales techniques de représentation des textes : occurrences des tokens, occurrences des N-grammes, TF-IDF des tokens et plongement lexical. Sont présents, trois exemples de représentation transparente (à gauche) où chaque dimension du vecteur de représentation peut-être expliquée et un exemple de représentation non transparente (à droite) où les dimensions du vecteur de représentation ne sont pas interprétables. Il est important de noter que la représentation d'un mot à l'aide du plongement lexical n'est pas transparente, ce qui induit la non-transparence du vecteur de représentation du texte.*

attribuent pour chaque token un vecteur de dimension définie. Puis, un réseau de neurones est entraîné afin d'attribuer à chaque token son vecteur de représentation. Ce qui différencie les deux modèles précédemment cités est la manière dont ils s'entraînent. Word2Vec a pour objectif soit : (1) de prédire le contexte d'un token à une certaine distance de celui-ci (Skip-Gram) ou à l'inverse (2) compte tenu du contexte d'un token inconnu, de prédire quel est ce token (CBOW). Glove s'entraîne à l'aide des co-occurrences des mots dans l'intégralité du corpus et donc pas forcément dans un contexte donné. En d'autres termes, ce sont bien les co-occurrences des tokens dans un contexte proche ou non qui construisent les représentations de ceux-ci. L'évolution du plongement lexical permet aujourd'hui de ne plus attribuer exclusivement aux mots ou aux caractères des vecteurs de représentation mais aussi aux syllabes ou aux parties de mots [Wu+16] ou même de représenter les mots en plusieurs parties superposables [Boj+17].

Récemment, ELMo [Pet+18b] et BERT [Dev+18] proposent une évolution dans le domaine du plongement lexical : le plongement contextualisé. ELMo se sert d'un réseau de neurones récurrents (section 2.2.2.3) et règle le problème de polysémie de certains mots dans certaines langues. Il fait ceci en attribuant pour chaque mot un vecteur en fonction du contexte de celui-ci, ce qui était impossible avec les méthodes classiques. À l'aide d'ELMo, un même mot peut donc avoir deux vecteurs de représentation différents dans deux contextes différents. Enfin, BERT fonctionne sur la même idée mais à l'aide d'un modèle auto-attentif (section 2.2.2.4). De plus, BERT réalise le plongement lexical sur des parties de mots, ce qui est plus efficace sur les formes verbales, la conjugaison des verbes ou les formes plurielles de noms par exemple.

2.2.2 Classification de textes à l'aide de réseaux de neurones

La tâche de classification de textes consiste à produire, pour chaque texte, une prédiction de classe, parfois associée à une probabilité et donc à une "certitude" de la prédiction. Les tâches de classification où il n'y a que deux classes sont dites binaires ou dichotomiques. Les tâches de classification où il y a plusieurs classes sont dites multiclasses, c'est-à-dire qu'une donnée peut être labellisée par zéro, une ou plusieurs classes différentes. Dans les travaux que nous avons conduits, nous nous sommes d'abord intéressés aux tâches de classification dichotomique.

Dans cet état de l'art, nous nous concentrons sur la tâche de classification de textes à l'aide de réseaux de neurones. Nous souhaitons néanmoins indiquer qu'il existe de nombreuses méthodes d'apprentissage automatique qui ne sont pas liées aux réseaux de neurones. Elles incluent notamment les classifieurs naïfs bayésiens, les machines à vecteurs de support (SVM), les régressions logistiques ou même plus simplement les classifieurs linéaires. Tous ces classifieurs doivent prendre en entrée un vecteur de dimension définie et se servent donc le plus souvent des représentations sans plongement lexical. Ces méthodes ne tirent donc pas parti de l'ordre des mots.

Il existe une grande variété de réseaux de neurones qui sont ou ont été appliqués aux tâches de classification [Min+21]. Dans les sous-sections suivantes, nous présentons les architectures de base appliquées à la tâche de classification de textes.

2.2.2.1 Perceptron multicouche

Le premier réseau de neurones à voir le jour est le **perceptron multicouche** (MLP) [RHW85]. C'est un réseau à propagation directe (qui ne contient pas de boucle). Au sein de chacune de ses couches, le réseau passe d'un espace d'entrée à un espace de sortie, à l'aide d'une transformation linéaire entre la matrice des données d'entrée et la matrice des poids de la couche (figure 2.2). Le réseau applique ensuite une fonction d'activation sur le résultat de cette transformation. La dimension de cet espace de sortie peut être différente de celle de l'espace d'entrée. Dans le cas de la classification de textes, la sortie finale du réseau est un nombre réel (dans le cas d'une classification dichotomique) ou une matrice de nombres réels. Cette sortie peut ensuite être modifié(e) de manière à obtenir des probabilités pour chaque classe, à l'aide de la fonction sigmoïde ou de la fonction exponentielle normalisée ("softmax" en anglais) par exemple. Dans les figures proposées au sein de cet état de l'art, nous passerons cette étape.

Pour une tâche de classification de textes, le MLP peut prendre, en entrée, la matrice des TF-IDF car l'ordre des mots n'est pas pris en compte par ce type de réseau (figure 2.2). Iyyer et al. [Iyy+15] ont choisi de prendre en entrée une moyenne des vecteurs de représentation des mots pour une tâche de classification de textes. Le et Mikolov [LM14] ont proposé Doc2Vec qui produit un plongement des textes plutôt que des tokens (plongement lexical), à l'image de ce qu'avaient proposé Mikolov et al. [Mik+13a] dans leur précédent article présentant Word2Vec. Doc2Vec introduit un token représentant le paragraphe et étant le token dont il faut trouver le vecteur représentatif. C'est ensuite cette représentation vectorielle du paragraphe qui peut-être utilisée dans la tâche de classification de textes faite par un MLP.

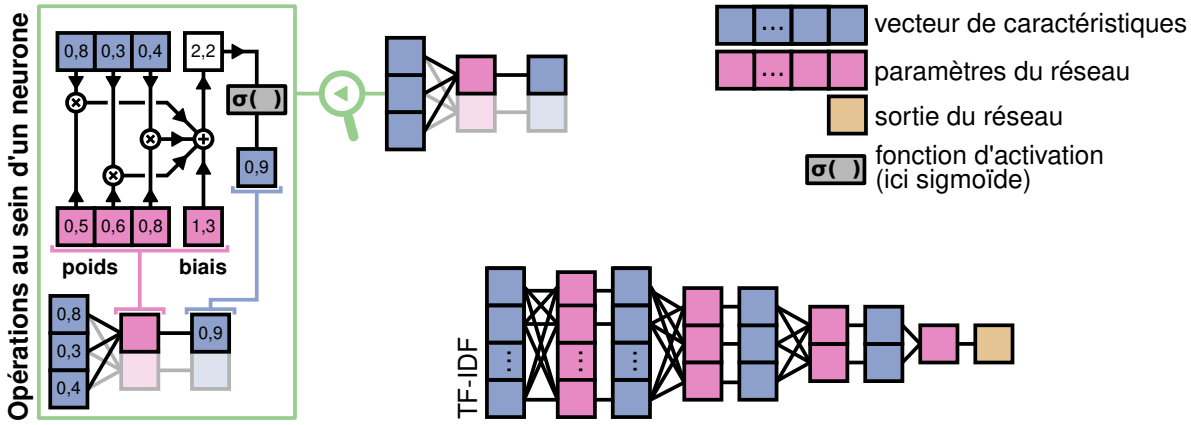


Figure 2.2 – *MLP pour une tâche de classification dichotomique de textes. Les opérations au sein d'un neurone classique sont présentées dans l'encadré vert. En bas de la figure, un MLP à quatre couches prenant en entrée le TF-IDF d'un texte.*

2.2.2.2 Réseaux de neurones convolutionnels

Parmi les premiers réseaux pour le traitement des données textuelles, ont émergé les réseaux de neurones convolutionnels [KGB14 ; Ouy+15] (CNN, voir figure 2.3) qui étaient déjà utilisés dans le traitement des images [Lec+98]. Ceux-ci parcourent les textes par fenêtre de mots.

Les CNN sont des réseaux à propagation directe. Les CNN classiques fonctionnent par système de filtres de convolution. Un filtre est une matrice de paramètres qui va appliquer certaines opérations à une donnée d'entrée et produire en sortie le résultat des opérations effectuées. Ces filtres vont successivement parcourir différentes fenêtres de représentation des tokens. À chaque passage, une opération est faite sur cette fenêtre et un nouveau vecteur est issu de cette opération. Par exemple, l'opération présentée dans la figure 2.3 dans l'encadré vert de droite est la somme des éléments de la matrice issue du produit matriciel d'Hadamard entre la fenêtre d'entrée et les poids du filtre. La fonction Max-Pooling présentée dans la partie gauche de la figure 2.3 prend en entrée une matrice et conserve certaines des plus grandes valeurs de celle-ci en fonction de leur position. Cette opération a pour objectif de sous-échantillonner la matrice d'entrée en réduisant sa dimension. En général, les CNN possèdent de nombreuses

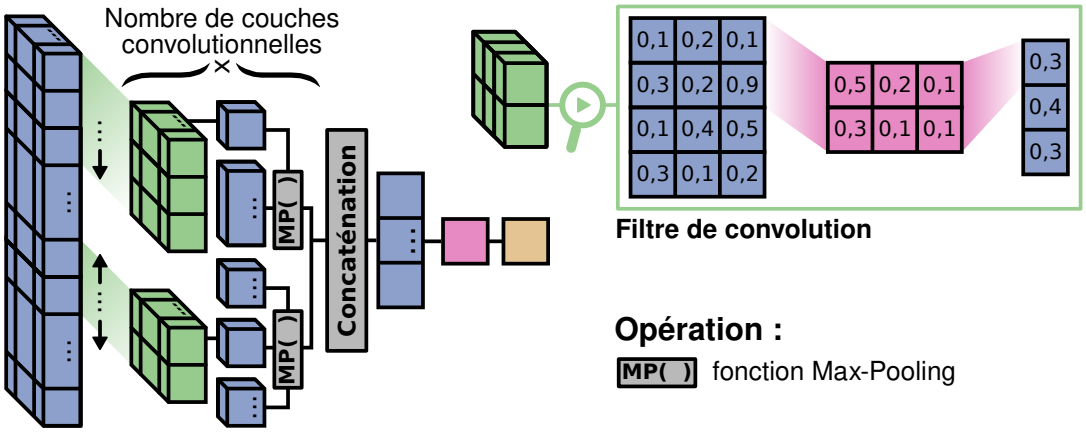


FIGURE 2.3 – *CNN pour une tâche de classification de textes.*

couches empilées les unes sur les autres, accompagnées de fonctions Max-Pooling (ou similaires). Les CNN possèdent également des filtres différents et parfois même de tailles différentes, comme dans la figure 2.3 où l'on peut voir deux filtres de tailles différentes dans l'architecture présentée. Les filtres peuvent aussi prendre en entrée des fenêtres sur-remplies (complétées sur une partie d'entre elles par des valeurs par défaut, nulles ou aléatoires), faire des écarts dans les fenêtres ou avoir des fenêtres à trous. Il existe ainsi de très nombreux paramètres pour un filtre convolutionnel [KGB14]. Les CNN agrègent ensuite les données à l'aide de concaténations pour pouvoir classifier les textes à l'aide de transformations linéaires. Les réseaux convolutionnels peuvent avoir une seule couche de convolution. En effet, KIM [Kim14] propose un réseau avec une seule couche de plusieurs filtres de convolution avec un Max-Pooling des vecteurs produits en fonction du temps avant une concaténation des données pour la prédiction, comme présenté dans la figure 2.3. KALCHBRENNER, GREFENSTETTE et BLUNSOM [KGB14], précurseurs dans l'utilisation des CNN pour la classification de textes, proposent un CNN à deux couches de convolutions alors que CONNEAU et al. [Con+17] poussent le nombre de couches jusqu'à vingt-neuf. Les tokens peuvent aussi être des caractères. Les travaux de ZHANG, ZHAO et LECUN [ZZL15], KIM et al. [Kim+16] ou CONNEAU et al. [Con+17] appliquent ce paradigme à la classification de textes.

2.2.2.3 Réseaux de neurones récurrents

Les **réseaux de neurones récurrents** (RNN, voir figure 2.4) [RHW86] parcourent les textes mot après mot, parfois dans les deux directions [SP97] et gardent en mémoire une plus ou moins grande partie des données déjà parcourues [HS97 ; Cho+14].

Les **RNN classiques** prennent simplement en entrée, au temps t, la représentation du token actuel x_t et le précédent état de la représentation du texte appelé état caché h_{t-1}. Ces entrées sont respectivement multipliées par les matrices W_{ih} et W_{hh} (poids du réseau).

Opération au sein d'une cellule RNN classique :

$$h_t = \tanh(W_{ih} \times x_t + b_{ih} + W_{hh} \times h_{t-1} + b_{hh}) \tag{2.1}$$

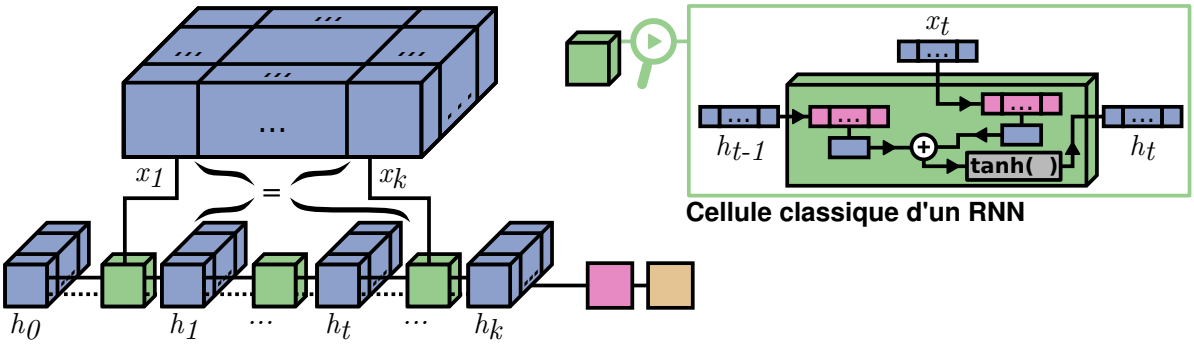


FIGURE 2.4 – *RNN pour une tâche de classification de textes. Les différents vecteurs de représentation h_t du texte aux temps t sont construits à l'aide de l'utilisation dans la cellule du RNN des représentations x_t des tokens. Les opérations au sein d'une cellule RNN sont présentées au sein du cadre **vert** à droite.*

Ici, h_t est l'état caché au temps t, x_t est l'entrée de la cellule au temps t, et h_{t-1} est l'état caché issu de la précédente cellule au temps $t-1$ (ou à l'initialisation de l'état caché au temps 0).

Les **réseaux de neurones récurrents à mémoire court-terme et long-terme** (LSTM) classiques [HS97] comportent une porte d'entrée (voir équation 2.2), une porte de sortie (voir équation 2.5), une porte d'oubli (voir équation 2.3) et une porte de cellule (voir équation 2.4). Cette architecture a pour objectif de mémoriser des valeurs sur différents intervalles de temps et les quatre portes régulent le flux d'informations entrant et sortant de la cellule. Les informations mémorisées sont passées de cellule en cellule à l'aide de l'état de la cellule (voir équation 2.6) qui vient compléter l'effet de l'état caché (voir équation 2.7). Les LSTM ont gagné en intérêt du fait qu'ils règlent le problème de rétro-propagation des gradients qui, à long-terme, peuvent "disparaître" et donc tendre vers zéro ou "exploser" et donc tendre vers l'infini [Hoc98].

Opérations au sein d'une cellule LSTM (figure 2.5) :

$$i_t = \sigma(W_{ii} \times x_t + b_{ii} + W_{hi} \times h_{t-1} + b_{hi}) \tag{2.2}$$

$$f_t = \sigma(W_{if} \times x_t + b_{if} + W_{hf} \times h_{t-1} + b_{hf}) \tag{2.3}$$

$$g_t = tanh(W_{ig} \times x_t + b_{ig} + W_{hg} \times h_{t-1} + b_{hg}) \tag{2.4}$$

$$o_t = \sigma(W_{io} \times x_t + b_{io} + W_{ho} \times h_{t-1} + b_{ho}) \tag{2.5}$$

$$c_t = f_t \odot c_{t-1} + i_t \odot g_t \tag{2.6}$$

$$h_t = o_t \odot \tanh(c_t) \tag{2.7}$$

Ici, h_t est l'état caché au temps t, c_t est l'état caché de la cellule au temps t, x_t est l'entrée de la cellule au temps t, h_{t-1} est l'état caché en sortie de la précédente cellule au

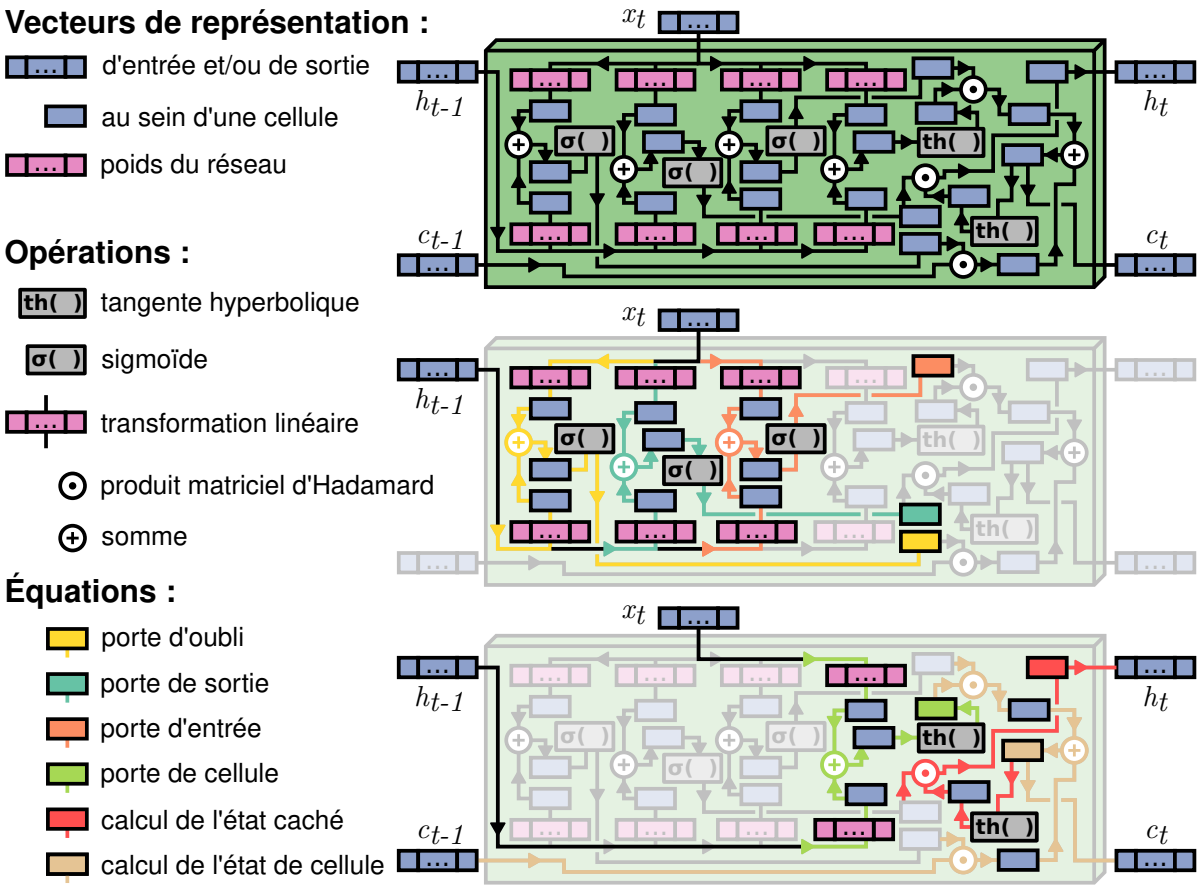


FIGURE 2.5 – *Cellule d'un réseau de neurones récurrent de type LSTM. Les flèches colorées montrent les opérations nécessaires pour produire la sortie d'une porte ou pour produire un état caché ou un état de la cellule.*

temps $t-1$ (ou est l'initialisation de l'état de la cellule au temps 0), i_t, f_t, g_t, o_t sont les sorties de la porte d'entrée, la porte d'oubli, la porte de la cellule et la porte de sortie, respectivement, σ est la fonction sigmoïde, $\odot$ est le produit matriciel de Hadamard et les matrices W et b sont les poids du réseau et les biais du réseau pour chacune des portes.

Les **GRU classiques** [Cho+14; Chu+14] ont vu le jour après les LSTM et comme ceux-ci, ils contiennent différentes portes mais ne font pas passer l'état de la cellule à la cellule suivante mais seulement l'état caché (voir équation 2.11). Ils possèdent donc une porte de réinitialisation (voir équation 2.8), une porte de mise à jour (voir équation 2.9) et une porte de nouveauté (voir équation 2.10). Leur avantage est d'avoir moins de paramètres. Ainsi, ils apprennent plus rapidement et sont plus transparents (section 2.3.1).

Opérations au sein d'une cellule GRU (figure 2.6) :

$$r_t = \sigma(W_{ir} \times x_t + b_{ir} + W_{hr} \times h_{t-1} + b_{hr}) \tag{2.8}$$

$$z_t = \sigma(W_{iz} \times x_t + b_{iz} + Whz \times h_{t-1} + bhz) \tag{2.9}$$

$$n_t = \tanh(W_{in} \times x_t + b_{in} + rt \odot (W_{hn} \times h_{t-1} + b_{hn})) \tag{2.10}$$

$$h_t = (1 - z_t) \odot nt + z_t \odot h_{t-1} \tag{2.11}$$

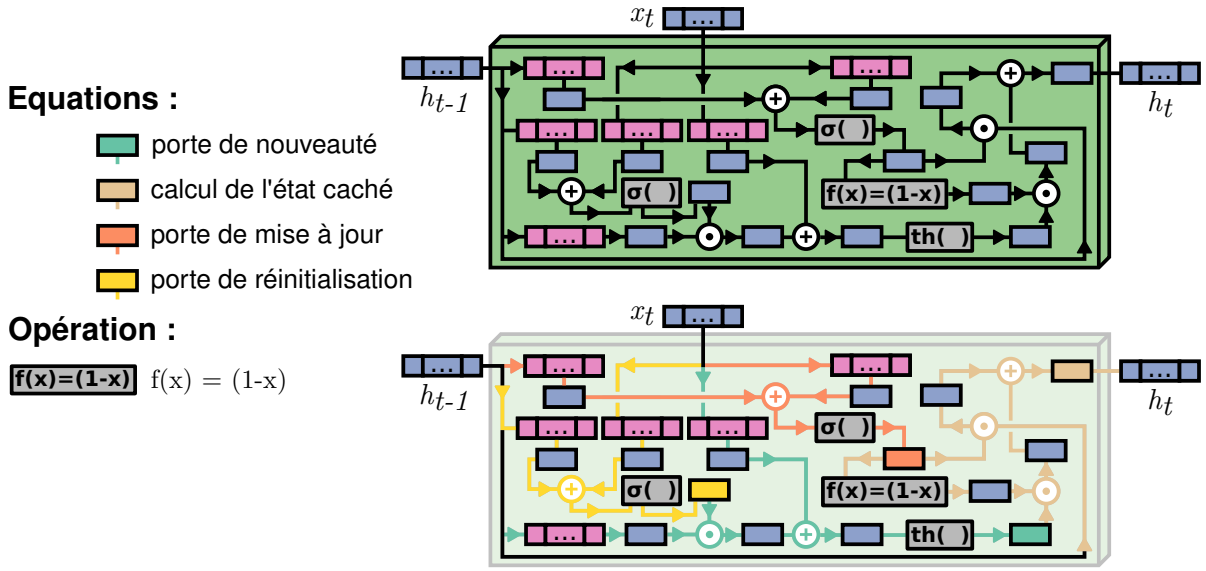


Figure 2.6 – *Cellule d'un réseau de neurones récurrent de type GRU. Les flèches colorées montrent les opérations nécessaires pour produire la sortie d'une porte ou pour produire un état caché.*

Ici, h_t est l'état de la cellule au temps t, x_t est l'entrée de la cellule au temps t, h_{t-1} est l'état de la cellule en sortie de la précédente cellule au temps $t-1$ (où h_0 est l'initialisation de l'état de la cellule au temps 0), et r_t, z_t, n_t sont respectivement la porte de réinitialisation, la porte de mise à jour et la porte de nouveauté, σ est la fonction sigmoïde, $\odot$ est le produit matriciel de Hadamard et les matrices W et b sont les poids du réseau et les biais du réseau pour chacune des portes.

Les principales variantes dans cette famille des RNN sont les RNN classiques, les LSTM et encore les GRU. Les RNN prennent en compte l'ordre des mots, ce que ne faisaient pas la plupart des méthodes qui ne sont pas basées sur les réseaux de neurones. De plus, la description très fine des mots proposée par les techniques de plongement lexical des mots (ou vectorisation des mots) est ainsi pleinement exploitée. Certaines variantes des RNN existent. Schuster et Paliwal [SP97] proposent de parcourir les tokens dans les deux sens pour construire ainsi deux représentations du texte. Cette technique peut être utilisée avec tous les types de RNN. Cheng, Dong et Lapata [CDL16] proposent de remplacer l'état de la cellule par un réseau de la cellule, ce qui consiste à conserver un nouvel état de la cellule à chaque étape. De cette manière, cette conservation des états peut ensuite servir dans un mécanisme d'**attention** (section 2.2.2.4) pour observer les liens entre les différents états et donc les liens entre différentes temporalités. On parle alors de réseau à mémoire [WCB15] et dans ce cas plus particulier de réseau LSTM à mémoire.

Les CNN et les RNN ont des performances comparables, avec un avantage à donner aux RNN [Yin+17]. Quelques travaux intéressants en TAL utilisant des RNN : ELMo

[Pet+18b] construit des représentations de mots et donc un plongement lexical de meilleure qualité. ULMFit [HR18] est un réseau de plusieurs variantes de LSTM efficace dans les tâches de classification.

2.2.2.4 Réseaux de neurones auto-attentifs

Les transformeurs, ou réseaux de neurones auto-attentifs, ou modèles auto-attentifs [Vas+17] comme BERT [Dev+18] ou GPT [RN18], sont les derniers nés dans la famille du TAL. Plus rapides que leurs prédécesseurs, les RNN, ils ont permis d'avoir accès à des modèles pré-entraînés sur un grand volume de données. De plus, ils utilisent des vecteurs d'attention [Vas+17] qui permettent de savoir sur quelle partie des données s'est appuyée la prédiction (en fonction des couches ou des têtes, voir figure 2.7). Les mécanismes d'attention, au sein d'une tête d'attention, prennent en entrée trois vecteurs q, k, v, respectivement une requête, une clef et une valeur. Un masque aléatoire des données peut-être exécuté, notamment pendant l'entraînement. Cela implique qu'une partie des données de la clef ne sera pas visible et donc égale à zéro. Chaque tête d'attention consiste donc en l'exécution de ces mécaniques avec des entrées q, k, v différentes. En pratique, pour la classification de textes, ces vecteurs sont tous issus du même vecteur de représentation sont modifiés par des transformations linéaires et des masques. Ces transformations (en général de plus petite dimension) et ces masques construisent les entrées des différentes têtes d'attention. Dans le cas de la traduction des textes, leurs vecteurs sources sont par exemple différents. Les têtes d'attention vont produire les matrices d'attention de dimension $\mathbb{R}^{(k+1)\times(k+1)}$ où k est la longueur de la séquence (le nombre de tokens du texte). Pour un texte de k tokens, plongés dans un espace de dimension d, on a $q, k, v \in \mathbb{R}^{(k+1)\times d}$. Dans notre exemple, nous avons choisi $k+1$ au lieu de k pour présenter ensuite l'exemple de BERT qui ajoute un token de classification au début de la phrase dans le cas d'un tâche de classification de textes.

Le calcul de l'attention est alors donné par :

$$\text{Attention}(q, k) = \sigma(\frac{q \times k^T}{\sqrt{d}}) \tag{2.12}$$

où σ est la fonction exponentielle normalisée ("softmax" en anglais).

Le calcul du contexte issu de la tête d'attention est alors donné par :

$$\text{Contexte}(q, k, v) = \text{Attention}(q, k) \times v \tag{2.13}$$

Les réseaux de neurones auto-attentifs, à la suite du passage de l'information dans les cellules d'attention, concatènent les différents contextes issus de celles-ci pour ensuite accomplir une nouvelle transformation linéaire. C'est ensuite sur la représentation obtenue du texte que les modifications, apportées par le processus d'attention, sont ajoutées au vecteur d'entrée qui représente le texte à la sortie. Les cellules d'attention n'ont donc pas pour objectif d'encoder la donnée dans un tout nouvel espace de représentation mais bien d'ajuster celui d'entrée en fonction de la tâche à accomplir. Ensuite, un réseau de neurones à propagation directe (sans récurrence) effectue aussi une modification des données. Ce réseau peut être un simple MLP ou un CNN [MFC19]. Cette

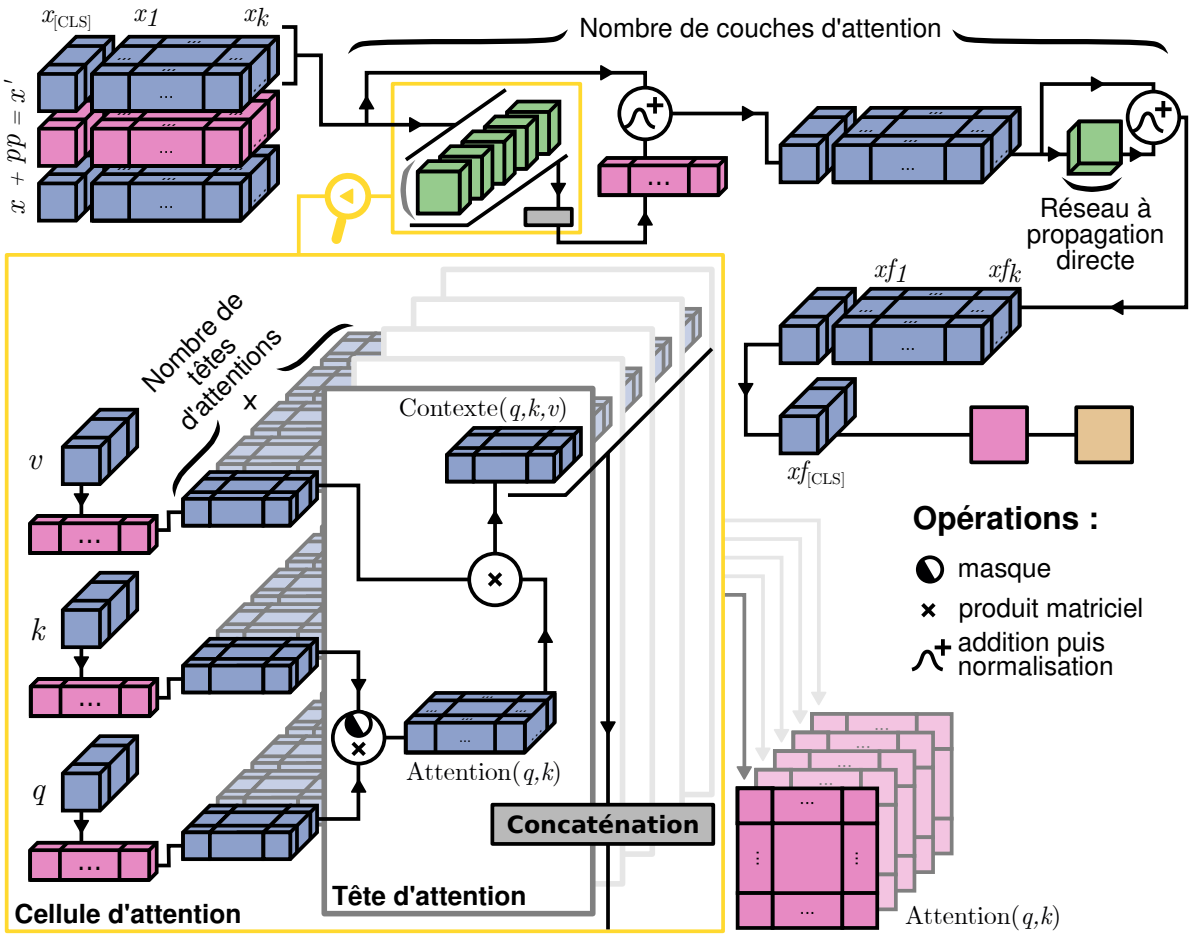


FIGURE 2.7 – *Adaptation d'un modèle auto-attentif (BERT) à une tâche de classification de textes.*

transformation n'a pour objectif que de modifier légèrement les données passées en entrée. Ainsi, dans le cadre de la classification de textes, pour une couche d'attention (CA), un triplet de vecteurs d'entrée q, k, v, un réseau de neurones à propagation directe RPD, et une procédure d'attention Contexte qui donne en sortie le vecteur concaténé des contextes, une fonction de normalisation Norm, la sortie $\text{CA}(q, k, v)$ est calculée de la manière suivante :

$$\text{CA}(q, k, v) = \text{Norm}(\text{RPD}(\text{Norm}(\text{Contexte}(q, k, v) + q)) + q) \tag{2.14}$$

Une illustration d'une classification automatique de textes utilisant une adaptation d'un réseau de neurones auto-attentif est proposée dans la figure 2.7. On peut voir que BERT ajoute un token supplémentaire au début de la représentation du texte $x \in \mathbb{R}^{d*(k+1)}$ où k est le nombre de tokens dans le texte et d la dimension du plongement lexical des tokens : $x_{[\text{CLS}]}$. Pour la suite, cette représentation est additionnée avec un encodage de position pp pour créer la représentation x' fournie au réseau. L'objectif est d'encoder la position des tokens. En effet, BERT traite tous les tokens en même temps et perd donc cette capacité des RNN à traiter les tokens dans l'ordre. L'encodage de position cherche donc à combler ce manque. Au fur et à mesure des couches successives d'attention, cette représentation va être modifiée. À la sortie de la dernière couche, on

a donc xf la représentation finale du texte et $xf_{[CLS]}$ le vecteur utilisé pour la tâche de classification. Une couche de transformation linéaire va ensuite produire la sortie du réseau. D'autres modèles auto-attentifs ne traitent pas tous les tokens dans le même ordre mais de manière séquentielle et se servent de la sortie générée pour prédire le mot suivant [RN18]. Ce type de modèle, comme GPT, se sert donc du dernier token produit pour ensuite effectuer la tâche de classification. BERT est un transformeur se servant de blocs d'encodage et traitant toute la donnée d'entrée simultanément. GPT est un transformeur se servant de blocs de décodage et traitant la donnée de manière séquentielle, en se servant de la précédente prédiction en tant que nouvelle clef dans les cellules d'attention.

L'attention est un mécanisme qui a également été utilisé dans les RNN [Lin+17b]. Il devient possible d'observer à quel point un mot a eu de l'influence sur la prédiction en récupérant les états cachés des cellules h avant de les utiliser pour prédire la sortie du réseau en conservant pour chaque token d'entrée au temps t une représentation vectorielle propre à lui à l'étape (v_t) jusqu'à avoir son score d'attention a_t. Enfin, cette matrice des scores d'attention a est multipliée par la matrice des états cachés h pour obtenir le vecteur de représentation du texte rep (figure 2.8).

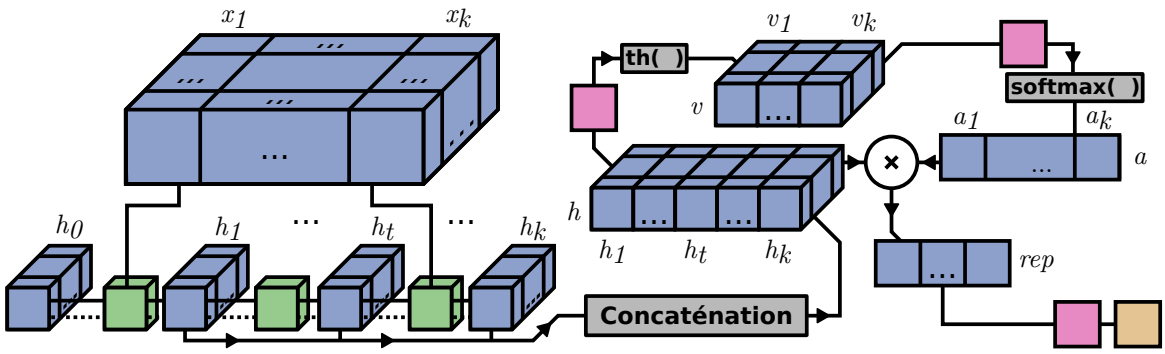


FIGURE 2.8 – *Adaptation d'un RNN pour utiliser les mécanismes d'attention.*

Un autre mécanisme intéressant en traitement automatique de langue est le mécanisme d'auto-encodeur (AE) [RHW85] qui a vu le jour dès l'apparition des premiers réseaux de neurones profonds. Un réseau AE a pour objectif d'encoder la donnée dans un espace de plus petite dimension avant de décoder celle-ci pour la reconstruire à l'identique. Cet espace de plus petite dimension peut ensuite servir dans une tâche de classification de textes. Un AE possède deux parties, une partie **encodeur** qui transforme la donnée d'entrée et une partie **décodeur** dont l'objectif est de reconstruire la donnée. Le premier réseau de neurones auto-attentif introduit par VASWANI et al. [Vas+17] est d'ailleurs un AE puisqu'il contient une partie encodeur et une partie décodeur. BERT ne contient que la partie encodeur et GPT ne contient que la partie décodeur. D'autres exemples d'AE, utilisant cette fois des RNN, sont présents dans les travaux de BOWMAN et al. [Bow+15].

Il existe donc un grand nombre de réseaux de neurones appliqués ou applicables aux tâches de classification comme le présentent MINAEE et al. [Min+21] dans leur état de l'art des techniques d'apprentissage profond appliquées à la tâche de classification de textes. Certains d'entre eux sont des réseaux hybrides se servant de plusieurs

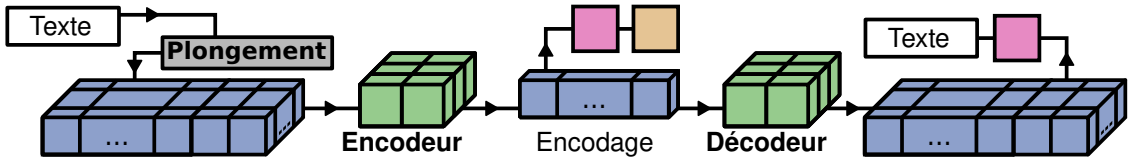


Figure 2.9 – *Adaptation d'un AE pour une tâche de classification de textes.*

approches. C-LSTM est un réseau hybride entre un LSTM et un CNN [Zho+15]. Il produit, à l'aide des fenêtres des filtres, des représentations des N-grammes qui sont ensuite passées successivement aux cellules LSTM. Kowsari et al. [Kow+17] proposent un réseau combinant un MLP, un CNN et un RNN (LSTM ou GRU) pour la classification de textes. Les approches les plus récentes et les plus efficaces pour les tâches de TAL sont probablement les adaptations de BERT et les nouvelles versions de GPT. RoBERTa [Liu+19b] est une version entraînée sur un plus gros volume de données que BERT. ALBERT est une version avec moins de paramètres et donc plus rapide que BERT [Lan+19] tout comme DistilBERT [San+19]. ELECTRA [Cla+20] masque certains tokens en les remplaçant par d'autres tokens probables durant la phase d'apprentissage. XLNet mixe les stratégies de BERT et GPT en masquant à la manière de GPT et de BERT les données dans les cellules d'attention [Yan+19]. BERTSUM utilise plusieurs tokens de classification [CLS] (section 2.2.2.4 ou figure 2.7) pour conserver des représentations du texte ou des morceaux du texte. Il est utile, par exemple, pour la tâche de construction de résumés. GPT-2 [Rad+19] et GPT-3 [Bro+20] sont des versions avec plus de paramètres et donc plus efficaces de GPT. Brown et al. [Bro+20] ont d'ailleurs, dans leur article sur GPT-3, alerté des risques que pouvait représenter un réseau de neurones comme celui-ci, qui pourrait produire des textes pour lesquels un humain ne pourrait dire si ils sont la production d'un humain ou non. Bien que les transformeurs proposent une profusion de nouveaux modèles, les précédentes architectures trouvent encore de l'intérêt [DVC21 ; Ish+21], une frise temporelle de l'apparition des techniques en TAL est proposée dans la figure 2.10, ainsi qu'une mise en lumière de quelques réseaux de neurones parmi la multitude de réseaux produits ces dernières années.

Le principal problème des réseaux de neurones est leur complexité et ainsi la capacité qu'ont leurs utilisateurs à expliquer les prédictions et à comprendre comment ils fonctionnent. On parle ici de problème d'explicabilité, d'interprétabilité, de transparence issu du caractère de boîte noire des réseaux de neurones. Nous développons et définissons ces problématiques dans la section 2.3.

2.3 Interprétabilité : Explicabilité et Transparence

Les notions d'interprétabilité, d'explicabilité et de transparence ne font pas consensus dans la littérature. Par exemple, Lipton [Lip18] définit deux concepts qui, ensemble, définissent l'interprétabilité : la transparence et les explications post-hoc. Waltl et Vogl [WV18] présentent la transparence et l'interprétabilité comme des sous-catégories de l'explicabilité. Beaudouin et al. [Bea+20] utilisent l'interprétabilité et l'explicabilité comme des synonymes. Enfin, Chatzimparmpas et al. [Cha+20], dans leur étude de la

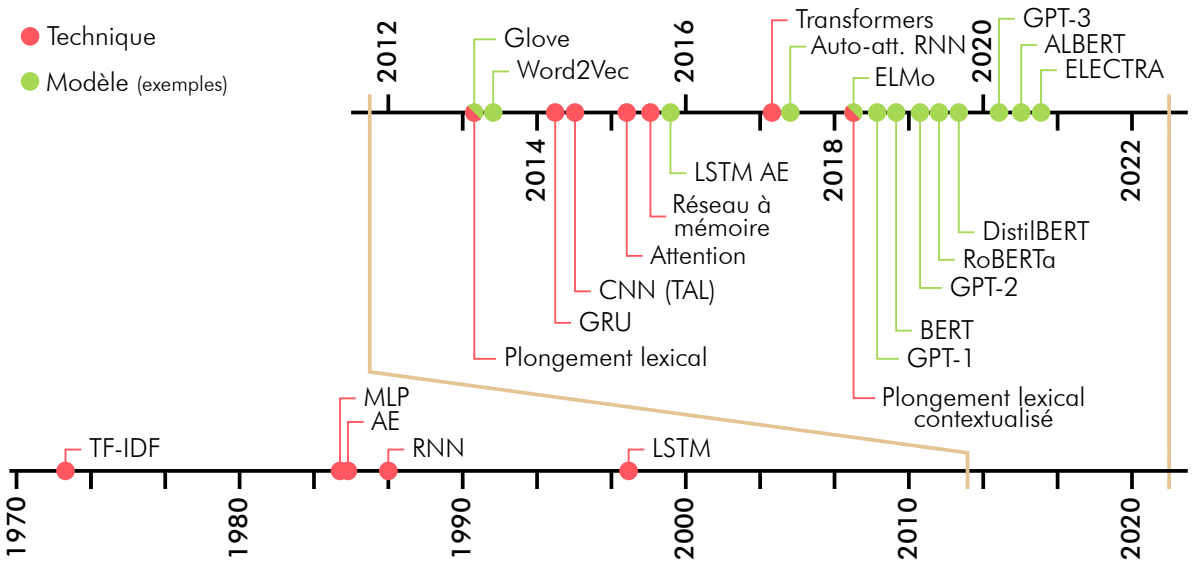


FIGURE 2.10 – *Apparition de techniques en TAL.*

littérature, utilisent les définitions de GILPIN et al. [Gil+18] qui présentent l'explicabilité comme la possibilité pour un modèle de résumer les raisons de son comportement et l'interprétabilité comme la compréhension de ce qu'un modèle a fait. Il est important de souligner, qu'au delà des définitions, se cache en réalité, derrière ces termes, une recherche de confiance des utilisateurs dans les réseaux de neurones qu'ils utilisent ou dont ils sont les cibles. Les critères de performance ne suffisent généralement pas. De ce fait, les utilisateurs cherchent à avoir des intuitions sur le fonctionnement des réseaux de neurones. Ces dernières servent à développer cette confiance de la part de l'humain envers le réseau de neurones.

Ces questions sont essentielles au vue de l'évolution des techniques en Intelligence Artificielle (IA). Comme le montre le communiqué, COM(2019) 168, du 9 avril 2019, de la commission européenne, intitulé "Renforcer la confiance dans l'intelligence artificielle axée sur le facteur humain" [1] :

> La traçabilité des systèmes d'IA doit être assurée; il est important d'enregistrer et de documenter les décisions prises par les systèmes, ainsi que l'ensemble du processus (description de la collecte et de l'étiquetage des données, description de l'algorithme utilisé) qui a abouti aux décisions. Dans le même ordre d'idées, il convient de veiller, dans la mesure du possible, à **l'explicabilité du processus de prise de décision algorithmique** à l'intention des personnes concernées, selon des modalités adaptées à leur cas. Il convient de poursuivre **les travaux de recherche en cours sur la mise au point de mécanismes d'explicabilité**. Des explications doivent également être fournies sur la manière dont un système d'IA influence et façonne le processus de prise de décision organisationnel, les choix opérés dans la conception du système, ainsi que la **justification de son déploiement** (de manière à assurer non seulement la transparence des données et du système, mais aussi la transparence du modèle économique). Enfin, il est important de communiquer des informations appropriées sur les capacités et les limites du système d'IA aux différentes parties concernées, selon des modalités adaptées au contexte d'utilisation concerné. En outre, les systèmes d'IA devraient être identifiables en tant que tels, de manière à ce que

1. https://www.eesc.europa.eu/fr/our-work/opinions-information-reports/opinions/renforcer-la-confiance-dans-lintelligence-artificielle-axee-sur-le-facteur-humain

> les utilisateurs sachent qu'ils interagissent avec un système d'IA et puissent identifier les personnes qui en sont responsables.
>
> (Commission européenne, COM(2019) 168)

Face au manque de consensus dans la littérature sur les définitions, nous précisons, dans la suite de cette section, ce qu'est l'explicabilité et la transparence d'un réseau de neurones dans les deux premières sections (section 2.3.1 et section 2.3.2). Puis, dans la dernière section, nous utilisons les précédentes définitions pour présenter ce qu'est l'interprétabilité (section 2.3.3). Enfin, nous présentons différentes méthodes participant à la transparence ou à l'interprétabilité des réseaux de neurones en ciblant les architectures présentées dans la section 2.2 auxquelles elles s'appliquent.

2.3.1 Transparence d'un modèle

Dans nos travaux, nous définissons la transparence comme **la facilité avec laquelle un humain peut comprendre et reproduire le fonctionnement d'un modèle, indépendamment d'une prédiction**. Par exemple, une régression linéaire est un modèle transparent car son fonctionnement est facilement compréhensible et reproductible par un humain. Par ailleurs, un réseau de neurones profond dépend de l'activation de millions de neurones, possédant chacun un grand nombre de paramètres. Il est donc bien moins compréhensible et reproductible, et donc moins transparent. C'est la raison pour laquelle les réseaux de neurones profonds sont des boîtes noires. Il est difficile, voire impossible, de comprendre précisément leur fonctionnement ou d'expliquer pourquoi telle variable dans un vecteur de représentation est définie à une valeur précise et ce que cette valeur représente. Pour un vecteur donné, un humain peut facilement traiter les données comme le fait une régression linéaire. En revanche, il lui est impossible de traiter les données comme le fait un réseau de neurones profond. Enfin, d'après Lipton [Lip18], la transparence d'un modèle peut être divisée en trois parties :

Tr.1 La compréhension globale du **fonctionnement du modèle** ;

Tr.2 La compréhension des différentes **parties du modèle** ;

Tr.3 La compréhension des **mécanismes d'apprentissage** et leur convergence vers une solution optimale.

Par exemple, il est plus facile de comprendre la convergence d'un réseau de neurones profond vers une solution optimale (*e.g.* à l'aide de la descente de gradient [RHW85]) ou à l'aide de la compréhension des parties de ce dernier que de comprendre son fonctionnement global.

Pour ce qui concerne **la compréhension globale du modèle** (Tr.1), lorsque sa complexité augmente, la compréhension possible d'un modèle diminue. Lorsque l'on compare le fonctionnement d'une cellule RNN classique et le fonctionnement d'une cellule LSTM, on peut observer que comprendre les mécanismes qu'utilisent les cellules LSTM (figure 2.5) sont bien moins préhensibles que les simples opérations d'une cellule RNN classique (figure 2.4). De la même manière, la construction des vecteurs de représentation d'un corpus de textes est complètement transparente si l'on utilise le

TF-IDF. En effet, la fonction mathématique et les paramètres de cette fonction pour une valeur d'un vecteur de représentation d'un texte est connue. En revanche, l'utilisation du plongement lexical ne l'est pas car on ne peut pas expliquer les multiples calculs qui ont amené une valeur à être définie (figure 2.1). Finalement, la compréhension globale du modèle passe par de nombreux mécanismes. Les techniques de visualisation de données permettent de visualiser le réseau de neurones et donc de comprendre comment évolue la dimension des vecteurs de représentation au sein de celui-ci [LeN19]. Dans la communauté scientifique, il n'est pas toujours simple de répondre à la question de la nécessité d'utiliser ou non des modèles très complexes [Rud19], spécifiquement pour les tâches à grands enjeux. Cela fait écho au texte proposé par la commission européenne qui précise qu'il faut être capable de "**justifier le déploiement en production**" des systèmes d'IA.

Pour ce qui concerne **la compréhension des parties du modèle** (Tr.2), des outils proposent la visualisation des réseaux de neurones et de leurs parties afin d'observer comment celles-ci se connectent. HARLEY [Har15] visualise l'évolution des vecteurs de représentation entre les couches d'un CNN. Cet aspect est donc assez proche de celui concernant la compréhension globale du modèle. Cependant, certains travaux accompagnent les utilisateurs dans la manipulation des couches spécifiques d'un réseau de neurones d'exemple afin de comprendre l'utilité de celles-ci [Smi+17]. Pour les réseaux de neurones, la compréhension des parties comprend celle des mécanismes des neurones, couches de neurones et fonctions d'activation par exemple (figure 2.2). Ce n'est qu'en comprenant les parties d'un modèle complexe que l'on peut rendre celui-ci plus transparent. Par exemple, comprendre le fonctionnement des cellules d'attention dans un réseau de neurones auto-attentif (figure 2.7) aide à comprendre comment celui-ci modifie, à chaque étape, l'espace de représentation des données, quelle est l'utilité des cellules et comment lire une matrice d'attention.

Pour ce qui concerne **la compréhension des mécanismes d'apprentissage et la convergence vers une solution optimale** (Tr.3), SMILKOV et al. [Smi+17] proposent également dans leur outil de comprendre les mécanismes et les paramètres d'apprentissage en effectuant l'apprentissage du réseau de neurones selon différents paramètres de manière interactive. Dans le cas des réseaux de neurones, le fait que ceux-ci n'aient pas une seule solution optimale et puissent proposer des paramètres différents pour un même modèle et une même tâche n'améliore pas leur transparence. La notion de descente de gradient insuffle une intuition à propos de quels mécanismes tendent à proposer des modèles efficaces mais laisse également imaginer que lors de l'apprentissage, les modèles puissent être bloqués dans des minimums locaux et ne pas être efficaces. En somme, c'est une des composantes les plus importantes à la non transparence des réseaux de neurones.

Finalement, ces trois aspects de transparence sont grandement liés à des problématiques de pédagogie. S'il n'est pas toujours possible de rendre un modèle transparent, c'est d'autant plus vrai pour les réseaux de neurones très profonds ou très complexes. On peut alors expliquer les prédictions pour insuffler une certaine confiance des utilisateurs dans ces modèles.

2.3.2 Explicabilité d'un modèle

Comme décrit précédemment, en apprentissage profond, les décisions prises par un réseau de neurones sont basées sur l'activation, ou non, de millions de neurones. Il est donc impossible pour un être humain de saisir toutes les nuances des décisions issues de ces réseaux. L'explication d'une prédiction donnée, ou **explication post-hoc**, est faite lorsque des indicateurs, issus ou non du fonctionnement d'un modèle, sont utilisés pour expliquer sa décision [Lip18]. Les explications post-hoc sont aussi appelées **explications locales** [Bea+20]. Le terme post-hoc fait référence au fait que les explications sont générées après l'inférence et sans ré-entraînement. Si une explication associée à une prédiction sert marginalement à interpréter un réseau, la multiplication des explications peut donner aux utilisateurs des intuitions sur le fonctionnement du modèle utilisé. Dans cette partie, nous allons spécifiquement nous intéresser aux méthodes traitant l'explicabilité et les explications post-hocs pour des réseaux de neurones en TAL. Lipton [Lip18] propose quatre catégories d'explications post-hoc :

Ex.1 Les explications **verbales** ou écrites justifiant les prédictions ;

Ex.2 Les explications **locales** qui peuvent donner accès à des explications plus simples ne concernant qu'un sous-ensemble de l'espace de données ;

Ex.3 Les explications de **complexité modérée** qui présentent des comportements pour des exemples similaires ;

Ex.4 Les techniques de **visualisation** pour explorer l'espace de représentation des données ou afficher des indications sur les parties, dans les données d'entrée, qui contribuent à la prédiction.

Certaines méthodes d'explication appartiennent à plusieurs de ces catégories.

2.3.2.1 Explications verbales ou écrites

Les explications verbales ou écrites (Ex.1) justifiant les prédictions sont les moins utilisées en TAL. Par définition, la donnée étant déjà textuelle, la décrire de nouveau avec du texte ne revêt que peu d'intérêt. En traitement des images, Zhang et al. [Zha+17] entraînent un réseau de neurones contenant un CNN et un LSTM qui prend en entrée des images et des textes décrivant ou possédant un lien avec des images. L'objectif est alors d'entraîner ce réseau à décrire les images en vue de justifier une prédiction. Ils proposent un exemple sur des images de cancer de la vessie et des diagnostics découlant de ces images. De même, Xu et al. [Xu+15] générerent la légende d'une image à l'aide d'un réseau similaire. On peut alors imaginer le même système associé à une tâche de classification, qui justifierait pourquoi une image est classée d'une manière ou d'une autre. L'explication met alors en lumière les liens entre les pixels d'une image et les mots composant la légende.

2.3.2.2 Explications locales

Les explications locales (Ex.2) construisent des explications à partir des règles utilisées par le modèle et qui ne sont valables que dans un espace donné. Ribeiro, Singh et Guestrin [RSG16] proposent LIME, un système indifférent au modèle utilisé

(agnostique) ouvrant l'accès à un modèle linéaire éparse (et donc plus transparent) dans un espace réduit des données perturbées (perturbées aléatoirement) depuis celles que l'on souhaite expliquer. Ainsi, l'utilisateur a une idée de l'impact de chacune des composantes de la donnée d'entrée sur sa prédiction. Cependant, c'est la multiplication de ces observations locales qui donne une idée globale du fonctionnement du modèle permettant de lui accorder une plus grande confiance. De plus, pour certaines techniques d'explication des prédictions, il faut que la donnée d'entrée soit elle-même interprétable (le TF-IDF est explicable alors qu'un plongement lexical ne l'est pas). Utiliser LIME avec une représentation de textes issue de plongement lexical revêt un moindre intérêt.

Certains travaux pointent les limites de LIME [Tan+19] quand d'autres essayent de proposer des variantes et améliorations : Anchors [RSG18] et LORE [Gui+18a]. Anchors [RSG18] diffère de LIME avec un modèle plus simple pour expliquer les prédictions, basé sur des "ancres" ("anchors" en anglais). Les ancres informe l'utilisateur en fonction des valeurs des attributs des données. Les ancres fonctionnent par système de SI-ALORS. Par exemple, pour une tâche d'analyse de sentiments de textes : **si** le texte contient une fois le mot "mauvais" **alors** sa prédiction sera négative avec une précision de 89%. Tout comme pour LIME, si la donnée d'entrée n'est pas interprétable, alors utiliser Anchors ne revêt que peu d'intérêt. LORE [Gui+18a] utilise un arbre de décision comme explication. Cela le rend plus efficace qu'Anchors avec les variables continues, ce dernier ne traitant que les données discrètes (ou discrétisées) en construisant des règles ne contenant que des tests d'égalité. Ces trois méthodes montrent l'intérêt de la communauté pour les stratégies agnostiques locales qui ne sont pas issues du fonctionnement du réseau de neurones pour en expliquer les prédictions.

2.3.2.3 Explications de complexité modérée

Les explications de complexité modérée, qui présentent des comportements pour des exemples similaires, peuvent insuffler une certaine confiance à la fois dans le comportement du modèle mais aussi dans la cohérence de l'espace de représentation des données construit (Ex.3). MIKOLOV et al. [Mik+13a ; Mik+13b], à l'introduction de Word2Vec, comparent les relations sémantiques entre des couples de mots similaires avec des relations sémantiques identiques et donc entre les représentations construites par le plongement lexical. Ils effectuent aussi une projection de ces relations dans un espace à deux dimensions à l'aide d'une ACP (section 2.4.1) pour montrer que les représentations construites ont du sens. Cette projection, présentée dans la figure 2.11, montre en effet qu'un lien sémantique d'un pays à sa capitale est visible et projette les données sur l'axe des abscisses principalement et d'une distance équivalente. L'ACP est une transformation linéaire, ce qui montre que cette projection sémantique existe aussi dans l'espace de représentation des données d'origine. De plus, on peut observer que cette projection semble avoir également ordonné géographiquement les couples pays-capitale sur l'axe des ordonnées. En effet, les couples semblent être placés grossièrement de ceux les plus à l'est à ceux les plus à l'ouest. Ce genre d'explication n'a donc pas pour objectif de présenter l'intégralité du modèle mais des exemples suffisamment solides pour donner confiance dans la cohérence du modèle ou dans l'espace de représentation construit.

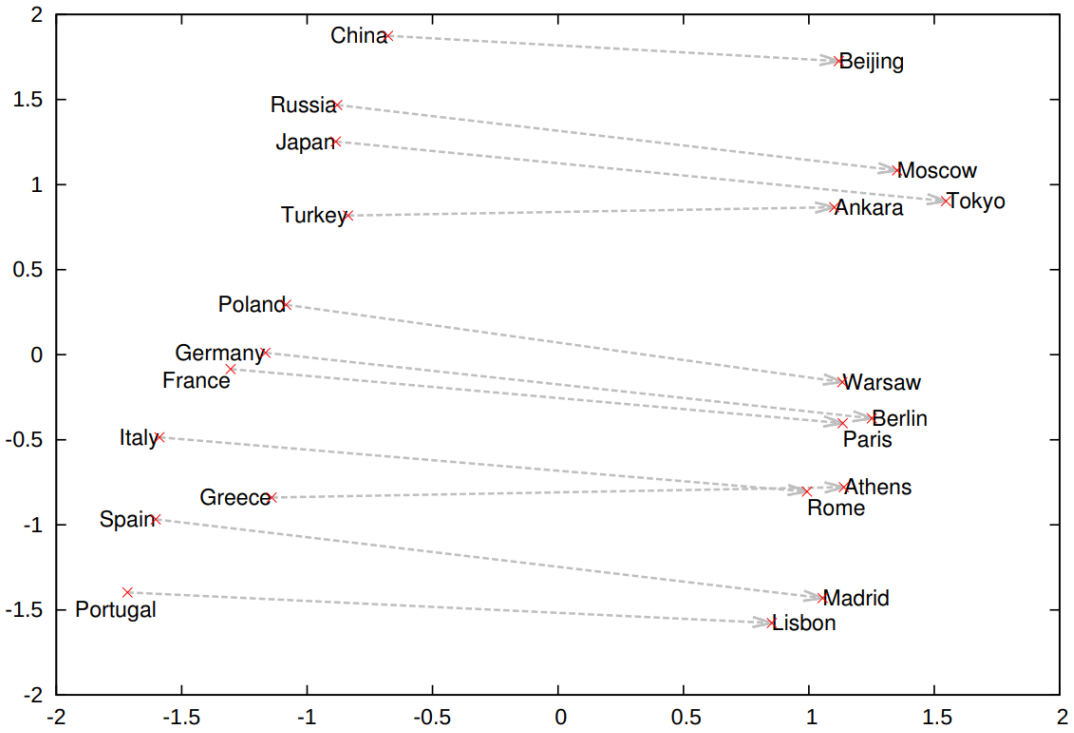


FIGURE 2.11 – *Vecteurs de représentation de pays et de leur capitale projetée par ACP laissant apparaître les relations sémantiques et géographiques entre les représentations construites par Word2Vec* [*Mik+13b*].

Dans la littérature, pour expliquer une prédiction, on se sert aussi bien de données voisines que de données générées à partir de la donnée à expliquer. Cependant, comme le montrent LAUGEL et al. [Lau+19], construire des explications ou des explications contre-factuelles pour une donnée à l'aide de ses voisins n'est pas chose aisée. Une explication contre-factuelle présente pour une modification d'une donnée d'entrée le comportement qu'aurait eu le modèle pour mettre en lumière l'importance d'une variable dans la prédiction. LAUGEL et al. [Lau+19] montrent que la plupart des approches de l'état de l'art ne font pas la différence entre les explications justifiées et injustifiées, ce qui conduit à des explications moins utiles. LAUGEL et al. [Lau+19] définissent l'explication de la prédiction $f(x_1)$ pour une donnée d'entrée x_1 à l'aide de la prédiction $f(x_2)$ d'un voisin x_2 comme justifiée si il existe un chemin continu h qui lie x_1 et x_2 dans l'espace de représentation des données et qui ne traverse pas la frontière de décision de f. Certaines méthodes locales comme Anchors [RSG18], et LORE [Gui+18a] dans une moindre mesure (section 2.3.2.2), expliquent les prédictions dans un espace dans lequel la frontière de décision n'est pas présente et pallient ce problème.

2.3.2.4 Techniques de visualisation pour l'explicabilité

Les techniques de visualisation utilisées dans tous les domaines de l'apprentissage automatique permettent d'explorer l'espace de représentation des données ou d'afficher des indications sur la partie, dans les données d'entrée, qui contribue à la prédiction (Ex.4). Elles servent d'autant plus dans le domaine du traitement des

images du fait que les données d'entrée sont visuelles [Hoh+19]. Nous listons ici de nombreuses contributions du domaine de l'explicabilité des prédictions des réseaux de neurones en TAL, utilisant les méthodes de visualisation de données, pour souligner l'intérêt croissant de la communauté pour ces sujets. Ces méthodes utilisent, le plus souvent, des cartes de chaleur pour présenter les contributions des tokens. L'ensemble des méthodes indiquant les parties, dans les données d'entrée, qui contribuent à la prédiction est présenté dans la figure 2.12. Cette figure montre les similitudes et le récent intérêt de la communauté pour ces méthodes. La visualisation et l'exploration des espaces de représentation des données seront traitées dans la section 2.4.1. Enfin, les réseaux auto-attentifs ayant supplantés les RNN pour de nombreuses tâches en TAL, nous décrivons les méthodes d'explication les concernant spécifiquement dans la section 2.4.4.

Dès l'avènement des réseaux de neurones, des travaux se sont intéressés à identifier les parties de la donnée d'entrée ayant contribué à la prédiction [Gar91 ; Mil95]. En TAL, de nombreuses méthodes existent. Elles visent à mettre en lumière les tokens ou groupes de tokens qui ont le plus participé à une prédiction. Du fait des nombreuses méthodes ayant émergées depuis les années 90, nous nous concentrons, ici, sur les plus récentes et montrons leur profusion aux cours des dernières années (figure 2.12).

L'**analyse de sensibilité**, ou "sensibility analysis" en anglais (SA) [Li+15 ; Arr+17], appliquée aux RNN et LSTM est l'une des premières méthodes. La SA consiste à produire, pour chaque token d'entrée, un score de pertinence issu des dérivées partielles au carré des dimensions du vecteur de représentation du token à l'aide de la rétro-propagation des gradients [RHW85]. Ensuite, ces scores sont sommés pour produire, pour chaque token, un score de pertinence. Dans la même idée, la LRP est une méthode issue du traitement d'images [Bac+15] qui a depuis été utilisée dans le domaine du TAL par ARRAS et al. [Arr+17]. Cette adaptation de la LRP est un autre des premiers pas dans l'explicabilité des prédictions en TAL, notamment des RNN. La LRP redistribue la part de responsabilité dans la prédiction à l'aide d'une donnée de référence, en partant de la couche de sortie du réseau et en rétro-propageant cette quantité jusqu'à la donnée d'entrée (une description d'une donnée de référence sera apportée dans le paragraphe suivant). Cette quantité rétro-propagée utilise les poids du réseau et les activations neuronales de la propagation classique pour rétro-propager la sortie du réseau à travers celui-ci jusqu'à la donnée d'entrée. DeepLIFT [SGK17] utilise un fonctionnement similaire mais inclut deux autres axiomes en plus de ceux énoncés dans la LRP, pour mieux définir ce qu'il faut rétro-propager. Enfin, il existe des stratégies plus rigoureuses mathématiquement que la LRP qui utilisent plus d'axiomes. Ces dernières utilisent les valeurs de Shapley.

Les méthodes utilisant **les valeurs de Shapley** pour l'apprentissage automatique [ŠK10 ; ŠK14] comparent le comportement d'une donnée spécifique vis-à-vis de la moyenne des comportements du modèle. Cette stratégie découle de la théorie des jeux de SHAPLEY [Sha53]. L'idée, pour expliquer des prédictions, est d'analyser les prédictions d'un modèle f en distribuant la différence entre la prédiction $f(x_k)$ et la moyenne des prédictions $\mathbb{E}[f(X)]$ pour une donnée x_k appartenant à un jeu de données X. La part de responsabilité $\phi_j(f, x_k)$ d'une variable $x_{k,j}$ dans une prédiction $f(x_k)$ est donc :

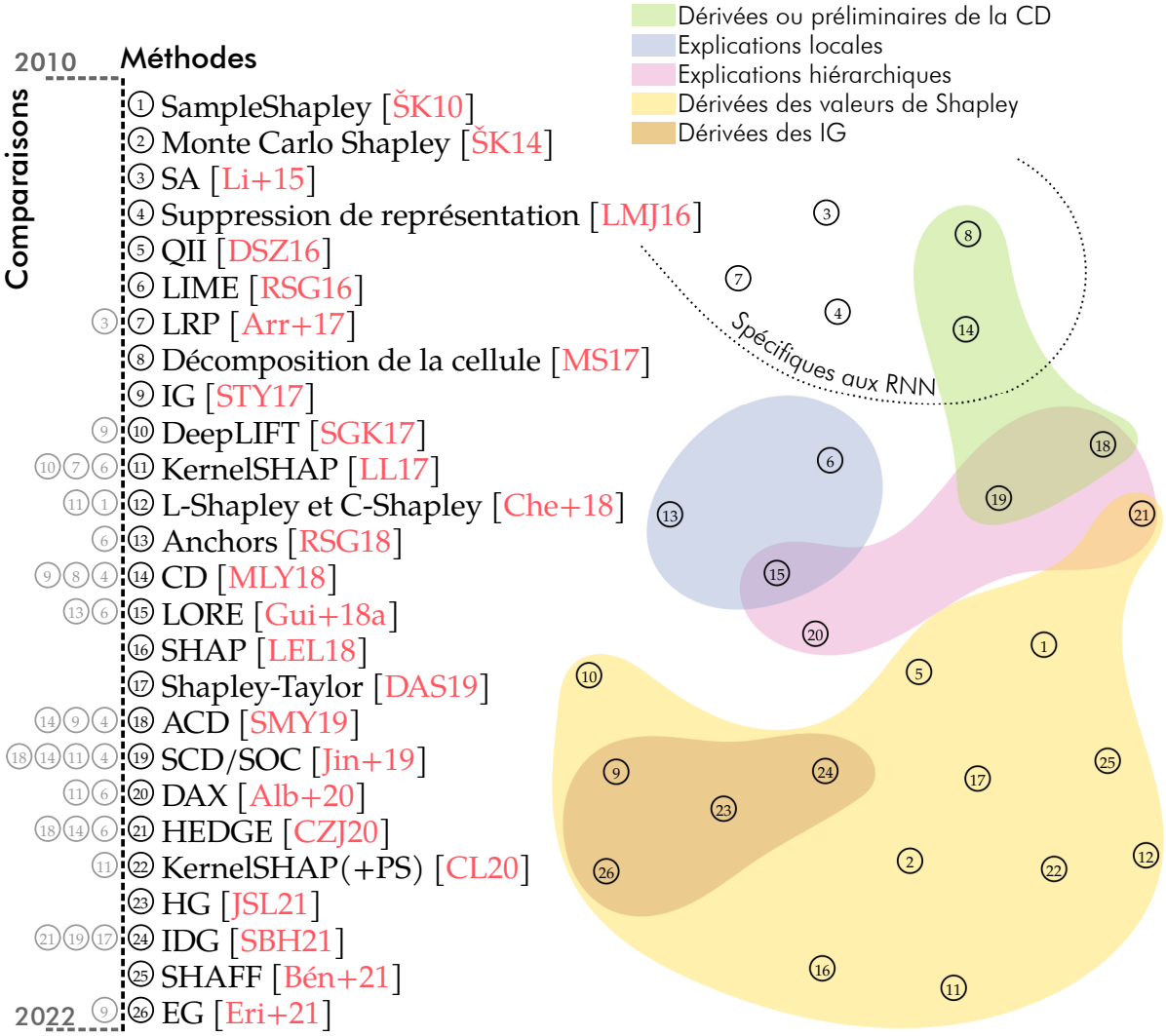


Figure 2.12 – *Projection à l'aide de l'algorithme UMAP [MHM18] des vecteurs TF-IDF des articles concernant l'explication des prédictions. À gauche se dessine une frise chronologique, allant de 2010 à 2022, listant les méthodes et les précédentes méthodes auxquelles elles se sont comparées dans leur article introductif (en gris). À droite sont présentés dans la projection les différents groupes de méthodes que nous avons identifiés. Un article peut appartenir à plusieurs groupes. En haut de cette projection, on peut observer la proximité qu'ont les articles spécifiques aux RNN.*

$$\phi_j(f, x_k) = \sum_{u \subseteq \{1,\dots,d\} \setminus j} \frac{(d - |u| - 1)!|u|!}{d!} (\mathbb{E}[f(X_{u \cup \{j\}})] - \mathbb{E}[f(X_u)]) \tag{2.15}$$

Ici, d est le nombre de variables dans le jeu de données X, X_u correspond à X, où pour un sous ensemble u de $\{1, \dots, d\} \setminus j$, $\forall i \in u, X_i = x_{k,i}$ et $X_{u \cup \{j\}}$ correspond à X, où pour un sous ensemble u de $\{1, \dots, d\} \setminus j$, $\forall i \in u, X_i = x_{k,i}$ et $X_j = x_{k,j}$. La somme des contributions des variables d'une donnée x_k est donc égale à la différence entre la moyenne des prédictions $\mathbb{E}[f(X)]$ et la prédiction $f(x_k)$ (voir équation 2.16). Chaque contribution (voir équation 2.15) présente donc comment une variable k d'une

donnée x éloigne la prédiction de cette donnée de la moyenne des prédictions. Au fur et à mesure que le nombre de variables dans le jeu de données augmente (ce qui est particulièrement le cas en apprentissage automatique profond en TAL), le nombre de sous-ensembles u devient exponentiellement plus grand. ŠTRUMBELJ et KONONENKO [ŠK14] proposent une première approximation de la solution optimale, à l'aide de la méthode d'échantillonnage de Monte Carlo.

$$f(x_k) - \mathbb{E}[f(X)] = \sum_{j=1}^{d} \phi_j(f, x_k) \tag{2.16}$$

DHAMDHERE, AGARWAL et SUNDARARAJAN [DAS19] utilisent la valeur de Shapley [Sha53], pour attribuer à la prédiction d'un modèle, une partie des données d'entrée, afin d'expliquer les prédictions d'un CNN. Leur méthode, appelée indice d'interaction de Shapley-Taylor, associe la prédiction du modèle aux interactions des sous-ensembles de la donnée d'entrée, à savoir pour le texte, aux sous-ensembles de tokens. CHEN, ZHENG et JI [CZJ20] présentent dans HEDGE, des explications hiérarchiques, et mesurent les interactions des sous-ensembles (de tokens par exemple) de la donnée d'entrée à l'aide de l'indice d'interaction de Shapley. Comme le présentent CHEN, ZHENG et JI [CZJ20], de nombreux autres travaux utilisent les valeurs de Shapley ou cherchent comment les approximer : SampleShapley [ŠK10], QII [DSZ16], KernelSHAP [LL17 ; CL20], L/C-Shapley [Che+18]. Ces différentes méthodes ont été comparées par CHEN et al. [Che+18] sur des données textuelles. Cependant, la méthode de référence aujourd'hui découlant des valeurs de Shapley, publiée la même année, est SHAP [LEL18]. SHAP (ou KernelSHAP) [LEL18 ; LL17] permet d'observer l'influence positive et négative des parties de la donnée d'entrée dans la prédiction. SHAP compare cette valeur de prédiction à la valeur moyenne des prédictions du jeu de données. SHAP met aussi en avant l'influence de chacune des variables de manière globale sur l'ensemble des données. SHAFF [Bén+21], est une méthode récente qui calcule les valeurs de Shapley à l'aide de forêts aléatoires plutôt qu'à l'aide de la méthode de Monte Carlo. De nombreuses méthodes en explication des prédictions utilisent plus ou moins directement les valeurs de Shapley [DSZ16 ; STY17]. Nous retrouvons dans la figure 2.12 dans la catégorie "Dérivées des valeurs de Shapley" les méthodes y faisant référence explicitement. Certaines méthodes, comme les gradients intégrés (IG) [STY17], s'en inspirent.

Les **IG** [STY17] sont une technique qui mesure, pour chaque composante de la donnée d'entrée, un score d'importance à l'aide de données d'interpolations entre la donnée à expliquer et une donnée de référence ne contenant pas d'information. Les IG donneront naissance à de nombreuses autres techniques pas toutes utilisées ou testées sur des données textuelles : les gradients espérés (EG) de ERION et al. [Eri+21] ou les gradients Hessians (HG) de JANIZEK, STURMFELS et LEE [JSL21] par exemple. L'objectif des IG est de calculer, pour une donnée x_k appartenant à un jeu de données X, l'importance de chacune de ses composantes en partant d'une donnée de référence x' ou x'_k. Cette donnée de référence ne doit pas contenir d'information et est donc, en général, la même pour toute donnée x_k du jeu de données X. En TAL, cela peut être, par exemple, un texte où tous les plongements lexicaux des tokens sont des vecteurs remplis de zéros.

Une fois cette donnée de référence choisie, un chemin entre la donnée de référence et la donnée à expliquer est tracé dans l'espace de représentation des données. Sur ce chemin, m données d'interpolations sont choisies pour approximer l'intégrale donnée dans l'équation 2.17. Pour chacune de ces données d'interpolations, on calcule le gradient de chacune des composantes de la donnée (une dimension d'un plongement lexical d'un des tokens de la donnée par exemple). Ensuite, il suffit de sommer ou de moyenner, pour chacune des données d'interpolations, les gradients des composantes de la donnée. Une approximation de l'influence d'une composante j d'une donnée x_k est donc le résultat de cette opération (voir équation 2.17). L'équation 2.18 est une approximation de l'équation 2.17. Cette approximation présente à quel point l'information contenue $(x_{k,j})$ dans une composante j pour une donnée x_k est différemment traitée par f (le modèle) qu'une absence d'information $x'_{k,j}$. Par conséquent, cette approximation représente à quel point cette composante apporte de l'information pour la prédiction. Sikdar, Bhattacharya et Heese [SBH21] proposent une adaptation des IG qui implique l'utilisation d'explications hiérarchiques : les IDG.

$$IG_j(f, x_k) = (x_{k,j} - x'_{k,j}) \int_0^1 \frac{\partial f(x'_k + \alpha(x_k - x'_k))}{\partial x_{k,j}} d\alpha \tag{2.17}$$

$$IG_j(f, x_k) \approx (x_{k,j} - x'_{k,j}) \times \sum_{n=1}^{m} \frac{\partial f(x'_k + \frac{n}{m}(x_k - x'_k))}{\partial x_{k,j}} \times \frac{1}{m} \tag{2.18}$$

Une autre méthode, **la suppression de représentation** [LMJ16], mesure la différence de probabilité de classification pour une classe donnée entre une donnée d'entrée et cette même donnée privée d'une de ses dimensions. Plus cette différence est grande, plus la dimension supprimée est importante.

D'autres travaux proposent la **décomposition de la cellule** (LSTM) [MS17] puis la décomposition contextuelle (CD) [MLY18]. La décomposition de la cellule se sert des probabilités p pour les données textuelles d'appartenir à une classe. Ainsi, $p_{k,c}$ est la probabilité que la donnée x_k appartienne à la classe c selon le modèle LSTM. Murdoch et Szlam [MS17] définissent alors la participation $\beta_{c,t}$ du token t à la prédiction pour la classe c de la manière suivante :

$$\beta_{c,t} = \exp(W_c(o_T \odot (\tanh(c_t) - \tanh(c_{t-1}))) \tag{2.19}$$

où W est la matrice de poids de la transformation linéaire qui projette le dernier état caché h_T vers une matrice de dimension $\mathbb{R}^{\mathbb{C}}$, où c_t et c_{t-1} sont respectivement les états de la cellule au token t et $t-1$ (voir équation 2.6) et o_T est la sortie de la porte de sortie (voir équation 2.5) du dernier token. Une démonstration de cette formule est présentée dans les travaux de Murdoch et Szlam [MS17]. La CD répond au même objectif, avec une méthode similaire de décomposition des cellules LSTM. Plutôt que de s'intéresser à la contribution d'un token à la prédiction, elle ouvre l'accès aux contributions de combinaisons de tokens [MLY18]. La CD a ensuite été adaptée à d'autres réseaux que les LSTM par Singh, Murdoch et Yu [SMY19]. Ils proposent une agglomération de la décomposition contextuelle (ACD). En plus des scores de CD, ils visualisent un

arbre pour voir les liens entre les mots et la CD pour chacun de ces groupes. C'est une des premières introductions d'arbres (ou d'explications hiérarchiques) dans la contribution à la prédiction en TAL. Elle insuffle une plus grande confiance chez les utilisateurs. Elle offre également une procédure de contrôle et de débogage simplifiée. JIN et al. [Jin+19] ont mis en lumière les limites de la CD et de l'ACD dans le calcul des interactions entre les phrases. Ils ont proposé un premier algorithme d'échantillonnage et de décomposition contextuelle (SCD) et un second algorithme d'échantillonnage et d'occlusion (SOC). Ces algorithmes quantifient l'importance des mots et des phrases indépendamment du contexte général, ce que n'étaient pas capable de faire la CD ou la ACD.

ALBINI et al. [Alb+20] développent DAX, une méthode d'observation pour les réseaux à propagation directe (CNN, MLP par exemple), de graphe des couches montrant lesquelles se sont activées à l'aide des nuages de mots. Ceux-ci correspondent aux mots ayant le plus activé ces couches dans les données d'entraînement. Cela met en lumière, notamment dans le cas de CNN, les spécificités de chaque filtre.

On remarque par le nombre et les associations des différentes méthodes, l'intérêt croissant de la communauté scientifique pour les méthodes d'explication des prédictions (figure 2.12). La plupart des travaux présentent d'ailleurs une comparaison de leur méthode avec les précédentes. Nous ne cherchons pas ici à les comparer mais à présenter l'effervescence autour de ces sujets. Des travaux cherchant à présenter [LT19] et comparer [Arr+19] ces méthodes, parfois non spécifiques au domaine du TAL [Gui+18b], ont déjà été proposés. Certains des travaux cités dans cette section sont plus centraux et servent de référence en termes de comparaison. On peut notamment citer la suppression de représentation [LMJ16], LIME [RSG16], les IG [STY17], la CD [MLY18] ou SHAP (KernalSHAP) [LL17]. Enfin, les explications hiérarchiques gagnent en intérêt [SMY19 ; Jin+19 ; CZJ20 ; SBH21]. Des méthodes spécifiques pour expliquer les prédictions des réseaux auto-attentifs prennent une grande place ces dernières années du fait de l'effervescence autour de ces réseaux. Nous décrirons ces méthodes dans la section 2.4.4.

2.3.3 Interprétabilité d'un modèle

Finalement, l'interprétabilité est définie dans cette thèse comme la façon dont un modèle est ou devient plus interprétable en fonction de sa transparence et des explications produites. De nombreux travaux s'intéressent, encore aujourd'hui, à définir et à présenter l'ensemble des solutions à cette problématique d'interprétabilité [Fan+21 ; Bar+20 ; Gui+18b] alors que d'autres mettent en lumière les futurs challenges à relever dans ce domaine [SO21]. Dans la taxonomie de LIPTON [Lip18], le concept principal d'interprétabilité inclut la transparence et les explications post-hoc. Un utilisateur comprend le fonctionnement d'un modèle, ou son processus caché, en analysant sa structure ou ses prédictions. Cette compréhension définit l'interprétabilité de ce modèle. Un modèle totalement transparent est interprétable. Les explications post-hocs ne sont utiles à l'interprétabilité que pour un modèle qui n'est pas ou que trop peu transparent. Enfin, les composantes d'un modèle peuvent avoir une transparence différente ou être plus ou moins explicables. Travailler sur l'interprétabilité de chacune de ces parties

rend le modèle plus interprétable. Un exemple est proposé dans la figure 2.7. BERT n'est pas un modèle transparent. En revanche comprendre comment une cellule d'attention fonctionne individuellement donne une meilleure idée de son fonctionnement global. Les cellules d'attention ne sont pas totalement transparentes mais permettent d'accéder à une matrice (pour chacune) avec un score d'attention de chacun des tokens avec les autres. Ces matrices servent à l'explicabilité du fonctionnement des cellules d'attention. Elles aident donc à l'interprétabilité globale du réseau BERT [Dev+18]. Nous abordons les techniques de visualisation qui utilisent les vecteurs d'attention des réseaux de neurones dans la section 2.4.4.

Dans cette thèse, nous nous sommes concentrés sur les tâches de classification dichotomiques de textes à l'aide de réseaux de neurones. Nous avons proposé des solutions à l'interprétabilité des modèles, indépendamment de leurs structures, en nous basant sur des techniques d'explicabilité des prédictions. Pour ce faire, nous avons proposé des outils de visualisation des données permettant d'utiliser les techniques d'explicabilité présentées dans cette section, en complément de la visualisation des espaces de représentation des données et de la frontière de décision (section 2.4). Nous avons utilisé comme exemple la visualisation de l'attention dans BERT et dans un RNN.

Le tableau 2.1 présente une classification des articles utilisant des techniques de visualisation de données pour l'interprétabilité et leur classification dans les classifications de Lipton [Lip18] et de Hohman et al. [Hoh+19] (section 2.4). Il montre que les techniques de visualisation participent entièrement au processus d'interprétabilité et pas seulement à la visualisation des espaces de représentation ou des parties de la donnée ayant participé à la prédiction.

2.4 Visualisation des réseaux de neurones et de leurs applications à la classification de textes

La visualisation de données a pour objectif de représenter visuellement celles-ci pour pouvoir déceler et comprendre les informations qu'elles contiennent. Elle cherche donc à résumer, à mettre en lumière des caractéristiques des données pour assister les utilisateurs dans l'analyse, la recherche des informations précises, le requêtage ou la production de nouvelles données [Mun14]. Dans la classification de Lipton [Lip18], la quatrième catégorie (Ex.4) est dédiée aux techniques de visualisation explorant l'espace de représentation des données ou affichant des indications sur la partie, dans les données d'entrée, qui contribue à la prédiction. Néanmoins, les techniques de visualisation ne se cantonnent pas uniquement à cette catégorie. Dans cette section, nous allons donc présenter les différentes applications des techniques de visualisation de données dans le domaine des réseaux de neurones, et plus spécifiquement dans le domaine du TAL. Celles-ci couvrent de larges missions, comme l'aide à la transparence, à l'explicabilité ou plus simplement à la présentation des résultats issus des réseaux de neurones.

Il existe de nombreux objets ou valeurs à visualiser pour expliquer le fonctionnement des réseaux de neurones. Hohman et al. [Hoh+19] proposent dans leur état de l'art

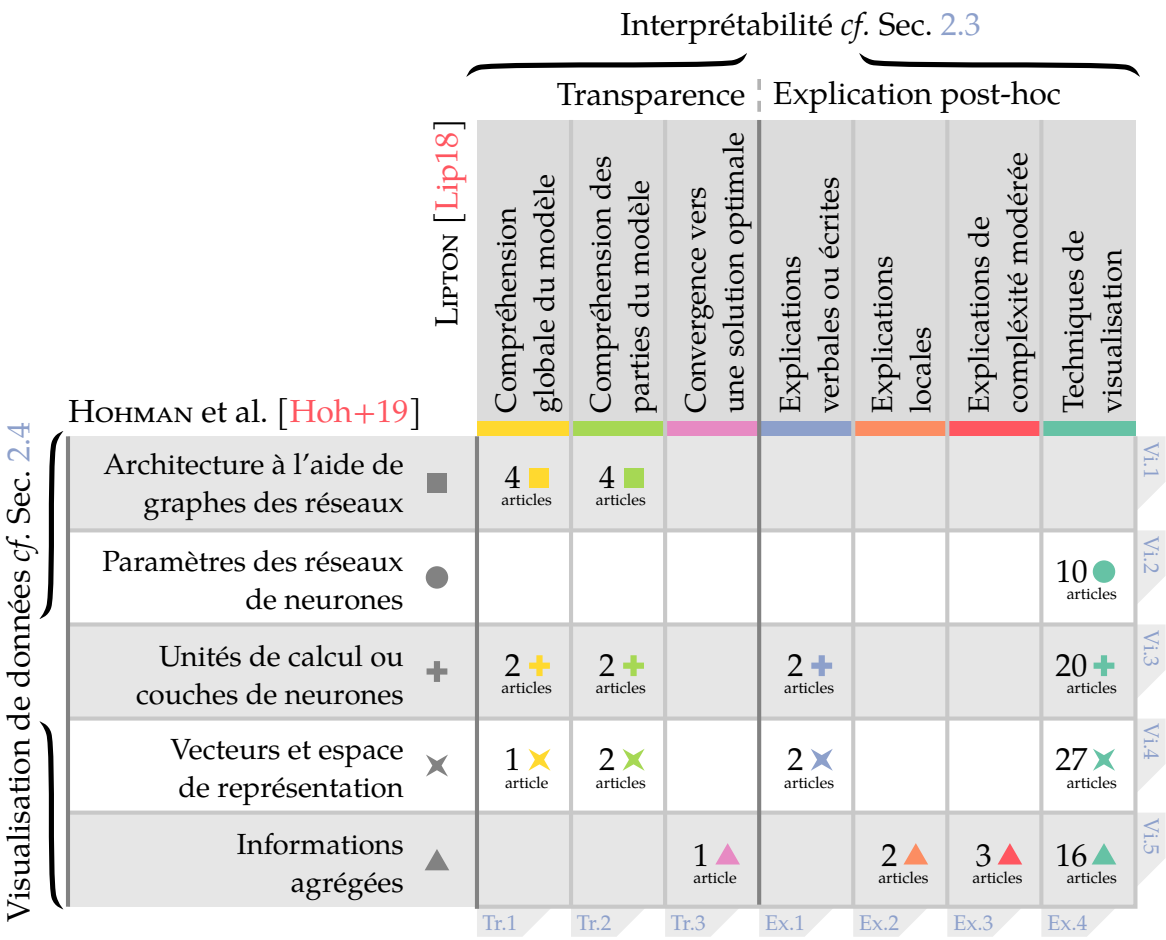


TABLE 2.1 – *Méthodes d'interprétabilité présentées selon les classifications de* HOHMAN *et al.* [*Hoh+19*] *et de* LIPTON [*Lip18*] *avec le nombre de travaux. On observe que les techniques de visualisation sont centrales. Les catégories peuvent ne pas être exhaustives, du fait qu'en fonction du sujet certaines catégories se confondent (i.e., les explications locales et les techniques de visualisation lors de la comparaison d'exemples dans une visualisation issue d'une technique de réduction de dimension, ou les unités de calcul et les vecteurs de représentation dans le cas d'utilisation de RNN ou réseaux auto-attentifs.)*

sur la visualisation de données dans l'apprentissage profond, cinq catégories d'objets à visualiser :

Vi.1 L'architecture des réseaux de neurones ;

Vi.2 Les paramètres des réseaux de neurones ;

Vi.3 Les unités de calcul ou couches de neurones ;

Vi.4 Les vecteurs de représentation des données dans des espaces à grandes dimensions ;

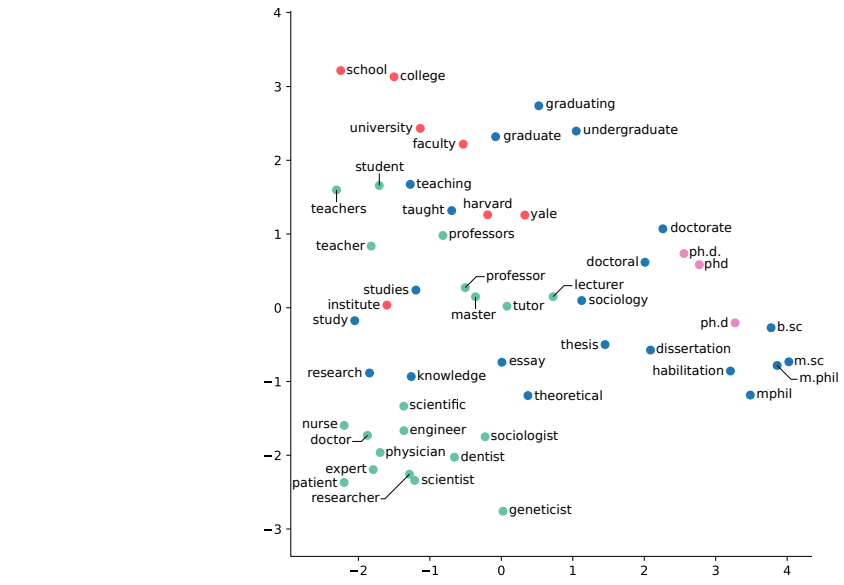
Vi.5 Les informations agrégées issues ou non du fonctionnement du modèle.

Certaines des méthodes présentées peuvent appartenir à plusieurs de ces catégories. La dernière catégorie couvre des travaux n'appartenant pas aux autres catégories. Dans cette section, nous traitons donc ces différents objets en nous intéressant à chaque fois aux réseaux de neurones concernés (MLP, CNN, RNN ou modèles auto-attentifs) par rapport à la tâche de classification de textes. Parfois, nous nous intéressons à d'autres tâches, notamment pour les MLP et les CNN, peu utilisés en TAL, et donc pour lesquels les travaux en visualisation de données sur des tâches de TAL sont quasiment inexistants. Ainsi, notre plan se découpe en six sections. La première traite de la visualisation des espaces de représentation des données de manière générale, avec une précision sur les espaces de représentation des tokens, mots ou textes. Les autres sections traitent ensuite successivement des différents types de réseaux présentés dans la section 2.2. Ces différentes sections présentent les travaux en visualisation de données concernant ces réseaux, de la visualisation de leur architecture, aux informations agrégées. Enfin, une dernière section traite des travaux en visualisation de données applicables à tous types de réseaux en TAL pour la classification de textes. Cette section évoque notamment les travaux de visualisation de la frontière de décision d'un réseau de neurones pour lesquels nous présentons une contribution dans le chapitre 3.

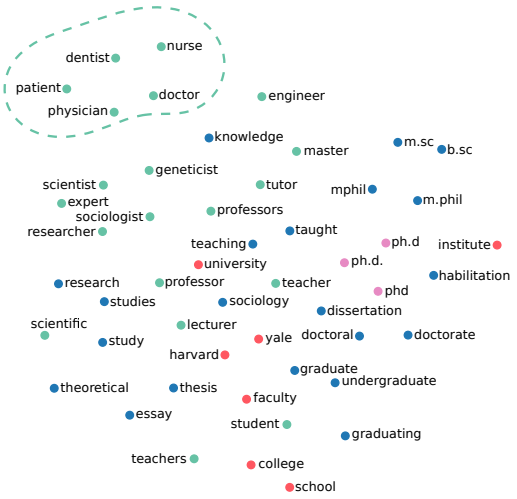
2.4.1 Visualisation des espaces de représentation

Les vecteurs de représentation des données sont créés dans des espaces à grandes dimensions. Par définition, ces espaces ne peuvent être visualisés directement. Dès lors, il est important de mettre au point des méthodes pour les visualiser en deux ou trois dimensions. L'Analyse en Composantes Principales (ACP) [Pea01], l'algorithme t-SNE [HR02 ; MH08] ou encore l'algorithme UMAP [MHM18] sont des techniques couramment utilisées [Smi+16 ; Li+15] pour les visualiser. Trois visualisations des données d'un même corpus de textes issues de ces trois méthodes sont proposées dans la figure 2.13.

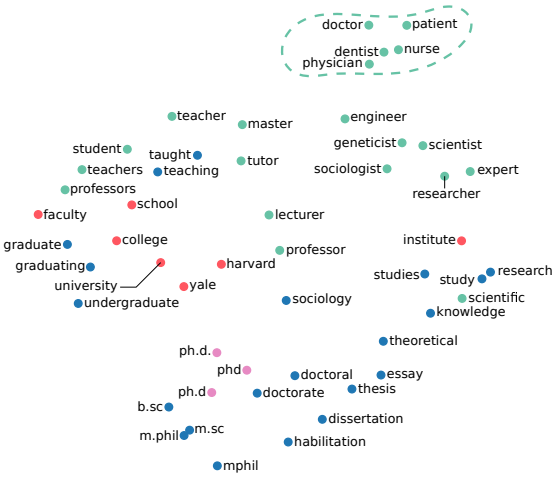
L'ACP [Pea01] utilise le degré de similarité (covariance) entre les variables pour projeter les dans le nouvel espace (généralement de dimension 2). Cette méthode perd le moins d'informations possible lors de la projection car elle construit les axes du nouvel espace pour qu'ils expliquent au maximum la dispersion des données. Un jeu de données avec des variables fortement corrélées entre elles ne perd, par exemple, que peu d'informations lors de la projection dans le nouvel espace, en comparaison avec un jeu de données sans variables corrélées.



(*a*)



(*b*)



(*c*)

FIGURE 2.13 – *Exemples de techniques de réduction de dimensions. (a) ACP. (b) t-SNE. (c) UMAP. Dans ces trois visualisations des données issues d'un même corpus de mots, on distingue différents groupes de données : les données en rouge sont relatives à l'université et aux études universitaires ; les données vertes sont relatives à des métiers ou des personnes ; les données roses sont trois écritures différentes pour l'acronyme anglais Ph.D ; et les données bleues ne semblent faire partie d'aucun groupe. On observe à l'aide de ces trois graphiques qu'au sein des données vertes, t-SNE et UMAP étaient plus efficaces pour identifier les données relatives au corps médical et à les séparer des autres. On peut également observer que UMAP semble avoir été plus efficace pour séparer les groupes entre eux. Ces techniques utilisent des principes différents et n'ont donc pas les mêmes performances, ni les même atouts. L'ACP semble avoir été, elle, capable de représenter, sur son axe des ordonnées, un ordonnancement des données les plus relatives à la recherche à celles les plus relatives à l'éducation.*

La t-SNE [HR02 ; MH08] est un algorithme non-linéaire itératif qui construit un nouvel espace de représentation des données. Au cours de ses itérations, la t-SNE place les données dans ce nouvel espace de manière à ce que les données proches dans l'espace original aient une probabilité élevée d'avoir des représentations proches dans le nouvel espace. Contrairement à l'ACP, cette technique ne construit pas de nouveaux axes interprétables mais une meilleure construction de clusters de données. Ces clusters voient le jour car la t-SNE n'est pas limitée par l'objectif de représentation de la dispersion des données.

UMAP [MHM18] est le dernier algorithme né dans la famille des techniques de réduction de dimension. UMAP commence par construire un graphe, à l'aide de complexes simpliciaux flous. Il s'agit d'une représentation d'un graphe pondéré dans lequel les poids des arêtes représentent la probabilité que deux sommets soient connectés. Pour déterminer la connexité, UMAP étend un rayon vers l'extérieur à partir de chaque sommet, reliant les sommets lorsque ces rayons se chevauchent. Le choix de ce rayon est crucial. Un choix trop petit conduit à de petits clusters isolés de sommets. Un choix trop grand connecte tout ensemble. Ce choix se fait en choisissant un rayon localement, sur la base de la distance au n-ième voisin le plus proche de chaque sommet. L'algorithme rend ensuite le graphe flou en diminuant la probabilité de connexion à mesure que le rayon augmente. Enfin, par défaut, chaque sommet doit être connecté au moins à son plus proche voisin. UMAP garantit ainsi que la structure locale est préservée en équilibre avec la structure globale. Au final, UMAP obtient un graphe de probabilité de voisinage entre les sommets, où chaque probabilité de voisinage dépend des deux voisins concernés et de leurs distances à leurs n autres voisins respectifs. Une fois ce graphe construit, la suite est presque identique à celle de la t-SNE. L'objectif est de se servir des probabilités de voisinage dans l'espace original de manière à ce que les données proches dans cet espace aient une probabilité élevée d'avoir des représentations proches dans le nouvel espace. EmbeddingProjector [Mar+15] visualise les espaces de représentation à l'aide de l'ACP, de la t-SNE et de UMAP (figure 2.14).

Dans le cas de la visualisation des espaces de représentation issus de plongements lexicaux, Mikolov et al. [Mik+13b] projettent dans un espace à deux dimensions à l'aide d'une ACP, différents mots de l'espace de représentation construit par Word2Vec pour présenter les liens sémantiques (figure 2.11). Un autre exemple plus récent, également appliqué aux espaces de représentation des tokens ou des mots, est celui de Liu et al. [Liu+18a]. Ils proposent le même type d'approche en utilisant SVM, ACP et des régularisations et comparent Word2Vec et Glove. Enfin, Li et al. [Li+18] ont élaboré un outil complet d'exploration de l'espace de représentation des mots en utilisant l'algorithme t-SNE. Liu et al. [Liu+19a] utilisent les différentes techniques de réduction de dimensions citées pour montrer l'existence de vecteurs d'attributs dans l'espace de représentation des tokens (ici des mots) qui contiennent du sens dans ces espaces. Ils identifient notamment un vecteur masculin-féminin et un vecteur participe présent. Ces constats s'approchent des travaux de Mikolov et al. [Mik+13a] avec la distribution ville-capitale dans l'espace construit par une ACP (figure 2.11). Les espaces de représentation des mots ne sont pas les seuls à être intéressants. Les documents et les N-grammes peuvent aussi être visualisés à l'aide de ces méthodes. Reif et al. [Rei+19] s'intéressent, par exemple, à l'espace de représentation des textes issu de

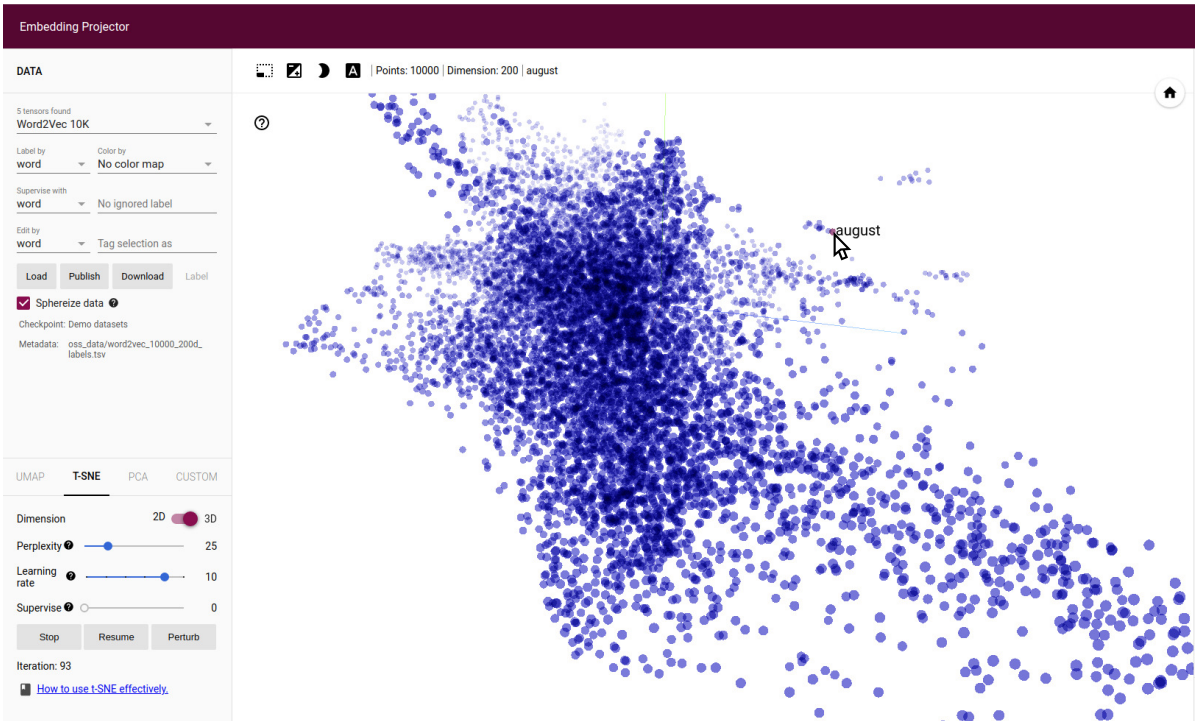


FIGURE 2.14 – *Visualisation de 10 000 mots issus du plongement lexical de Word2Vec [Mik+13b] à l'aide de UMAP présent dans EmbeddingProjector [Mar+15] dans un nouvel espace à trois dimensions. Le mot survolé fait partie d'un petit cluster de mots contenant les mois de l'année.*

BERT. BRUNNER et al. [Bru+18] s'intéressent à celui issu d'un LSTM AE. Embedding comparator [BCS19] (figure 2.15) compare tous les espaces de plongement lexical et même, plus largement en dehors du TAL, d'encodage des données. Les techniques de plongement lexical contextualisé gagnant en intérêt (section 2.2.1.2) des travaux se sont intéressés à l'analyse du plongement des tokens produits [Ber20 ; Pet+18a]. SEVASTJANOVA et al. [Sev+21] proposent LMExplorer, un outil de visualisation facilitant l'observation, dans un espace à deux dimensions, de l'évolution des représentations des tokens au travers des différentes couches de BERT d'un texte lors de son traitement.

Les techniques de réduction ne sont pas les seules solutions utilisables pour obtenir des informations sur les espaces de représentation. HEIMERL et GLEICHER [HG18] utilisent un simple axe où ils affichent les distances entre un mot choisi et ses plus proches voisins pour obtenir des informations sur ces voisinages dans différents espaces de représentation des tokens. Ils utilisent aussi des visualisations de co-occurrences ou de distances de tokens à deux autres tokens (espace à deux dimensions) pour obtenir des informations sur ces espaces de représentation.

2.4.2 Perceptrons multicouches et réseaux de neurones convolutionnels

Visualiser l'architecture des réseaux de neurones est primordial pour les rendre plus transparents et donc aider à leur interprétabilité. Les figures présentes dans la section 2.2, par exemple, servent à les rendre plus transparents. Tensorboard, présent dans

FIGURE 2.15 – *Visualisation de quatre mots dans l'espace de représentation du plongement lexical de deux réseaux différents : FastText* [Mik+17] *et LSTM* [HS97]. *On observe ici les plus proches voisins des mots sélectionnés et donc les similarités et différences entre les voisins de ces mots dans les plongements lexicaux. On observe par exemple qu'un plus grand nombre de voisins est identique entre les deux plongements pour le mot "word" (61%) que pour le mot "work" (49%). À gauche, sont présents les paramètres utilisés pour cette visualisation, à savoir une projection à l'aide de la t-SNE, où les 50 plus proches voisins de chaque mot sont considérés à l'aide d'une similarité cosinus* [BCS19].

la bibliothèque logicielle appliquée aux réseaux de neurones Tensorflow [Mar+15], permet de voir les matrices de données traverser les différentes couches et de connaître leurs dimensions. Toutes ces informations représentent comment évoluent les données au sein d'un réseau de neurones.

En ce qui concerne les MLP, la totalité des méthodes de visualisation d'architecture de réseaux de neurones permettent de les visualiser (Vi.1). Les outils comme Netron [2], Keras Sequential ASCII [3], Netscope CNN Analyzer [4], présentent des stratégies très similaires à Tensorboard. Les travaux de HARLEY [Har15] et SMILKOV et al. [Smi+17] (voir TensorFlow Playground [5] présenté dans la figure 2.16) présentés dans la section 2.3.1 permettent aussi de visualiser les architectures des réseaux de neurones simples dans une démarche pédagogique. En effet, les perceptrons multicouches étant les réseaux avec l'architecture la plus simple, c'est en général les premiers qui sont utilisés par les novices en apprentissage profond. DAX, introduit par ALBINI et al. [Alb+20], représente, pour les MLP et les CNN, l'influence des tokens, et à travers quels filtres ou neurones leur influence a été augmentée (qu'elle soit dans le sens de la prédiction ou non). DAX montre également quels sont les tokens qui activent le plus les filtres ou neurones concernés à l'aide de nuages de mots (voir Fig 2.17).

La plupart des méthodes présentées en amont sont applicables aux CNN. Elles

2. https://github.com/lutzroeder/Netron
3. https://github.com/stared/keras-sequential-ascii/
4. http://dgschwend.github.io/netscope/quickstart.html
5. http://playground.tensorflow.org/

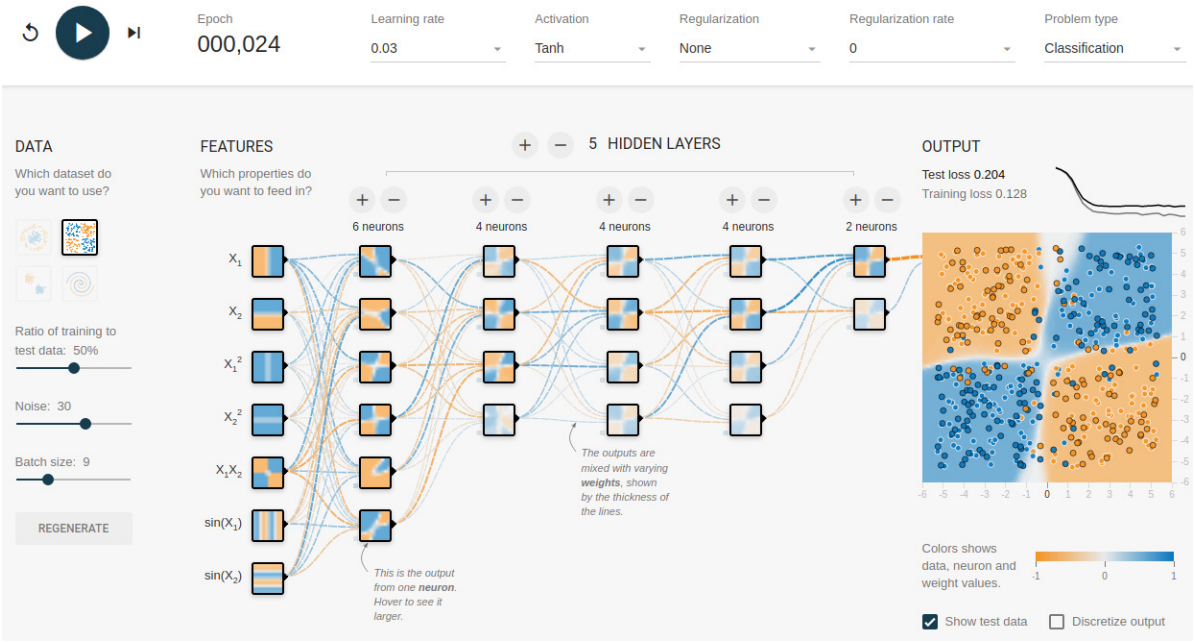


FIGURE 2.16 – *TensorFlow Playground introduit par SMILKOV et al. [Smi+17] sert à comprendre et tester les concepts relatifs à l'apprentissage profond à l'aide de perceptrons multicouches. Le nombre de couches, de neurones, les paramètres d'apprentissage et les données d'entrée sont paramétrables.*

sont d'ailleurs la plupart du temps développées pour être appliquées aux CNN plutôt qu'aux perceptrons multicouches au vu des plus grandes performances des CNN.

2.4.3 Réseaux de neurones récurrents

Les RNN ont été pendant longtemps les réseaux incontournables en TAL. Depuis l'avènement des modèles auto-attentifs, ils sont un peu moins utilisés mais ont toujours des performances comparables sur certaines tâches [Kar+19]. On peut d'ailleurs constater, à l'aide des figures 2.18 et 2.20, qu'après l'arrivée des modèles auto-attentifs, les recherches en visualisation de données appliquées aux réseaux de neurones se sont détournées des RNN au profit des modèles auto-attentifs. Les différents travaux en visualisation de données à propos des RNN concernent parfois seulement certaines architectures (GRU ou LSTM). Dans cette partie, nous traitons donc des outils de visualisation appliqués à tous les RNN et précisons les architectures concernées.

Il n'y a pas de travaux spécifiques sur **la visualisation de l'architecture** des RNN (Vi.1) mais Tensorboard [Mar+15] construit la visualisation d'une cellule de type RNN et les processus la composant. Du fait de leur caractéristique de récurrence, ils ne présentent que peu d'intérêt à être visualisés car chaque cellule d'un RNN est identique aux autres (dans le cas d'un réseau unidirectionnel avec une unique couche de RNN). Les **paramètres des RNN** (Vi.2) ne sont eux-aussi que peu visualisés pour la même raison.

En ce qui concerne les **unités de calcul ou couches de neurones** des RNN (Vi.3), KARPATHY, JOHNSON et FEI-FEI [KJF15] observent l'activation de la fonction d'activation (tangente hyperbolique) qui met à jour le nouvel état caché (voir équation 2.7) à l'aide

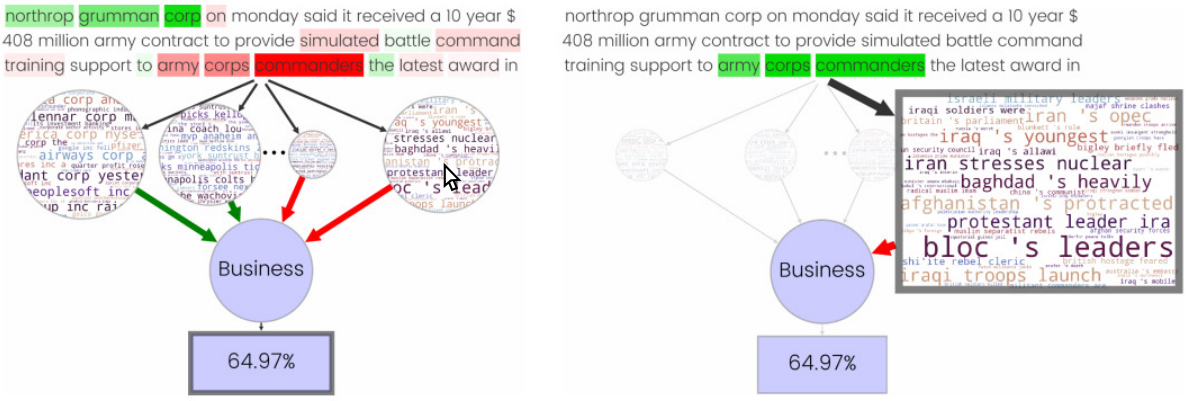


Figure 2.17 – *Visualisation issue de DAX [Alb+20] pour une tâche de classification de textes. À gauche, le graphique DAX pour la classe de sortie (correcte) "Business" (déterminée avec une probabilité de 64,97%) par un CNN pour un texte. On y observe grâce aux couleurs les mots qui ont participé positivement (vert) ou négativement (rouge) à la prédiction. À droite, la vue de DAX après avoir cliqué sur le nuage de mots le plus à droite. On observe le groupe de mots qui active les mêmes filtres de convolutions, à savoir des termes relatifs à la guerre et donc pas au "Business", ce qui explique le lien rouge entre ce nuage de mots et la prédiction "Business".*

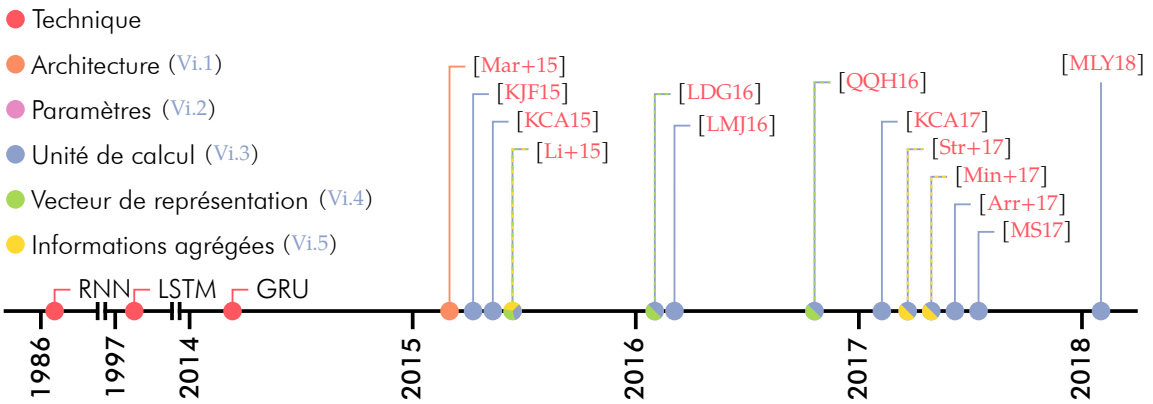


Figure 2.18 – *Apparition des techniques de visualisation appliquées aux RNN ces dernières années.*

du précédant contexte dans un LSTM travaillant sur les caractères plutôt que les tokens. Qian, Qiu et Huang [QQH16] visualisent la même activation tout en inspectant également l'évolution précise de certaines dimensions de l'état caché dans un LSTM traitant sur les tokens et les caractères. Ils mettent ainsi en évidence l'activation de certaines portes dans les cellules en fonction du rôle du mot dans la phrase. Certaines portes sont ainsi sensibles au déterminant quand d'autres le sont aux verbes et à leur préposition. Travailler à l'aide de tokens plutôt que de caractères permet de constater si les catégories lexicales et les fonctions grammaticales, et donc l'information sémantique que peuvent porter les mots, sont capturées par les RNN. C'est notamment ce que montrent les travaux de Kádár, Chrupała et Alishahi [KCA15; KCA17]. En omettant certains mots en entrée, ils observent des différences dans les représentations des données construites. Enfin, Linzen, Dupoux et Goldberg [LDG16] montrent que les LSTM capturent des structures grammaticales à l'aide d'apprentissage supervisé. Là encore,

ils se basent sur l'étude des valeurs dans les états cachés de LSTM. LSTMVis [Str+17] et RNNVis [Min+17] analysent les états cachés des RNN à l'aide d'outils interactifs. LSTMVis se concentre sur l'évolution des états cachés dans les LSTM. L'utilisateur peut sélectionner une fenêtre de tokens dans la donnée d'entrée de manière à observer l'évolution de l'état caché dans cette fenêtre. RNNVis (voir Fig 2.19), disponible pour tout type de RNN, les compare entre-eux. Certaines méthodes présentées dans la section 2.3.2 sont spécifiques aux RNN. Elles utilisent des cartes de chaleur pour identifier les mots ayant le plus participés à une prédiction [Li+15; LMJ16; Arr+17; MS17; MLY18] (figure 2.12). Il est important de noter que pour les RNN, les catégories Vi.3 et Vi.4 se confondent car on peut considérer l'état caché h_t comme une représentation de la donnée d'entrée au token t et non comme une unité de calcul.

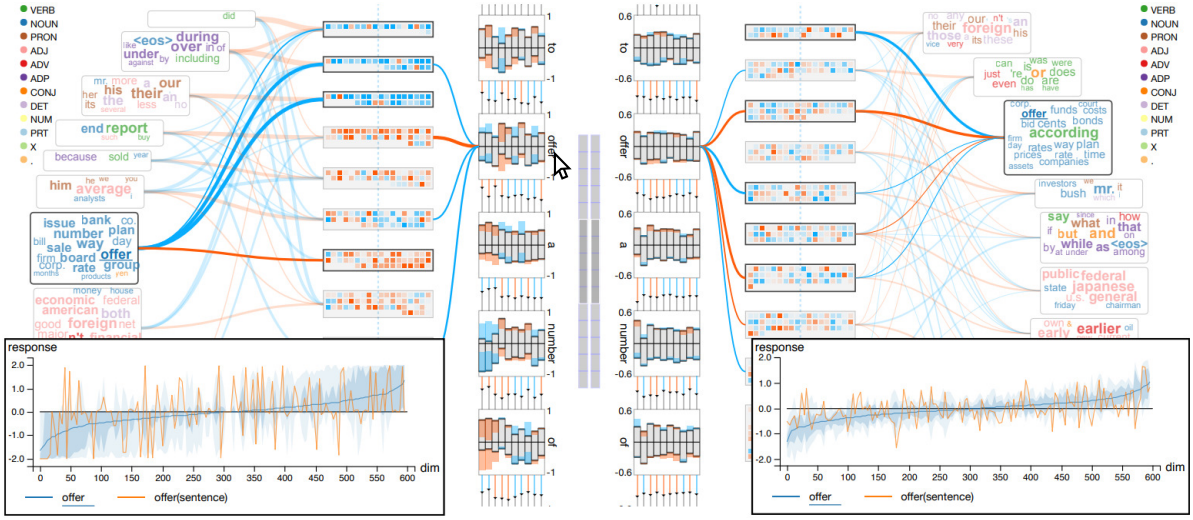


FIGURE 2.19 – *Comparaison d'un RNN classique (à droite) et d'un LSTM (à gauche) dans RNNVis [Min+17]. Les états cachés sont présentés dans le haut de la visualisation. En bas est présenté comment les différentes dimensions des représentations du texte d'entrée sont influencées par le mot "offer" pour chacun des réseaux.*

En ce qui concerne les **vecteurs de représentation des données dans des espaces à grandes dimensions** (Vi.4), QIAN, QIU et HUANG [QQH16] visualisent à l'aide d'une ACP l'espace de représentation des mots construit par un LSTM et encodent à l'aide de la taille des mots, l'activation résultant de leur traitement pour certaines dimensions de portes du LSTM. LINZEN, DUPOUX et GOLDBERG [LDG16] montrent que les représentations des mots construites par un LSTM sont capables de connaître le caractère pluriel ou singulier d'un mot alors que cette information était omise pendant l'entraînement (sans le "s" pour l'anglais pas exemple) à l'aide d'une ACP de quelques mots sélectionnés. LI et al. [Li+15] utilisent également une ACP dans l'espace de représentation des groupes de tokens pour montrer l'influence des comparatifs et des superlatifs sur la représentation d'un groupe de tokens issu d'un RNN.

Pour visualiser **des informations agrégées issues ou non du fonctionnement du modèle** (Vi.5), les outils LSTMVis [Str+17] et RNNVis [Min+17], comme de nombreux autres travaux, analysent des POS (ou étiquetage morpho-syntaxique) en plus des états cachés, par exemple en complément de la visualisation des états

cachés. L'étiquetage morpho-syntaxique (POS) regroupe les mots en fonction de leurs propriétés grammaticales. Au niveau le plus élevé, les parties du discours sont le nom, le verbe, ou le déterminant par exemple. D'autres travaux se concentrent sur l'apport des tokens dans une représentation de texte. Li et al. [Li+15] visualisent l'intensification et la négation, en inspectant les états cachés des RNN à différents instants t.

2.4.4 Réseaux de neurones auto-attentifs

Les réseaux de neurones auto-attentifs ou modèles auto-attentifs ou transformeurs ont supplantés les RNN dans la plupart des tâches en TAL. On peut observer l'intérêt immédiat des chercheurs en apprentissage profond et en visualisation de données pour ces réseaux notamment dès la parution des travaux sur BERT [Dev+18] (figure 2.20).

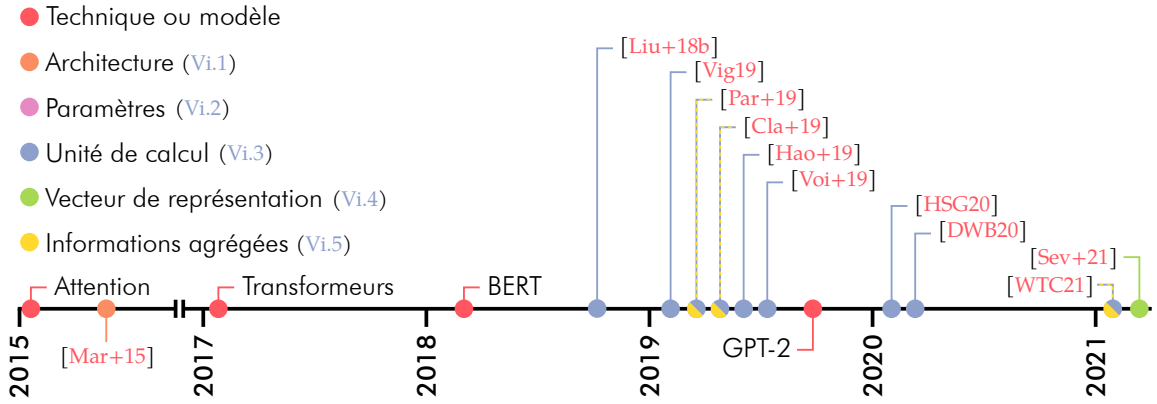


Figure 2.20 – *Apparition des techniques de visualisation appliquées aux modèles auto-attentifs depuis 2015.*

Tout comme pour les RNN, il n'y a que peu de travaux concernant **la visualisation de l'architecture** des transformeurs (Vi.1) ou les **paramètres** (Vi.2) de ceux-ci. La principale raison est sûrement la multiplication des couches (BERT en a par exemple 12 pour l'attention et 12 pour les couches à propagation directe). Néanmoins, comme pour les RNN, Tensorboard [Mar+15] visualise l'architecture de BERT. L'analyse des matrices d'attention qui peuvent être vues comme des **unités de calcul ou des couches de neurones** (Vi.3), revêt un grand intérêt. Les transformeurs contiennent de nombreuses couches d'attention présentes à différentes étapes du traitement de la donnée d'entrée. De plus, chacune de ces couches contient plusieurs têtes d'attention. Ainsi, on dénombre $k \times j$ matrices d'attention où k est le nombre de têtes d'attention par couche et j le nombre de couches d'attention. Nlize [Liu+18b] visualise des matrices d'attention pour une tâche d'inférence en langage naturel. Les cases de ces matrices sont colorées avec un encodage séquentiel des couleurs pour identifier les liens faibles ou forts d'attention entre deux tokens en fonction de la couleur foncée ou claire. Nlize visualise aussi à l'aide de bump chart des informations similaires. Les bump chart [Liu+18b ; Vig19 ; Par+19 ; Cla+19] et les matrices [Liu+18b ; Par+19] sont très largement utilisés dans l'analyse de l'attention issue de transformeurs. Les bump charts sont des graphiques utilisés pour comparer deux dimensions, l'une par rapport à l'autre, en utilisant une valeur de mesure comme l'encodage d'une arête entre les valeurs des

dimensions. Dans la cas de l'attention, les bump chart comparent un groupe de mots à un autre groupe de mots (le même le plus souvent) avec comme liens entre les deux le score d'attention entre les mots. Le plus souvent, plus ce score est grand, plus l'arête est opaque ou foncée.

VIG [Vig19] propose BERTViz, une visualisation de l'attention dans chacune des têtes pour chacune des couches dans le réseau de neurones BERT, ainsi qu'une visualisation des représentations des tokens au sein et à la sortie des cellules d'attention (pour la requête et la clef de celles-ci, voir section 2.2.2.4). PARK et al. [Par+19] proposent SANVis, un outil complet de visualisation de l'attention, couche par couche, avec des moyennes sur la valeur d'attention par tête à chaque couche. SANVis utilise, ce que PARK et al. [Par+19] ont appelé une HeadLens. Elle permet d'avoir des informations sur le fonctionnement d'une tête d'attention précise en termes de types de mots sur lesquels elle se concentre. CLARK et al. [Cla+19] visualisent dans un espace à deux dimensions les cellules d'attention de BERT et observent les similarités de fonctionnement de celles-ci. HAO et al. [Hao+19] s'intéressent plutôt à l'entraînement de BERT. Ils utilisent des visualisations de la surface d'erreur (error surface) lors de l'entraînement. Ils démontrent ainsi que grâce au pré-entraînement et au réglage fin (fine-tuning) de BERT, celui-ci est robuste au sur-apprentissage. Ils démontrent aussi plus globalement l'efficacité d'utiliser BERT pré-entraîné. Certains de ces outils ne sont pas spécifiques à BERT. Un exemple est exBERT [HSG20] qui applique cette même méthode à GPT-2. WANG, TURKO et CHAU [WTC21] visualisent l'attention en adaptant les bump charts avec une disposition radiale. Ils visualisent aussi les différentes têtes au sein des couches pour montrer leur importance dans la tâche mais aussi la manière dont elles capturent des règles syntaxiques et sémantiques. L'analyse de l'attention ou de l'importance des cellules d'attention peut aussi se faire de manière classique. VOITA et al. [Voi+19] utilisent par exemple la LRP pour l'analyse de l'importance des têtes d'attention dans une tâche de traduction. DEROSE, WANG et BERGER [DWB20] proposent AttentionFlows, un outil qui explore la façon dont l'attention des modèles auto-attentifs est affinée pendant le réglage fin, et comment l'auto-attention informe les décisions de classification. Pour cela, ils visualisent 12 couches d'attention pour chacun des mots à l'aide d'une visualisation radiale.

Enfin, pour **visualiser les vecteurs de représentation des données** (Vi.4) la plupart des méthodes s'appuient sur des travaux déjà effectués pour le plongement lexical classique ou la visualisation de données classiques. LMExplorer [Sev+21] visualise les mots dans une réduction en deux dimensions de l'espace de représentation construit par BERT à chaque sortie des couches d'attention. Certains travaux visualisent les matrices d'attention et d'autres informations. Ces travaux entrent aussi dans la catégorie des outil de visualisation avec des **informations agrégées** (Vi.5) de HOHMAN et al. [Hoh+19].

2.4.5 Méthodes agnostiques

Dans cette section, nous présentons tout d'abord quelques travaux ne relevant pas des catégories précédentes, avant de nous intéresser plus spécifiquement aux travaux concernant la contribution à la prédiction dans la section 2.4.5.1 et la visualisation de

la frontière de décision dans la section 2.4.5.2.

En TAL, il est souvent nécessaire d'utiliser des réseaux AE de manière à pouvoir encoder dans un espace de plus petite dimension un long texte. Seq2Seq-Vis [Str+18] visualise les scores d'attention entre un mot en entrée et un mot en sortie et affiche les espaces de représentation des mots mais aussi les prédictions les plus probables à chaque instant t. L'objectif d'un tel outil est d'identifier ce qui a été appris par les réseaux AE et de détecter les éventuelles erreurs dans la procédure de décodage de ceux-ci pour ensuite les déboguer ou les corriger.

2.4.5.1 Contribution à la prédiction

Les méthodes agnostiques permettant de visualiser les parties de la donnée d'entrée qui contribuent à la prédiction ont été présentées dans la section 2.3.2.4. Tout comme la visualisation de l'attention dans les modèles auto-attentifs (section 2.4.4) ou la visualisation des activations des portes dans les RNN (section 2.4.3), la visualisation de la participation des tokens à la prédiction se fait majoritairement à l'aide de cartes de chaleur [Li+15; LMJ16; Arr+17] (figure 2.21). Dans le cadre de l'utilisation des valeurs de Shapley, de SHAP [LL17; LEL18] ou de LIME [RSG16], des diagrammes en bâtons sont également utilisés (figure 2.22).

true	predicted	N°	Notation: -- very negative, - negative, 0 neutral, + positive, ++ very positive
		1.	do n't waste your money .
		2.	neither funny nor suspenseful nor particularly well-drawn .
		3.	it 's not horrible , just horribly mediocre .
		4.	... too slow , too boring , and occasionally annoying .
		5.	it 's neither as romantic nor as thrilling as it should be .
	--	6.	the master of disaster - it 's a piece of dreck disguised as comedy .
		7.	so stupid , so ill-conceived , so badly drawn , it created whole new levels of ugly .
		8.	a film so tedious that it is impossible to care whether that boast is true or not .
		9.	choppy editing and too many repetitive scenes spoil what could have been an important documentary about stand-up comedy .
--		10.	this idea has lost its originality ... and neither star appears very excited at rehashing what was basically a one-joke picture .
	++	11.	ecks this one off your must-see list .
	-	12.	this is n't a `` friday '' worth waiting for .
	-	13.	there is not an ounce of honesty in the entire production .
	-	14.	do n't expect any surprises in this checklist of teamwork cliches ...
	-	15.	he has not learnt that storytelling is what the movies are about .
	-	16.	but here 's the real damn : it is n't funny , either .
	+	17.	these are names to remember , in order to avoid them in the future .
	-	18.	the cartoon that is n't really good enough to be on afternoon tv is now a movie that is n't really good enough to be in theaters .
	++	19.	a worthy entry into a very difficult genre .
++		20.	it 's a good film -- not a classic , but odd , entertaining and authentic .
	--	21.	it never fails to engage us .

Figure 2.21 – *Cartes de chaleur issues de la méthode d'explication LRP de textes. La tâche effectuée est une analyse de sentiments. L'influence positive est représentée en rouge, la négative en bleu, et l'intensité de la couleur est normalisée par rapport à l'influence absolue maximale par phrase. La classe de la phrase, et la classe prédite par le classifieur, sont indiquées à gauche* [*Arr+17*].

Certains travaux cherchent à montrer l'activation précise de certaines variables des vecteurs de représentation des tokens [Li+15]. Or, ne pouvant pas extraire l'information à propos de ce qu'est censé présenter une dimension, cela ne revêt que peu d'intérêt. Enfin, les explications hiérarchiques sont présentées sous forme d'arbre de décision montrant comment un groupe de tokens influe sur la prédiction à un instant t [SMY19;

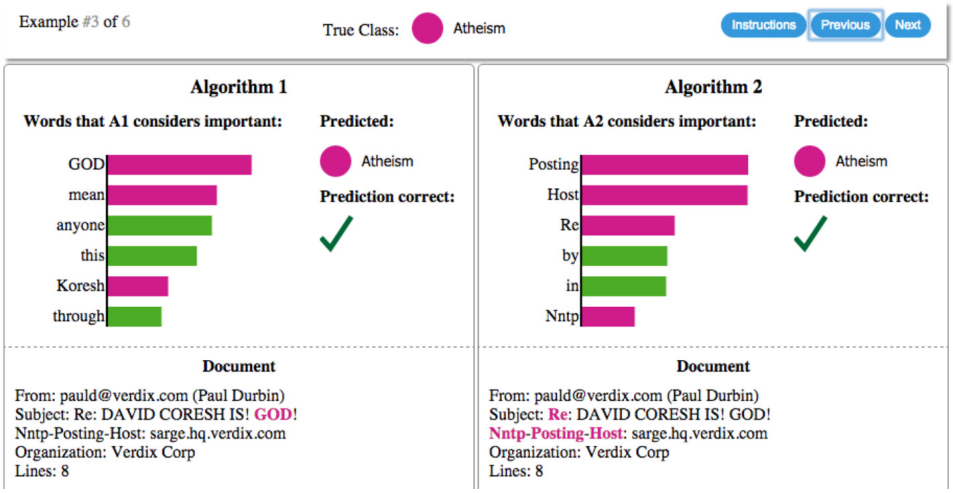


FIGURE 2.22 – *Explication des prédictions individuelles de classifieurs de substitution (plus simple que le modèle à expliquer et se substituant à lui) essayant de déterminer si un document porte sur le christianisme ou l'athéisme. Les diagrammes en bâtons représentent l'importance donnée aux mots les plus pertinents, également mis en évidence dans le texte. La couleur indique la classe à laquelle le mot contribue (vert pour "christianisme", magenta pour "athéisme")* [RSG16].

Jin+19 ; CZJ20] (figure 2.23) ou comment un modèle de substitution traite l'information à l'aide d'un arbre de décision classique [Gui+18a].

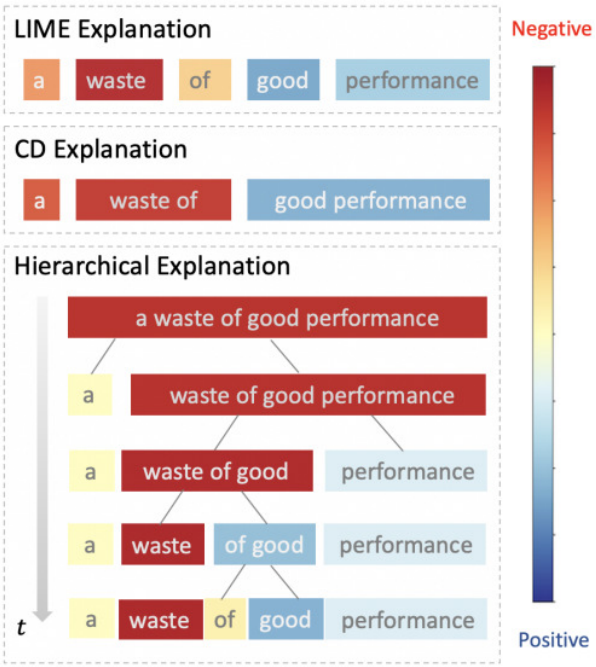


FIGURE 2.23 – *Différentes explications pour une critique de film négative. La couleur des groupes de tokens représente la contribution des tokens correspondant à la prédiction du modèle. À partir de l'explication hiérarchique, nous obtenons pour chaque groupe, à chaque instant t, sa contribution, dont la plus importante est pour le groupe "waste of good"* [CZJ20].

2.4.5.2 Visualisation de la frontière de décision

La frontière de décision d'un classifieur dichotomique est l'hyperplan dans l'espace de représentation des données à partir duquel le classifieur prédit un label ou un autre. Plus une donnée est proche de la frontière, plus sa prédiction est d'une faible confiance. La visualisation de la frontière peut-être soit une méthode agnostique dépendant des processus utilisés soit une méthode spécifique à un type d'architecture. Elle est dans le domaine de la classification à l'aide de réseaux de neurones intéressante du fait de la grande dimension dans laquelle les classifieurs représentent leurs données. La visualisation de la frontière doit donc adapter ces espaces de représentation en deux ou trois dimensions. Migut, Worring et Veenman [MWV15] visualisent la frontière de décision dans l'espace d'entrée (ou espace d'origine), mais les distances à la frontière ne sont pas significatives. Avoir des distances à la frontière significatives permet de comparer les données entre elles en termes de confiance ou force des prédictions. C'est un point essentiel à l'interprétabilité des réseaux de neurones. Zhiyong et Congfu [ZC08] proposent un algorithme pour trouver les données de la frontière de décision, mais les distances utilisées dans leur visualisation ne sont à nouveau pas assez significatives. Zhang et al. [Zha+19] intègrent des distances significatives pour comparer deux classifieurs mais toute autre information sur les voisins ou le contenu des données est perdue. Rodrigues et al. [Rod+19] visualisent la frontière de décision des classifieurs en utilisant des techniques de réduction de dimension dans l'espace d'entrée des classifieurs. Parmi les cinq techniques les plus efficaces qu'ils identifient pour visualiser la frontière de décision des CNN dans la classification binaire, on retrouve UMAP et t-SNE présentées dans la section 2.4.1. Cependant, ils utilisent un ensemble de données linéairement séparables, ce qui n'est pas un ensemble de données réaliste. De plus, leur méthode ne sert pas à l'exploration de différentes parties dans cet espace mais juste de l'espace comme un tout.

Il est souvent intéressant dans le cadre de la visualisation d'un espace de représentation de ne s'intéresser qu'à certaines parties. Mikolov et al. [Mik+13b] choisissent, par exemple, de ne s'intéresser qu'à quelques mots lorsqu'ils visualisent le lien sémantique entre pays et ville (figure 2.11). La partie de l'espace sélectionnée n'a pas forcément vocation à être arbitraire mais souvent pour les grands jeux de données, visualiser tout l'espace revêt un intérêt moindre aussi bien en termes de lisibilité que d'information à extraire. Dans leur démonstration disponible en ligne [6] de EmbeddingProjector, Martín Abadi et al. [Mar+15] proposent des sous-ensembles de l'espace de représentation des mots à visualiser. Melnik [Mel02] analyse la connectivité des données dans l'espace d'entrée. Cela garantit qu'aucune frontière de décision n'existe entre deux données si elles appartiennent aux mêmes régions de décision. Ces régions de décision sont des zones de l'espace de représentation dans lesquelles les prédictions sont toutes identiques. Différentes régions de décision sont calculées et comparées à travers différents classifieurs tels que des classifieurs neuronaux ou des SVM. Ramamurthy, Varshney et Mody [RVM19] proposent une méthodologie pour comparer, pour différents modèles, la complexité de la frontière de décision et donc la capacité de généralisation de ces modèles pour un ensemble de données. Enfin, Ma et Maciejewski [MM21] visualisent la décision relative à la frontière de décision en utilisant le SVM sur les données proches

6. urlhttps ://projector.tensorflow.org/

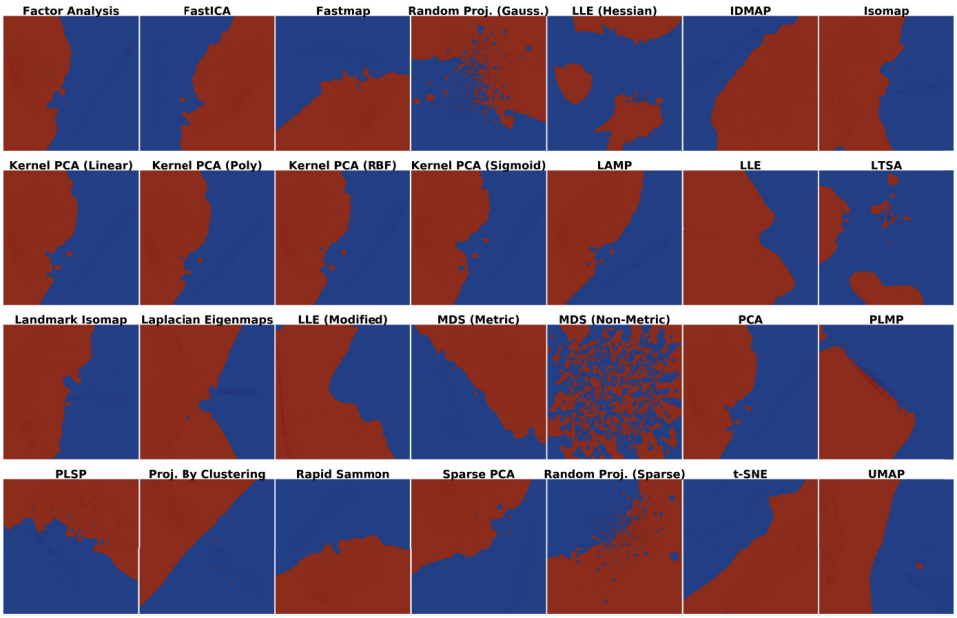


FIGURE 2.24 – *Utilisation de techniques de réduction de dimensions pour visualiser la frontière de décision d'un réseaux de neurones depuis l'espace d'entrée des données sur un jeu de données linéairement séparable proposée par* [Rod+19]. *Leurs travaux montrent que la t-SNE et UMAP sont les techniques les plus efficaces pour visualiser la frontière comme une ligne. Cette frontière de décision est la bordure séparant les deux couleurs. Chacune d'entre elles correspondant à une classe.*

de la frontière de décision. Ils construisent plusieurs segments linéaires de la frontière de décision avec des mises en lumière de certaines parties de la frontière de décision.

La plus grande différence entre tous ces travaux et celui que nous présentons dans le chapitre 3 est que nous construisons la frontière de décision dans l'espace construit par le réseau de neurones. Cet espace est plus intéressant pour explorer les données traitées par les réseaux de neurones et construire des parties de l'espace de représentation plus significatives. De plus, nous pouvons calculer la distance réelle à la frontière de décision et explorer ces parties significatives de l'espace de représentation (ces parties seront appelées "localités", voir section 3.1).

2.5 Conclusions

En résumé, dans ce chapitre, nous avons présenté dans la section 2.2 la tâche de classification de textes à l'aide de réseaux de neurones, dans la section 2.3 les problématiques d'interprétabilité et dans la section 2.4 l'apport des techniques de visualisation de données combinées aux réseaux de neurones.

Pour commencer, dans la section 2.2, avons montré la multitude d'architectures disponibles en apprentissage profond, en nous focalisant sur les architectures fondatrices. Pour ce faire, nous avons produit différentes figures originales pour les comparer mais aussi afin de comprendre la problématique inhérente aux réseaux de neurones : leur non-transparence. Nous avons également montré la profusion de travaux autour de ces architectures à l'aide de frises chronologiques.

Ensuite, dans la section 2.3, nous avons présenté les problématiques d'interprétabilité, de transparence et d'explicabilité en notant les méthodes qui expliquent les prédictions issues des réseaux de neurones en TAL et augmentent leur interprétabilité. Nous avons conçu des figures soulignant les liens entre ces méthodes et leur chronologie d'apparition, ce qui montre notamment le grand intérêt de la communauté pour ces recherches.

Pour finir, dans la section 2.4, nous avons présenté les nombreuses applications des techniques de visualisation de données dans l'utilisation des réseaux de neurones. Ces techniques aident aussi bien pour la transparence des modèles que pour l'explicabilité des prédictions pour chacune des architectures présentées dans la section 2.2. Les méthodes de visualisation de la frontière de décision des classifieurs sont essentielles à leur interprétabilité, ainsi que les méthodes permettant de représenter des espaces à grande dimension. Nos travaux traitant de ces aspects sont présentés dans le chapitre suivant. Ils montrent comment, la visualisation de la frontière de décision dans l'espace construit par le réseau de neurones, accroît l'interprétabilité de ces derniers.

CHAPITRE 3

EBBE-TEXT : EXPLIQUER LES PRÉDICTIONS D'UN CLASSIFIEUR NEURONAL POUR LES DONNÉES TEXTUELLES

Sommaire

3.1 Introduction

Dans ce chapitre, nous présentons EBBE-Text (acronyme issu de l'anglais : Explanations By Boundary Exploration in the Text representation space), un outil de visualisation, d'exploration et d'explication des prédictions des réseaux de neurones en classification dichotomique de données textuelles. EBBE-Text intègre deux visualisations : une première permettant de repérer les localités contenant des données similaires et une seconde pour visualiser la frontière de décision et d'autres informations pour une localité choisie. Pour cette visualisation de la frontière de décision, nous proposons une nouvelle méthode, basée sur la création d'un graphe de proximité des données réelles.

Dans la section 3.1, nous introduisons globalement EBBE-Text. Dans la section 3.3, nous décrivons les attentes des utilisateurs en termes d'explicabilité des réseaux de neurones, en identifiant différentes questions. Puis, nous transposons ces questions sous forme de besoins pour l'outil. Dans la section 3.4, nous développons les manipulations de données nécessaires pour pouvoir visualiser les localités et la frontière de décision. Dans la section 3.5, nous montrons l'utilité des techniques de visualisation que nous avons sélectionnées pour répondre aux besoins identifiés dans la section 3.3. Dans la section 3.6, nous évaluons notre méthode à travers des questionnaires et présentons trois études de cas. Enfin, dans la section 3.7, nous discutons des modifications, améliorations et perspectives pour EBBE-Text et plus globalement pour la visualisation de la frontière de décision et concluons sur l'apport de notre méthode.

3.2 Contexte

La classification de textes est la tâche fondatrice du traitement automatique de la langue (TAL). Elle vise à attribuer des étiquettes aux textes en fonction de leur contenu. Les contenus peuvent être abstraits par de nombreuses techniques, mais l'essor de l'apprentissage par représentation des mots a permis aux chercheurs d'utiliser diverses techniques d'apprentissage profond comme les réseaux de neurones récurrents (RNN) et les modèles auto-attentifs pour abstraire les textes (section 2.2). Ces représentations abstraites des textes ne sont pas interprétables directement par les utilisateurs, qui ont toujours du mal à comprendre les décisions du modèle. Ce manque de compréhension constitue une barrière à l'interprétation et n'encourage pas la confiance des utilisateurs. Par conséquent, certains n'utilisent pas les modèles modernes de TAL dans les domaines à risque ou à fort enjeu. La visualisation des localités et de la frontière de décision est une solution pour une plus grande confiance dans les modèles de TAL pour la classification automatique de textes.

Comme présenté dans les sections 2.3 et 2.4, les techniques de visualisation de données servent tout au long de l'interprétation et de l'explication des réseaux de neurones [Hoh+19] et pendant leur débogage. EBBE-Text se concentre sur les explications des prédictions selon les trois dernières catégories d'explications proposées dans la classification de Lipton [Lip18] (Ex.2, Ex.3, Ex.4, voir section 2.3.2).

Dans cette première contribution, nous concentrons nos travaux sur la classification

binaire. Les distances des données à la frontière de décision montrent à quel point un réseau de neurones est certain de sa prédiction. La visualisation des données positionnées autour de la frontière de décision permet aux utilisateurs de voir s'ils auraient classé les données comme le classifieur l'a fait, par exemple, avec le même ordonnancement, de la donnée la plus proche à la donnée plus éloignée de la frontière de décision. Les données sont comparées entre elles, ce qui permet de construire des explications concernant le fonctionnement du réseau. La confiance de l'utilisateur peut être encouragée par le sentiment que les modèles en TAL adoptent un comportement semblable à celui des humains et donc par la multiplication des explications. Pour visualiser la frontière de décision, EBBE-Text construit au préalable des **localités** dans l'espace de représentation. Nous définissons ces **localités** comme des zones de l'espace de représentation des données dans lesquelles les données sont proches les unes des autres. Nous construisons ces localités pour permettre à l'utilisateur de se concentrer sur des structures textuelles précises. Ces localités se construisent en fonction du sens, de la présence, du nombre et de la place des tokens. Elles simplifient également l'exploration de l'espace de représentation des données. Les explications à l'échelle des localités sont importantes car elles permettent une comparaison fine des données similaires. Comme RIBEIRO, SINGH et GUESTRIN [RSG16], nous proposons une explication locale. Mais plutôt que de produire un modèle local pour expliquer la prédiction, nous comparons localement les données à l'intérieur d'une localité. La visualisation de la frontière de décision est ensuite proposée pour chacune de ces localités. Nous définissons une explication locale comme une explication faite dans une **localité** spécifique.

La principale contribution d'EBBE-Text est une exploration visuelle multi-échelle de l'espace de représentation des textes avec la frontière de décision du classifieur. La première échelle est une vue globale du corpus de données à travers les différentes localités qui le composent. Pour chaque localité, on affiche la distribution des données autour de la frontière de décision. La deuxième échelle est une vue détaillée montrant les localités sur demande. Cette vue détaillée est composée d'autres informations complétant la visualisation de la frontière de décision (telles que la liste des mots les plus pertinents, les visualisations classiques de l'espace de représentation en deux dimensions comme UMAP, t-SNE et l'ACP, les scores d'attention des mots et les matrices de confusion) pour donner plus d'informations sur les données. Les scores d'attention (section 2.2.2.4) sont les composants d'un réseau de neurones qui gèrent et quantifient l'interdépendance entre les éléments d'entrée et de sortie et entre les éléments d'entrée. Ces caractéristiques visuelles sont reliées par diverses fonctionnalités interactives permettant d'explorer les prédictions du réseau à partir de perspectives différentes et complémentaires.

3.3 Caractérisation du problème

Pour produire des explications post-hoc des prédictions, nous identifions certains besoins. Les utilisateurs doivent explorer l'espace de représentation composé de différentes localités. Choisir une localité permet d'inspecter les textes proches les uns des autres ainsi que leur position par rapport à la frontière de décision. L'utilisateur comprend la qualité de la prédiction et la manière dont la transformation des données

affecte les prédictions du modèle, en produisant des explications. Prenons l'exemple de la comparaison, sur une tâche d'analyse de sentiments, de deux textes proches et de part et d'autre de la frontière : l'un prédit comme positif et l'autre prédit comme négatif. Cela aide l'utilisateur à comprendre comment le modèle fait ses prédictions et l'incertitude associée à chacune de ses prédictions.

3.3.1 Questions des utilisateurs pour l'explicabilité

Selon les besoins des utilisateurs de réseaux de neurones précédemment mentionnés, nous identifions six questions pour lesquelles une réponse aide les utilisateurs dans l'interprétation des processus dont sont issues les prédictions :

Qu.1 Pour un texte donné, sa prédiction est-elle correcte et quelle est l'incertitude associée ?

Qu.2 Pour un texte donné, dans une localité donnée, existe-t-il des textes similaires ou légèrement différents dans cette localité et comment sont-ils classés ?

Qu.3 Existe-t-il des différences entre les prédictions dans le corpus entier et dans les localités et quelles sont les localités les plus intéressantes à inspecter ?

Qu.4 Quels mots caractérisent une localité et les textes contenant ces mots sont-ils classés différemment dans cette localité que dans le corpus entier ?

Qu.5 Quels mots ou co-occurrences de mots ont le plus d'influence sur les prédictions ?

Qu.6 Le réseau fonctionne-t-il comme prévu et existe-t-il des anomalies dans les données étiquetées ?

Les questions Qu.1, Qu.2 et Qu.5 sont directement liées à la production d'explications post-hoc. Y répondre permet aux utilisateurs de comparer les prédictions et parfois d'identifier des différences et/ou des similitudes entre elles. Les questions Qu.3, Qu.4 et Qu.5 présentent dans quelle mesure une explication locale peut être généralisée à toutes les localités. Enfin, Qu.6 est la conclusion du fonctionnement du réseau de neurones suite à des explications post-hoc et/ou des observations sur les prédictions. Cette question met aussi en lumière des anomalies dans les données.

3.3.2 Besoins identifiés pour l'outil

À partir des questions de la section 3.3.1, nous identifions sept besoins pour notre outil :

Be.1 Visualiser les différentes localités dans l'espace de représentation et identifier l'efficacité du classifieur dans chacune, pour Qu.3 et Qu.6 ;

Be.2 Visualiser la frontière de décision et les distances par rapport à celle-ci, pour Qu.1, Qu.2, Qu.3, Qu.5 et Qu.6 :

Be.3 Visualiser le voisinage du texte choisi et des chemins vers la frontière de décision (avec le contenu textuel associé), pour Qu.2, Qu.4 et Qu.5 ;

Be.4 Calculer des métriques de score de prédiction et des matrices de confusion, pour Qu.3, Qu.5 et Qu.6 ;

Be.5 Extraire des mots pertinents dans les localités et les textes, localiser ces mots dans la visualisation de la frontière de décision et inspecter leur influence sur les prédictions, pour Qu.4, Qu.5 et Qu.6 ;

Be.6 Localiser le texte choisi dans les espaces de représentation construits par différentes méthodes de réduction de dimension et visualiser sa place autour de la frontière de décision, pour Qu.2 ;

Be.7 Calculer le score de prédiction d'un texte entré par l'utilisateur, pour Qu.2, Qu.5 et Qu.6.

3.4 Manipulation des données

Dans cette section, nous présentons les manipulations des données nécessaires pour la visualisation de l'espace de représentation à travers différentes localités et pour la visualisation de la frontière de décision : (1) encodage des données à l'aide d'un réseau de neurones (section 3.4.1), (2) création de données de frontière qui sont des projections des données réelles sur la frontière de décision du réseau de neurones (section 3.4.2), (3) création du graphe de proximité contenant à la fois les données réelles et les données projetées (section 3.4.3), (4) division en plusieurs composantes connexes du graphe de proximité en effaçant les arêtes les plus faibles (section 3.4.4), (5) première connexion entre les données réelles et leurs projections dans le graphe de proximité (section 3.4.5), (6) division des grandes composantes pour créer des graphes de proximité (section 3.4.6), (7) connexion finale entre les données réelles et les projections (section 3.4.7) et (8) simplification de la frontière de décision (section 3.4.8). Les étapes (5) et (6) forment l'optimisation qui itère jusqu'à ce qu'une sortie réussie soit fournie (section 3.4.5 et 3.4.6). Ces étapes de manipulation des données garantissent principalement qu'EBBE-Text est capable de proposer différentes localités significatives et interprétables (Be.1). Pour ce faire, les localités ont besoin de données de frontière, de données d'entrée et d'une taille raisonnable. La figure 3.1 donne un aperçu de notre approche. Nous détaillons ces étapes dans les sections suivantes.

3.4.1 Encodage des données à l'aide d'un réseau de neurones

EBBE-Text permet l'exploration des classifications et la visualisation de la frontière de décision pour n'importe quel réseau de neurones à condition qu'il construise un vecteur de représentation pour chaque texte. Les réseaux de neurones prennent usuellement en entrée des représentations de textes via leurs représentations des tokens. Selon le réseau de neurones, les représentations des tokens dans les textes sont utilisées les unes après les autres (*e.g.*, RNN) ou toutes ensembles en une seule fois (*e.g.*, modèles auto-attentifs) pour produire la représentation du texte. C'est l'encodage de textes qui produit le vecteur de représentation de grande dimension. Ce vecteur est ensuite utilisé pour prédire une classe dans la tâche de classification du réseau de neurones. Dans notre méthode, la dernière couche du réseau de neurones est considérée comme

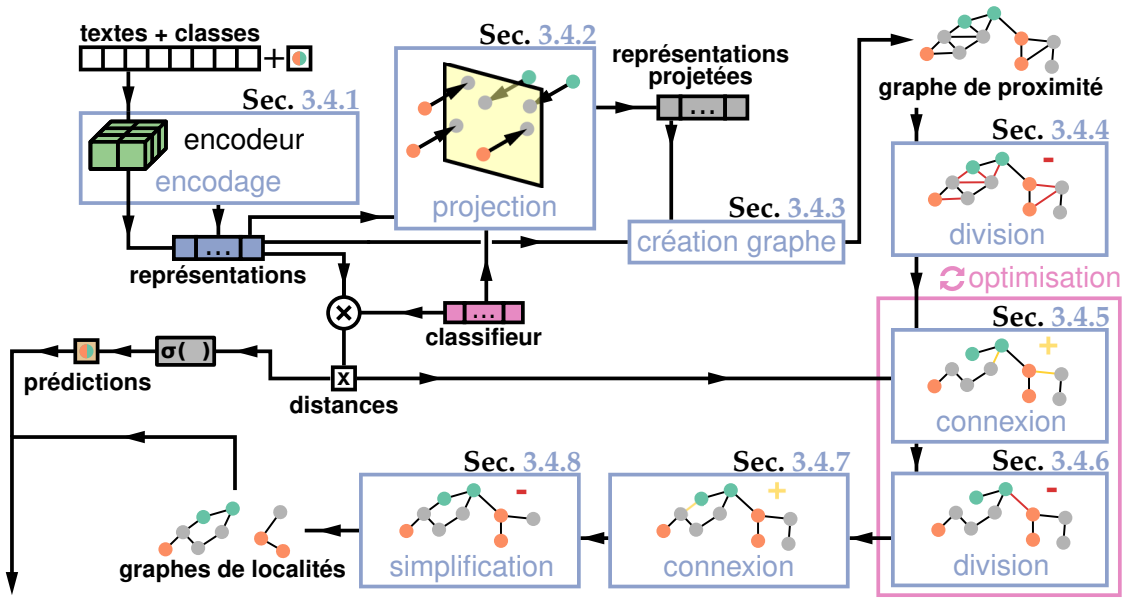


Figure 3.1 – *Étapes de manipulation des données. En* ***bleu****, les différentes étapes décrites dans la section 3.4. Premièrement, chaque texte est encodé par le réseau de neurones (section 3.4.1). Ensuite, chaque vecteur de représentation issu de l'encodage est projeté sur la frontière de décision (section 3.4.2) du réseau de neurones, ce qui signifie que pour chaque entrée, il existe une unique projection sur la frontière de décision. Un graphe de proximité est construit en utilisant la méthode UMAP (section 3.4.3). Ce graphe est ensuite divisé en plusieurs composantes connexes (section 3.4.4). Dans les étapes d'optimisation (****encadré rose****), chaque composante connexe n'ayant pas suffisamment de sommets de frontière est liée à d'autres composantes en fonction des distances entre les données d'entrée et leurs projections (section 3.4.5). Ensuite, les grandes composantes connexes sont divisées si nécessaire (section 3.4.6). Ces deux étapes d'optimisation sont répétées tant qu'il y a des composantes connexes sans suffisamment de sommets de frontière ou de grandes composantes connexes. Enfin, pour chaque composante connexe, les sommets sont liés à leur projection s'ils appartiennent à la même composante connexe (section 3.4.7) et les sommets de frontière inutiles sont supprimés de la composante connexe (section 3.4.8). Pour finir, chaque composante connexe représente une localité. Les localités, les distances à la frontière de décision et les prédictions sont utilisées pour produire les visualisations dans la suite de notre méthode.*

le classifieur et donc les autres couches sont considérées comme l'encodeur. Les sorties de l'encodeur sont les vecteurs utilisés en tant que représentations des textes. Ces vecteurs appartiennent à l'espace de représentation des textes. Nous utilisons dans la section 3.6.3 différents réseaux de neurones : un modèle auto-attentif, un RNN auto-attentif et un RNN AE.

3.4.2 Projection des données réelles sur la frontière

Après l'encodage du texte, pour répondre aux besoins Be.1 et Be.2, nous créons des données situées sur la frontière de décision en utilisant la structure linéaire du classifieur. La frontière de décision dans l'espace de représentation du texte est représentée par un hyperplan. L'orientation de cet hyperplan est contrôlée par un vecteur normal β, (*i.e.*, orthogonal à l'hyperplan) qui définit également les poids du classifieur du réseau de neurones. Chaque encodage de texte, *i.e.*, chaque vecteur z représentant

une donnée peut être décomposé comme la somme d'un vecteur $\boldsymbol{u}$ qui appartient à l'hyperplan (*i.e.*, $\langle \boldsymbol{\beta}, \boldsymbol{u} \rangle = 0$) et d'un vecteur v colinéaire à $\boldsymbol{\beta}$ ou nul. Nous calculons les vecteurs de représentation des données projetées (sur la frontière de décision) en projetant orthogonalement sur l'hyperplan l'ensemble des vecteurs de représentation des données : $\boldsymbol{u} = \text{proj}_{\beta}(\boldsymbol{z}) = \boldsymbol{z} - \langle \boldsymbol{\beta}, \boldsymbol{z} \rangle \boldsymbol{\beta} / \|\boldsymbol{\beta}\|_2^2$. Les données projetées sur la frontière sont aussi nombreuses que les données d'entrée puisque pour chaque donnée d'entrée sa projection est créée.

3.4.3 Création du graphe de proximité

Nous créons le graphe de proximité en utilisant l'algorithme de réduction de dimension UMAP [MHM18]. UMAP est basé sur des variétés mathématiques et est issu de l'analyse de données topologiques. Dans cette section, nous proposons différentes figures pour un espace de deux dimensions. L'algorithme UMAP s'applique aux espaces de grandes dimensions aussi bien en général que dans nos travaux.

La première étape de UMAP consiste à fournir des objets de mathématiques combinatoires appelés simplexes. Géométriquement, un simplexe est un objet à k dimensions appelé k-simplexe. Un k-simplexe est formé en prenant l'enveloppe convexe de $k+1$ points indépendants. Ainsi, un 0-simplexe est un point, un 1-simplexe est un segment de droite (entre deux 0-simplexes), un 2-simplexe est un triangle (avec trois 1-simplexes comme "côtés") et un 3-simplexe est un tétraèdre (avec quatre 2-simplexes comme "faces"). Les différents simplexes présentés sont visualisables en figure 3.2. Une construction aussi simple permet une généralisation facile à des dimensions arbitraires. Par exemple, à partir de la dimension 4, un 4-simplexe est aussi appelé 5-cellules et évolue donc en dimension 4 (difficile à se représenter car nous vivons dans un espace de dimension 3). Ce qu'il faut retenir de la notion de simplexe est qu'un simplexe est une généralisation d'un triangle en dimension k. Il représente donc une partie de l'espace de représentation et peut contenir des points au sein de son enveloppe convexe.

(a) (b) (c) (d)

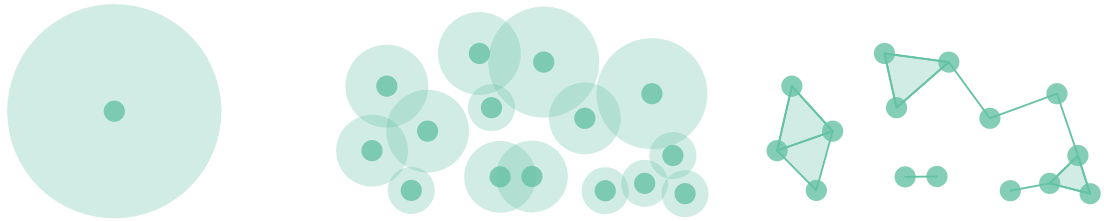
FIGURE 3.2 – *Exemples de k-simplexes. (a) 0-simplexe. (b) 1-simplexe. (c) 2-simplexe. (d) 3-simplexe. Un 3-simplexe est défini par quatre 0-simplexes ou six 1-simplexes ou quatre 2-simplexes. On comprend ainsi que l'on peut définir un k-simplexe à partir d'autres simplexes de dimension inférieure.*

De la même manière que les simplexes peuvent contenir des points, il est possible de définir des simplexes autour de points (des ensembles de simplexes). La figure 3.3 présente visuellement cette étape de création d'ensembles de simplexes appelés ensembles simpliciels. Nous appelons ensemble simpliciel d'un point, l'ensemble qui contient des simplexes ayant pour barycentre le point en question. UMAP construit des ensembles simpliciels pour chacun des points.

(a) (b) (c) (d)

Figure 3.3 – *Création des ensembles simpliciels.* (a) *Un point dans un espace à deux dimensions.* (b) *Un 2-simplexe ou triangle dans un espace à deux dimensions ayant pour barycentre un point.* (c) *Quatre 2-simplexes ou triangles ayant pour barycentre un même point formant un ensemble de cardinalité 4 de simplexes autour d'un point.* (d) *Un ensemble simpliciel de cardinalité 16 d'un point. Les opacités des simplexes des figures* (b-c) *et* (d) *sont différentes à des fins de facilitation de lecture.*

Une fois ces ensembles simpliciels construits pour chaque point dans l'espace de représentation des données, UMAP construit des complexes simpliciels qui résultent de l'union des différents ensembles simpliciels des points. Comme présenté en figure 3.4, les complexes simpliciels sont formés de différents ensembles de simplexes de dimensions différentes.



(a) (b) (c)

Figure 3.4 – *Création des complexes simpliciels.* (a) *Un point dans l'espace de représentation des données avec son ensemble simpliciel.* (b) *Plusieurs points dans l'espace de représentation des données avec, pour chacun d'entre eux, un ensemble simpliciel.* (c) *Des complexes simpliciels construits par l'union des ensembles simpliciels. L'objectif de ces complexes est de représenter une partie, ou dans certains cas l'intégralité, de l'espace de représentation des données. Dans cet exemple, trois complexes simpliciels ont été construits.*

Mathématiquement, le complexe simpliciel capture la topologie fondamentale de l'ensemble de données et représente donc la distribution des données dans leur espace de représentation. La plus grande partie du complexe simpliciel est composée de 0-simplexes et de 1-simplexes. Les simplexes de plus grande dimension peuvent être définis à l'aide de ceux-ci. La structure d'un complexe simpliciel peut être ramenée à celle d'un graphe classique possédant des sommets et des arêtes. En pratique, pour adapter à des données réelles cette notion de complexe simpliciel et produire un graphe de proximité, UMAP construit pour chaque point de données (sommet dans un graphe ou 0-simplexe dans le complexe simpliciel) une sphère (n-sphère où n est la dimension de l'espace de représentation) ouverte floue. Cette sphère représente l'ensemble simpliciel de la donnée (une intuition du caractère flou de la construction de cette sphère est visible en Fig 3.3(d)). Afin de conserver la connexité du graphe, UMAP attribut à chaque sphère un rayon en fonction de la distance entre la donnée et son plus proche voisin. Ainsi, pour chaque donnée, il existe au moins une autre donnée telle que leurs sphères ouvertes floues se superposent. La superposition de sphères

floues est vue comme une probabilité de voisinage. Dans le graphe de proximité de UMAP, on a donc pour chaque donnée un sommet et pour chaque probabilité de voisinage supérieure à zéro, entre deux données, une arête pondérée à cette probabilité. La distance jusqu'à laquelle s'étendent les sphères ouvertes floues dépend du nombre de voisins voulus dans l'algorithme UMAP.

Dans notre cas de classification dichotomique de textes, pour construire un graphe de proximité, nous appliquons la méthode de construction de graphe de proximité de UMAP à notre espace de représentation des textes dans lequel se trouvent aussi les projections (voir Fig 3.5).

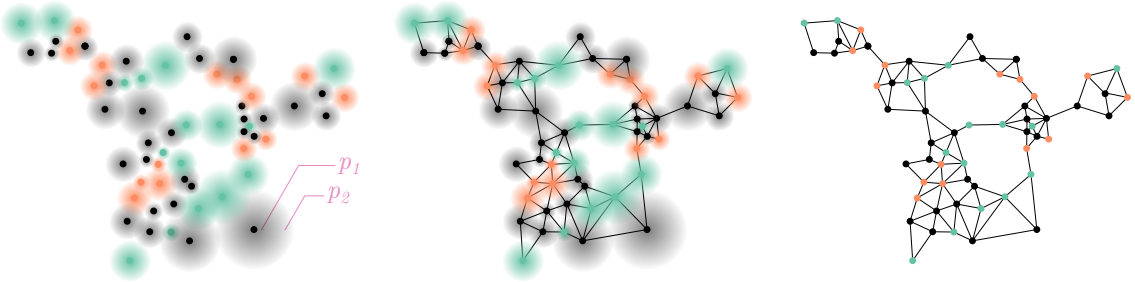


FIGURE 3.5 – *Création du graphe de proximité à partir de l'espace de représentation des textes.* (*a*) *Données* (*représentées par* • *et* • *selon leurs classes dans une tâche de classification dichotomique*) *et données projetées* (*représentées par* •) *avec leurs ensembles simpliciels* (*symbolisés par des cercles fondus*). *Ici,* p_1 *et* p_2 *sont deux probabilités de voisinage avec le sommet le plus proche, pour une donnée placée à la position indiquée par les lignes* ***roses*** ($p_1 > p_2$). (*b*) *Création des liens entre les sommets lorsque les ensembles simpliciels se superposent.* (*c*) *Liens entre les données créées par l'union d'ensembles simpliciels. Ce graphe est le graphe de proximité final issu de la méthode UMAP.*

3.4.4 Division du graphe de proximité sur les faibles connexions

Plus le paramètre du nombre de voisins dans UMAP est élevé, plus la probabilité d'obtenir une seule composante connexe dans le graphe de proximité G est élevée (ou du moins la probabilité d'obtenir quelques grandes composantes connexes). En d'autres termes, il existe un chemin entre pratiquement tous les couple de sommets dans G. Pour construire de plus petites composantes connexes, nous supprimons dans G les arêtes en partant de celles ayant les poids les plus faibles aux plus élevés et appartenant à une composante connexe dont le nombre de sommets ou d'arêtes est supérieur aux seuils fixés (par défaut, nous utilisons 800 sommets et 3 200 arêtes). Autrement dit, nous supprimons les arêtes à faible poids liant des couples de sommets éloignés. De cette manière, si une composante connexe se crée tôt dans le processus de suppression et n'excède aucun des seuils, elle conserve autant d'arêtes que possible, et ainsi, autant d'informations sur sa topologie fondamentale. À la suite de cette suppression, nous dérivons de G un ensemble de graphes de localité. Chaque graphe est une composante connexe de G. Nous notons $LG_i = (V_i, E_i)$ le i-ème graphe de localité dans cet ensemble.

3.4.5 Connexion des données réelles et de leurs projetés

La méthode décrite dans les sections précédentes ne garantit pas que chaque sommet correspondant à une donnée réelle est connecté à sa projection sur la frontière. Par conséquent, elle donne de petits graphes de localité LG_i avec un petit nombre de sommets correspondant à des données réelles et un trop petit nombre de sommets de frontière. Cela induit un faible potentiel d'analyse. En effet, les sommets correspondant à des données réelles ont besoin de sommets de frontière pour être placés autour de la frontière de décision (section 3.4.2). Pour surmonter ce problème, nous fusionnons les graphes de localité, qui possèdent un nombre de sommets sous un certain seuil, avec d'autres graphes de localité en les connectant par des arêtes. Plus précisément, nous ajoutons des arêtes entre les sommets correspondant à des données réelles et leurs sommets les plus proches de la frontière dans G, *i.e.*, leurs projections.

Pour ce faire, pour chaque graphe de localité $LG_i = (V_i, E_i)$, nous considérons le sous-graphe $LG_i[R_i]$ où R_i est l'ensemble des sommets correspondant à des données réelles de V_i, et un seuil $th(|R_i|)$ dépendant de la taille de R_i. Un processus d'itération commence avec des graphes de localité qui contiennent moins de données réelles que les autres. Si le nombre de sommets de frontière dans le graphe de localité (noté $|B_i|$ où B_i est l'ensemble des sommets des données de frontière dans LG_i) est inférieur au seuil $th(|R_i|)$, on ajoute une arête pondérée à 1 entre le sommet de données réelles v de R_i qui est le plus proche de sa projection $p(v) \in G \backslash LG_i$ et sa projection $p(v)$.

Les connexions fusionnent certains graphes de localité. Ainsi, pour chaque connexion, nous redéfinissons l'ensemble des LG_i. Chaque fois qu'une connexion est créée, le processus d'itération recommence depuis le début pour garantir que les plus petits graphes de localité sont toujours traités avant les plus grands. Le seuil utilise une fonction logarithmique binaire afin de garantir un nombre minimal de sommets de données de frontière dans chaque graphe de localité en fonction du nombre de sommets de données réelles. La fonction logarithmique assure également qu'un petit nombre d'arêtes est créé afin de conserver le graphe de proximité le plus similaire possible de ce qu'il était avant l'étape de connexion des données réelles avec leur projetés.

3.4.6 Division du graphe de proximité sur les grandes composantes

La connexion des données réelles et de leurs projetés précédemment présentée produit de très grandes composantes qui contiennent de nombreux sommets. Ces dernières limitent ainsi l'interprétabilité de la localité. Par exemple, les mots les plus pertinents de ce type de grande composante ne contiennent pas beaucoup d'informations et cachent certains mots importants qui pourraient être mieux identifiés dans des composantes plus petites. Pour faire face à ce phénomène, nous divisons les plus grands graphes de localité.

Le processus de division des grandes composantes (ou grands graphes de localité) est motivé par l'observation que parfois, deux parties différentes d'un graphe de localité sont liées par peu d'arêtes. Ces deux parties peuvent concerner des textes très différents dans leur structure et/ou leur vocabulaire, mais n'étant liées que par quelques paires de

textes similaires. Ces paires peuvent contenir des mots identiques ou similaires, ou une structure très proche. Toutefois, cela ne justifie pas le regroupement des deux parties de la composante. Nous supprimons ces liens pour construire des localités significatives de l'espace de représentation et, par conséquent, une liste plus significative de mots-clés. Pour diviser, nous fixons des seuils maximum pour le nombre de sommets et le nombre d'arêtes (par défaut, nous utilisons 800 sommets et 3 200 arêtes). Ensuite, nous essayons de diviser les composantes qui dépassent l'un des seuils. La définition des seuils dépend des données et de la proximité des données entre elles. La division supprime les arêtes ayant la plus grande valeur de centralité intermédiaire [Ulr01] et dont le poids est différent de 1. L'objectif est d'assurer la connectivité des sommets apportée par la méthode UMAP. À chaque suppression d'arête, la valeur de la centralité intermédiaire est recalculée.

Certaines localités nouvellement créées peuvent dépasser le seuil $th(|R_i|)$ défini par la fonction logarithme binaire dans la section 3.4.5 qui assure un nombre minimal de sommets de données de frontière dans chaque graphe de localité. Si c'est le cas, la connexion aux sommets de données de frontière est à nouveau effectuée. Cette étape est toujours suivie de la division des grandes composantes. Ces deux étapes constituent les étapes d'optimisation. Elles se terminent lorsqu'aucun seuil n'est dépassé (nombre de sommets ou d'arêtes par localité et nombre minimal de données de frontière pour chaque localité).

À la fin des étapes d'optimisation, chaque graphe de localité correspond à une localité dans la vue globale proposée dans EBBE-Text (figure 3.6).

3.4.7 Connexion interne aux composantes des données réelles et de leurs projections

Pour les mêmes raisons que lors de la première connexion de frontière (section 3.4.5), une fois les étapes d'optimisation faites, nous connectons chaque sommet des données réelles à sa projection, si sa projection appartient au même graphe de localité. Cette étape est similaire à l'étape de connexion aux données de la frontière de décision sauf qu'elle ne considère que les sommets de la frontière de B_i (section 3.4.5) au lieu de $G \backslash LG_i$ et que les distances à la frontière de décision ne sont pas prises en compte. Cette étape facilite la visualisation des données autour de la frontière de décision car elles sont placées de manière plus éparse.

3.4.8 Simplification de la frontière de décision

Comme mentionné auparavant, les étapes de connexion à la frontière ne garantissent pas que tous les sommets des données initiales sont connectés à leurs projections sur la frontière. Pour éliminer les sommets de la frontière non connectés aux données initiales et non utiles pour l'analyse, nous sélectionnons d'abord les voisins des sommets des données initiales qui se trouvent sur la frontière, ainsi que les sommets et sommets des arbres situés sur les plus courts chemins entre eux. Ensuite, nous ajoutons à la sélection toutes les arêtes reliant les sommets appartenant déjà à la sélection. Enfin, nous supprimons tous les sommets et les arêtes de la frontière qui ne sont pas sélectionnés.

Cette étape clôture les étapes de manipulation de données. En sortie, nous obtenons les différents graphes de localités comprenant des sommets de données réelles et des sommets de données projetées (aussi appelées données de frontière tout au long de cette thèse).

3.5 Visualisations de données à deux échelles

Nous proposons deux vues principales dans EBBE-Text : la vue globale et la vue de localité. La vue globale, présentée dans la figure 3.6, répond à trois besoins (Be.1, Be.2, Be.4). La vue de localité, présentée dans la figure 3.7, répond à tous les besoins sauf le premier (Be.2-Be.7). Dans ces vues, nous codons les données avec trois couleurs : orange, vert et gris. Nous utilisons une palette de couleurs issue de ColorBrewer [HB03] pour les données qualitatives afin de choisir les couleurs verte et orange. Ces deux couleurs représentent les classes ou labels des données et le gris représente les données projetées sur la frontière de décision. Les positions gauche et droite autour de la frontière représentent les classes prédites.

3.5.1 Vue globale

La vue globale d'EBBE-Text (figure 3.6) permet de comparer le coefficient de corrélation de Matthews (MCC) [Mat75] entre les classes et les classes prédites par le réseau de neurones pour l'ensemble du corpus et pour les différentes localités de l'espace de représentation du texte. Le MCC du corpus entier et la matrice de confusion sont présentés dans la figure 3.6 dans le volet de gauche.

Les graphes de localité, dérivés du graphe de proximité (section 3.4.3), sont représentés par des diagrammes de flux sur le volet de droite de la figure 3.6 et montrent la distribution du sous-ensemble de données associée à chaque localité, autour de la frontière de décision. L'axe des x représente la distance d par rapport à la frontière renvoyée par le classifieur du réseau de neurones. L'axe des y représente le nombre de données correspondant à cette distance. Les diagrammes de flux ont été choisis au lieu des diagrammes en bâtons parce que la distance est une mesure continue. Par exemple, dans le diagramme de flux en haut à gauche de la figure 3.6, la localité n'a pas de données orange mal classées. Elles sont toutes placées à gauche de la frontière de décision (ligne grise). En revanche, il y a des données vertes des deux côtés de la frontière de décision, ce qui indique que beaucoup de données vertes sont mal classées.

Les diagrammes de flux sont organisés dans une matrice avec sept colonnes et suffisamment de lignes pour représenter toutes les localités. Les localités sont assignées aux colonnes en fonction de leur distribution par classe : les localités avec la plus grande proportion dans une classe sont dans les colonnes de gauche et les localités avec la plus grande proportion dans l'autre classe sont dans les colonnes de droite. Une colonne contient un septième des localités. Les localités sont ordonnées, au sein des colonnes, de la plus grande à la plus petite.

Nous proposons également quatre autres critères pour ordonner les localités dans les lignes et les colonnes de la matrice : le MCC, le nombre de mauvaises classifications,

FIGURE 3.6 – *Vue globale de l'espace de représentation du texte. Dans le volet de droite, dans l'encadré bleu, un diagramme de flux représente la distribution des données autour de la frontière de décision pour une localité. La position de la ligne grise dans chaque diagramme de flux indique la position de la frontière de décision. Les diagrammes de flux sont organisés dans une matrice et ordonnés en fonction de plusieurs mesures (taille de la localité, nombre de mauvaises classifications, etc.). Dans le volet de gauche, dans l'encadré rose, on retrouve la matrice de confusion pour les prédictions, et dans l'encadré vert, les différents types d'ordonnancement disponibles.*

le score de mono-classification (qui calcule si les données sont distribuées de manière égale autour de la frontière de décision), et la valeur du test de Shapiro-Wilk [SW65]. La valeur du test de Shapiro-Wilk calcule si la distribution des distances à la frontière de décision suit ou non une distribution normale. Cette valeur est utile, par exemple, dans le cas de données non étiquetées. En effet, une distribution normale indique que le réseau n'est pas efficace pour répartir les données autour de la frontière de décision pour une localité précise ou qu'une ambiguïté sur les données utilisées existe.

3.5.2 Vue de localité

Lorsque l'on sélectionne un diagramme de flux dans la vue globale, la vue de la localité correspondante s'ouvre (figure 3.7). La vue de la localité présente la sous-vue de la frontière de décision dans son volet gauche. La section 3.5.2.1 décrit cette vue en

détail. Dans le volet central, une liste de textes structurée en arbre montre les chemins d'un élément de donnée [1] sélectionné vers les éléments de données de la frontière de décision. Le volet de droite montre la liste des 10 mots les plus pertinents, les espaces de représentation construits avec différentes méthodes de réduction de dimension, une matrice de confusion interactive avec le score MCC de localité et un formulaire de saisie de texte permettant aux utilisateurs de classer leur propre texte. Toutes ces caractéristiques sont présentées dans la section 3.5.2.2.

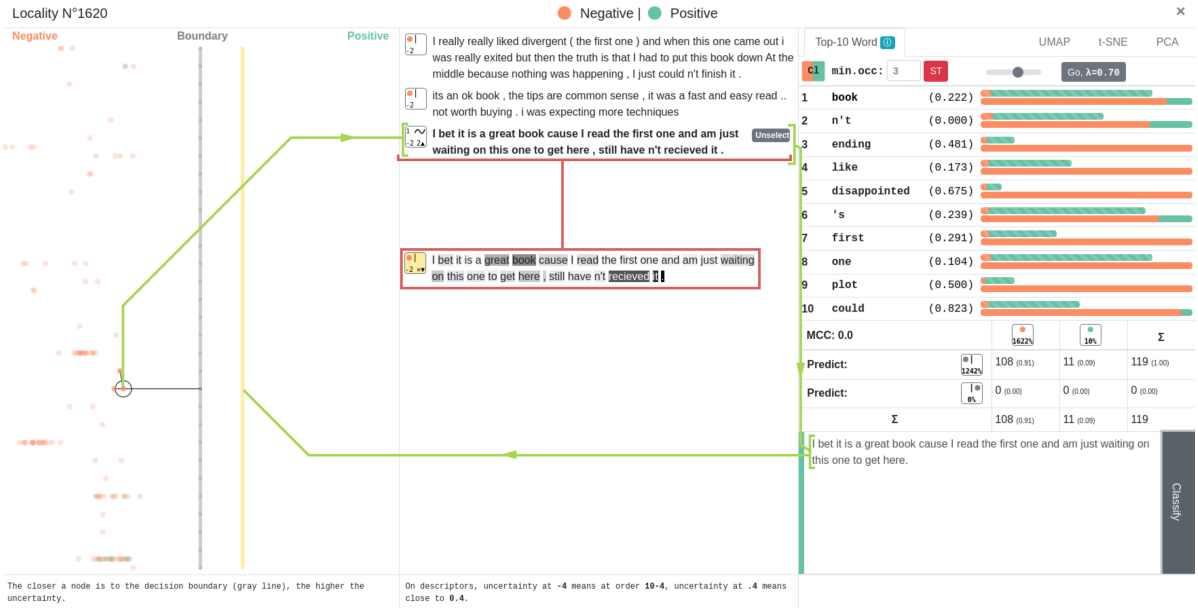


Figure 3.7 – *Exploration d'une localité avec EBBE-Text. De gauche à droite et de haut en bas : sous-vues de la frontière de décision, de la liste de textes et des chemins, de la liste des 10 mots les plus importants (les représentations de réduction de dimension peuvent être affichées à la place), de la matrice de confusion pour la localité, du formulaire de saisie à la volée de texte et de la commande de classification. Les* ***chemins verts*** *montrent comment une modification dans un texte peut entraîner un changement important dans la prédiction. La* ***ligne jaune*** *montre où le nouveau texte serait positionné dans la sous-vue de la frontière de décision. L'****encadré rouge*** *montre comment le score d'attention d'une prédiction s'affiche au passage de la souris.*

3.5.2.1 Sous-vue de la frontière

La sous-vue de la frontière de décision place les données autour de la frontière de décision en conservant la distance à celle-ci en abscisse et en assurant la proximité entre les données proches dans l'espace de représentation en ordonnée. Dans cette section, nous développons les différentes étapes de la visualisation de la frontière de décision à l'aide des différentes abstractions de données issues de la manipulation des données : les classes, les prédictions (et donc les distances à la frontière de décision) et les graphes de localités. Cette sous-vue répond aux besoins Be.2, Be.3, Be.5, Be.6 et Be.7.

1. Pour simplifier la lecture, "un élément de donnée" sera appelé "donnée" tout au long de la lecture. De même, les données de frontière sont les éléments de données issus de la projection et se trouvant donc sur la frontière de décision.

Tout d'abord, nous extrayons le sous-graphe des données de frontière, à partir du graphe de localité. Ensuite, nous arrangeons linéairement chaque composante connexe de ce sous-graphe avant d'empiler leurs éléments pour représenter l'ensemble de la frontière de décision de la localité. La frontière de décision est ainsi dessinée verticalement au centre de la vue. Les autres données sont dessinées à gauche ou à droite en fonction de la classe prédite. La distance entre une donnée et la frontière de décision le long de l'axe des abscisses représente l'incertitude de la prédiction (la probabilité issue de la prédiction du réseau de neurones). Plus une donnée est proche de la frontière, plus l'incertitude de sa prédiction est grande. Les données peuvent être sélectionnées et les textes associés sont affichés dans le volet central. Les chemins entre une donnée sélectionnée et la frontière de décision sont affichés dans la sous-vue de la frontière de décision. Les liens sont colorés de trois façons. Les liens noirs représentent les voisins directs de la donnée sélectionnée. Les liens bleus représentent les liens appartenant aux chemins vers la frontière de décision. Les liens roses sont similaires aux bleus, sauf qu'ils traversent la frontière de décision. Cela signifie que les données appartenant à la paire liée sont prédites dans les deux classes.

Nous avons choisi de représenter la frontière de décision sous la forme d'une ligne. Les distances des données par rapport à une ligne droite sont facilement interprétables et comparables. Pour construire la ligne représentant la frontière de décision de la localité, nous considérons le sous-graphe de données de frontière $LG_i[B_i]$ du graphe de localité LG_i où B_i est l'ensemble des sommets de données de frontière. Nous commençons par positionner les sommets de chaque composante connexe de $LG_i[B_i]$ indépendamment. Ensuite, nous ordonnons les composantes connexes le long de l'axe des ordonnées. Enfin, nous ajoutons les autres sommets de LG_i autour de la frontière créée.

Pour produire l'**arrangement linéaire des données projetées d'une composante connexe**, nous positionnons les sommets de chaque composante connexe de $LG_i[B_i]$ dans un espace discret à une dimension. Trouver un ordre qui minimise les distances entre les sommets reliés est connu comme le problème de l'arrangement linéaire minimum, qui est NP-complet. Nous utilisons l'algorithme TSSA-Φ proposé par RODRIGUEZ-TELLO, HAO et TORRES-JIMENEZ [RHT08] pour approcher la solution optimale pour chaque composante connexe.

Soit $C_{i,j} = (V_{i,j}, E_{i,j})$ la j-ième composante connexe de $LG_i[B_i]$ (voir les zones colorées dans la figure 3.8(a)). Pour chaque $C_{i,j}$, nous calculons une minimisation de l'accroissement frontal [Mca99] pour produire l'arrangement linéaire initial $\phi_{i,j}$. Ensuite, l'algorithme TSSA-Φ itère à travers les données, tant que la réduction des distances montre une amélioration significative. À chaque étape, il sélectionne dans le graphe $C_{i,j}$ un sommet aléatoire $u \in V_{i,j}$ et tente d'échanger sa place avec d'autres par deux procédures, exécutées respectivement lors de 10% et 90% d'itérations :

- test de toutes les combinaisons possibles $u, v \in V_{i,j}$, tel que $u \neq v$ et choix de la combinaison qui améliore le plus la fonction de coût.
- test de toutes les combinaisons possibles $u, v \in V_{i,j}$, tel que $v \in M(u)$ et $u \neq v$ et choix de la combinaison qui améliore le plus la fonction de coût.

Dans la deuxième procédure, $M(u) = \{v : med(u) - 2 \leq \phi_{i,j}(v) \leq med(u) + 2\}$ où $\phi_{i,j}(v)$ est la position de v dans $\phi_{i,j}$ et $med(u)$ est la position médiane des sommets adjacents de u (figure 3.8(b)).

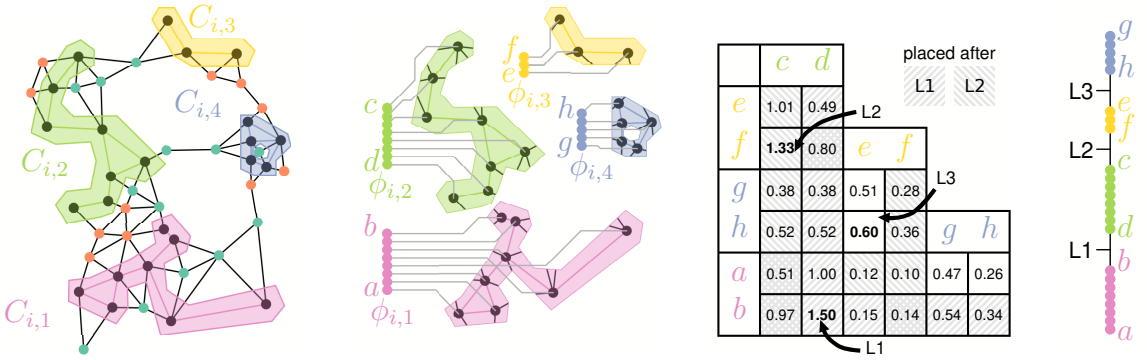


FIGURE 3.8 – *Arrangement linéaire ϕ_i d'une composante connexe LG_i (graphe de localité). (a) Identification des différentes composantes connexes $C_{i,j}$ des données projetées (ici, quatre) dans un graphe de localité LG_i. (b) Arrangements linéaires $\phi_{i,j}$ issus de l'algorithme TSSA-Φ pour chaque composante connexe des données projetées $C_{i,j}$. (c) Matrice de proximité entre les extrémités a, b, c, d, e, f et g des arrangements linéaires $\phi_{i,j}$. L1, L2, L3 représentent respectivement le premier, le deuxième et le troisième choix de placement ou d'empilement. À chaque étape, l'empilement est uniquement recherché entre les extrémités des arrangements linéaires non placés et les extrémités réelles de l'arrangement linéaire final. Les nouvelles paires d'extrémités inéligibles sont représentées par des striages sud-ouest nord-est pour L1 et par des striages sud-est nord-ouest pour L2 (e.g., il n'existe plus que quatre possibilités différentes de placement à la dernière étape). (d) Arrangement linéaire final ϕ_i produit par l'empilement ordonné des arrangements linéaires $\phi_{i,j}$.*

Lorsque l'arrangement linéaire $\phi_{i,j}$ a été réalisé pour chaque $C_{i,j}$ de LG_i, nous produisons un **ordonnancement des composantes connexes** (figures 3.8(c) et 3.8(d)) en les empilant les unes sur les autres. Soit $\text{sides}(\phi) = \{v : v = \phi^{-1}(1) \vee v = \phi^{-1}(|V|), v \in V\}$ l'ensemble contenant les deux extrémités d'un arrangement linéaire ϕ parmi un ensemble de sommets V. Soit $u \in \text{sides}(\phi_{i,j})$ et $w \in \text{sides}(\phi_{i,k})$ les côtés de deux arrangements linéaires $\phi_{i,j}$ et $\phi_{i,k}$. Nous définissons une mesure de proximité entre u et w comme suit :

$$\text{px}(u, w) = \sum_{v_1 \in V_{i,k}} \frac{1}{\text{SP}_i(v_1, u)^2} + \sum_{v_2 \in V_{i,j}} \frac{1}{\text{SP}_i(v_2, w)^2} \tag{3.1}$$

où $\text{SP}_i(p, q)$ est la longueur du plus court chemin entre p et q avec $p, q \in B_i$.

Cette mesure de proximité donne un score plus élevé aux composantes qui sont proches les unes des autres car la valeur ajoutée au score pour chaque lien est divisée par le carré de la longueur du lien. Elle donne également un score plus élevé entre les grandes composantes car la somme des poids des liens n'est pas pondérée par la taille des composantes. Enfin, elle donne un score plus élevé entre les composantes dont les sommets sont proches des extrémités des autres composantes. Un exemple de résultats du score de proximité est présenté dans la figure 3.8(c).

Sur la base de cette mesure de proximité entre les extrémités des arrangements linéaires, nous pouvons maintenant calculer l'ordonnancement ϕ_i des sommets des

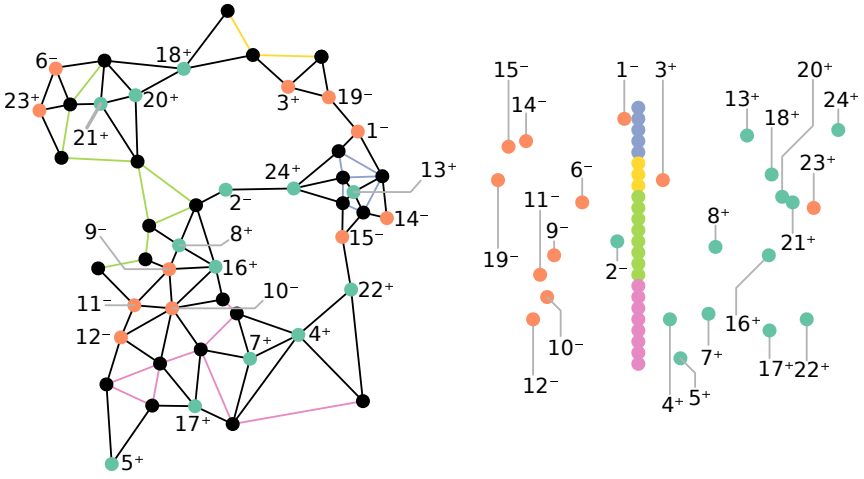
composantes connexes de $LG_i[B_i]$. Soit $P_{\phi_i} = \{u : \exists \phi_i(u), u \in B_i\}$ l'ensemble des sommets positionnés. Initialement, $P_{\phi_i} = \emptyset$. Nous commençons par sélectionner la plus grande composante connexe $C_{i,j} \in LG_i[B_i]$ et définissons $\phi_i(v) = \phi_{i,j}(v), \forall v \in V_{i,j}$. Ensuite, nous choisissons de manière itérative u et v de telle sorte que :

- $u \in \text{sides}(\phi_i)$,
- $v \in \text{sides}(\phi_{i,k})$, $v \notin P_{\phi_i}$,
- $\text{px}(u, v)$ est la valeur la plus élevée parmi tous les candidats pour v.

Si $u = \phi_i^{-1}(1)$, nous inversons l'ordonnancement de l'arrangement linéaire ϕ_i, donc $u = \phi_i^{-1}(|P\phi_i|)$. Ensuite, $\forall w \in V_{i,k}$, nous fixons sa position dans ϕ_i comme suit :

$$\phi_i(w) = \begin{cases} \phi_{i,k}(w) + |P\phi_i| & \textbf{si } v = \phi_{i,k}^{-1}(1) \\ 1 - \phi_{i,k}(w) + |V_{i,k}| + |P\phi_i| & \textbf{si } v = \phi_{i,k}^{-1}(|V_{i,k}|) \end{cases} \tag{3.2}$$

Le processus se termine lorsque $|P_{\phi_i}| = |B_i|$.



(*a*) (*b*)

FIGURE 3.9 – *Placement des données autour de la frontière. (a) Les données non projetées sont étiquetées de 1 à 24 en fonction de leur proximité avec la frontière de décision. Chaque étiquette est assortie d'un moins (-) ou d'un plus (+) : le classifieur prédit une classe ou une autre. (b) Placement β_i du graphe de localité LG_i. Les données situées à gauche de la frontière sont classées d'une certaine manière, et celles situées à droite d'une autre manière (e.g., trois erreurs de classification ici). Les données sont positionnées de la plus proche à la plus éloignée (étiquetées de 1 à 24).*

Pour produire le **positionnement final des données et des données projetées**, nous positionnons les sommets de V_i dans la vue de la frontière de décision à deux dimensions finale β_i d'une localité LG_i. Pour positionner ces sommets, nous sélectionnons d'abord tous les sommets dans B_i et définissons leur position en abscisse à zéro et leur position en ordonnée à leur position dans l'ordonnancement de ϕ_i. Une fois toutes les données de frontière positionnées, nous trions les sommets de $V_i \backslash B_i$ par ordre croissant en fonction de leur distance d_v à la frontière de décision issue du réseau de neurones. Ensuite, nous itérons à travers les données ordonnées et pour chaque $u \in V_i \backslash B_i$, nous

fixons sa position en abscisse à d_u et sa position en ordonnée à la position médiane de ses voisins déjà positionnés (y compris les sommets des données projetées de B_i). Si un sommet u n'a pas de voisins déjà positionnés, nous l'ignorons, *e.g.*, un sommet sans lien avec un sommet de B_i et avec d'autres voisins plus éloignés de la frontière de décision. Une fois la première itération terminée, tous les sommets ne sont pas positionnés. On répète la procédure en utilisant ces sommets jusqu'à ce qu'ils soient tous positionnés. Un exemple de ce processus est donné dans la figure 3.9.

3.5.2.2 Autres sous-vues

Dans cette section, nous présentons les sous-vues de la vue de la localité. Ces sous-vues donnent un aperçu du processus de classification du réseau de neurones dans la localité choisie. Les interactions entre les sous-vues sont décrites dans la figure 3.10.

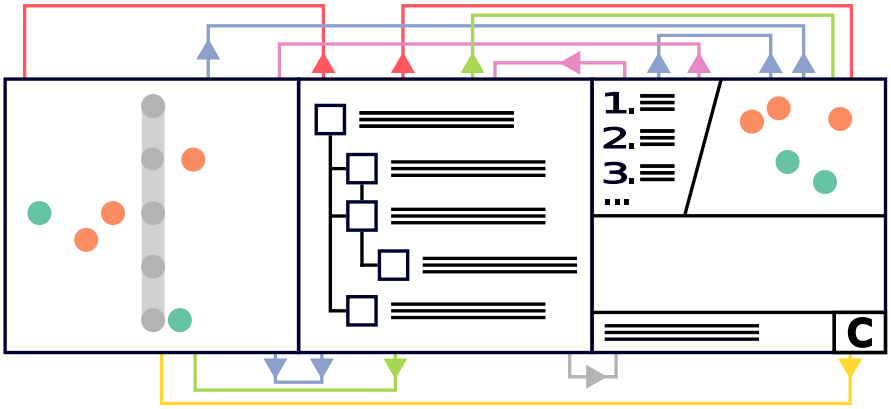


FIGURE 3.10 – *Interactions entre les différentes sous-vues de la vue de localité. De gauche à droite et de haut en bas : la sous-vue de la frontière de décision, la liste de textes et leurs descripteurs, la liste des mots les plus pertinents, les volets des espaces de représentation de textes résultant des différentes techniques de réduction de dimension, la matrice de confusion et le MCC (sans interaction) et la zone de saisie de texte accompagnée de son bouton pour lancer la classification. **Les liens bleus** : sélectionner une donnée dans cette sous-vue la sélectionne également dans les autres sous-vues. Les données survolées dans la sous-vue affichent leurs liens de voisinage dans les autres sous-vues. Les liens de voisinage sont toujours affichés dans la liste des textes des chemins en raison de sa structure arborescente. **Les liens roses** : survoler un mot dans la liste des mots les plus pertinents met en évidence à l'aide de leurs descripteurs les données contenant ce mot dans la liste des textes. Les données contenant le mot survolé sont également encerclées dans les autres sous-vues. **Les liens verts** : survoler un descripteur de texte met en évidence la position de la donnée associée dans les autres sous-vues. **Le lien gris** : un clic sur un texte dans la liste de textes remplit la zone de saisie de texte de la sous-vue de classification. **Le lien jaune** : un clic sur le bouton calcule la prédiction pour le texte saisi par l'utilisateur et affiche sa position dans la sous-vue de la frontière de décision. **Les liens rouges** : un mot survolé dans un texte montre la position des données contenant ce mot dans d'autres sous-vues.*

La **liste des textes accompagnés de leurs descripteurs** répond aux besoins Be.3, Be.4, Be.5 et Be.6. Lorsqu'une donnée est sélectionnée dans la vue de la frontière de décision, son texte, les textes de ses voisins et les textes se trouvant sur les chemins de celui-ci à la frontière de décision sont affichés dans une liste arborescente. Nous avons choisi d'utiliser une liste arborescente car elle permet aux utilisateurs de savoir rapidement quelles sont les données qui sont proches de la frontière de décision et

de visualiser les chemins vers la frontière de décision. Elle répond au besoin Be.3. Pour chaque texte, un descripteur (décrit dans la figure 3.11) donne des informations sur la prédiction associée résultant du réseau de neurone. Lors d'un "double-clic" sur un descripteur, la donnée associée est sélectionnée. Au "survol" d'un descripteur, la donnée associée est entourée dans la vue de la frontière de décision et dans les vues des espaces de réduction de dimension (voir ci-dessous) avec un cercle de couleur pâle selon la classe de la donnée. Enfin, un "clic" sur un descripteur permet de masquer ou d'afficher les chemins jusqu'à la frontière de décision, dans la liste arborescente, et donc les textes se trouvant sur ces chemins.

(a) (b) (c)

FIGURE 3.11 – *Descripteurs des données placées dans la liste des textes. Les descripteurs sont placés hypothétiquement du plus éloigné de la frontière de décision à gauche en (a) au plus proche de la frontière de décision à droite en (c). Le second descripteur, en (b), est le descripteur de la donnée sélectionnée. Le descripteur à sa gauche, en (a), et les descripteurs à sa droite, en (c), concernent respectivement ses voisins plus éloignés ou plus proches de la frontière de décision. Les* ***rectangles bleus*** *montrent la classe de la donnée en la colorant. La position du point informe l'utilisateur sur la classification (à gauche pour une classe et à droite pour une autre). Les* ***données oranges (points)*** *sont corectement classifiées à gauche de la ligne centrale. Inversement, les* ***données vertes (points)*** *sont correctement classifiées à droite de la ligne centrale. On observe ici par exemple que la donnée la plus à droite en (c) n'est pas correctement classifiée. Le* ***rectangle rose*** *informe l'utilisateur sur le nombre de chemins entre la donnée sélectionnée et la frontière de décision. Les* ***carrés jaunes*** *indiquent le degré d'incertitude du réseau de neurones par rapport à une prédiction. Une valeur précédée d'un symbole moins signifie à $\times 10^{-valeur}$. Un point signifie la valeur elle-même (e.g., "-4" signifie une incertitude d'ordre 10^{-4}, ".4" signifie une incertitude proche de .4). Le* ***carré rouge*** *indique le nombre de voisins de la donnée sélectionnée plus éloignés de la frontière de décision que la donnée sélectionnée (× signifie aucun). Les* ***carrés verts*** *indiquent le nombre de voisins plus proches de la frontière de décision (× signifie aucun).*

Si le réseau de neurones utilisé possède un ou plusieurs vecteurs d'attention ou qu'un score indiquant l'impact des mots dans la prédiction a été mesuré, on peut visualiser ces scores pour chaque mot dans les textes. Les scores sont représentés par une carte de chaleur selon deux modes en fonction de la structure des scores : (1) Si le score est une matrice carrée qui donne les scores d'attention (ou autre score) dans la prédiction pour les mots par rapport aux autres, alors les scores d'attention sont montrés par rapport au mot survolé; (2) Sinon, le survol du texte montre le score d'attention pour chaque mot. De plus, les textes contenant le mot survolé sont mis en évidence dans la sous-vue de la frontière de décision et dans la sous-vue des espaces de réduction de dimension. Comme nous voulons visualiser facilement les scores d'attention (ou de contribution) entre les données, nous utilisons des cartes de chaleur plutôt que des graphiques en relief ou autres pour faciliter la visualisation des différentes données et pour la comparaison des scores. Cela répond au besoin Be.5.

La **liste des mots les plus pertinents** [SS14], pour la localité sélectionnée, est affichée pour répondre au besoin Be.5. Nous avons choisi d'ignorer les mots utilisés de

manière commune dans tous les sujets en anglais comme "a", "to", "the" (stop words) et nous avons défini la pertinence d'un mot w pour la localité l en fonction d'un paramètre de poids λ (où $0 \leq \lambda \leq 1$) comme suit :

$$\mathbf{r}(w, l|\lambda) = \lambda \log(p_{w,l}) + (1 - \lambda) \log(\frac{p_{w,l}}{p_w}) \tag{3.3}$$

λ détermine le poids donné à la fréquence du mot w sous la localité l par rapport à son "lift" (en mesurant les deux sur l'échelle logarithmique). Le "lift" d'un mot est défini comme la fréquence d'apparition du mot dans une localité par rapport à sa fréquence marginale d'apparition dans le corpus. En fixant $\lambda = 1$, on obtient le classement des mots dans l'ordre décroissant de leur fréquence spécifique à la localité, et en fixant $\lambda = 0$, on classe les mots uniquement en fonction de leur "lift". La fréquence marginale du mot w dans l'ensemble du corpus est notée p_w. $p_{w,l}$ désigne la fréquence marginale du mot w dans la localité l. La liste des mots à ignorer peut être modifiée par les utilisateurs. Cette modification lui permet d'ignorer les mots non importants dans son corpus.

Afin de répondre au besoin Be.5, nous proposons trois modes différents de comparaison de la prédiction et de la classe à l'échelle des données de la localité et/ou du corpus entier (figure 3.12). Des barres de distribution empilées les unes sur les autres permettent de comparer l'influence des mots sur la prédiction en fonction du mode choisi. Le passage d'un mode à l'autre se fait en cliquant sur le bouton en haut à gauche du tableau des mots les plus pertinents. Le premier mode montre deux barres de distribution. Celle du haut présente la distribution des classes des textes contenant le mot associé dans l'ensemble du corpus. Celle du bas présente la distribution des classes des textes contenant le mot associé dans la localité uniquement. Le deuxième mode concerne la distribution des prédictions au lieu de la distribution des classes. Le troisième mode, la barre supérieure présente les distributions des prédictions des données de la localité et la barre inférieure présente la distribution des classes des données de localité pour le texte contenant le mot associé.

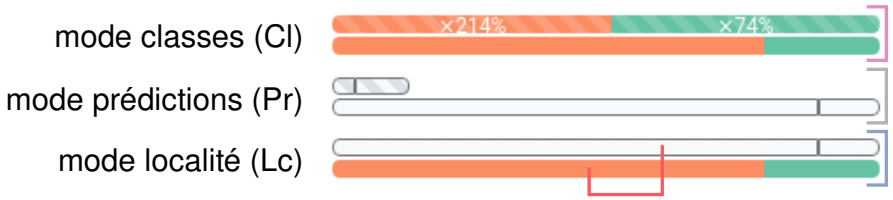


Figure 3.12 – *Barres de distribution pour chaque mot dans la liste des mots les plus pertinents. Dans le crochet rose, nous pouvons voir les distributions par classe des textes contenant le mot associé. Les barres sont à la même échelle et permettent les comparaisons entre le corpus entier des données (barre du haut) et la localité (barre du bas). Dans le crochet gris, nous pouvons voir les distributions des prédictions. Les barres peuvent être mises à l'échelle pour être comparées, comme présenté dans le mode classes. Par défaut, le ratio de la longueur des barres correspond au ratio de la probabilité d'apparition du mot associé. La barre du haut concerne le corpus en entier et la barre du bas uniquement les textes de la localité. Dans le crochet bleu, se trouve le mode de localité. Il est utilisé pour inspecter jusqu'à quel point le classifieur prédit une classe plus souvent qu'il ne le devrait, en comparant les fréquences dans la barre de classe et la barre de prédiction (montrée ici par les segments rouges).*

Le nombre d'occurrences de chaque mot est indiqué dans le mode de localité. Les valeurs-p des mots sont indiquées dans les modes prédictions et classes. Ces valeurs-p, obtenues par des tests de χ^2 [Pea00]. Elles représentent la probabilité que les fréquences observées des classes ou des prédictions entre la localité et l'ensemble du corpus suivent la même règle de distribution. Par conséquent, une faible valeur-p signifie qu'il existe une probabilité élevée que le réseau de neurones ait un comportement spécifique avec un mot dans cette localité. Cela permet aux utilisateurs d'identifier des synergies entre les mots et leurs sens pour certaines localités.

Les **sous-vues des techniques de réduction de dimensions** répondent au besoin Be.6 en permettant aux utilisateurs de comparer les distances entre les données dans des espaces alternatifs à celui construit pour visualiser la frontière de décision. Nous proposons trois méthodes de réduction de dimension UMAP [MHM18], la t-SNE [HR03; MH08] et l'ACP [Pea01] pour des raisons de popularité mais aussi d'efficacité. Proposer des méthodes de réduction de dimension est essentiel pour comprendre réellement la proximité des voisins et observer ensuite dans quelle mesure les voisins peuvent être classés différemment. Lorsqu'une donnée est sélectionnée dans la vue de la frontière de décision, les liens entre cette donnée et ses voisins directs sont affichés dans les vues des espaces issues des techniques de réduction de dimension. Les données peuvent également être sélectionnées dans les vues des réductions de dimensions. Lors d'un survol d'un descripteur de texte, la donnée est entourée d'un cercle de couleur faiblement opaque en fonction de la classe de celle-ci (figure 3.10). Enfin, des diagrammes de Voronoï [For87] dépendant des classifications sont proposés pour chacune de ces méthodes de réduction. Ils permettent d'identifier si un groupe de voisins proches contient des données classées différemment, directement dans les espaces de réduction de dimension [MWV15]. Leur utilisation confirme le besoin d'une technique de visualisation des frontières qui présente des distances significatives par rapport à la frontière de décision, ce que cette méthode ne permet pas. Dans le cas de frontière de décision linéaire avec une utilisation de la méthode ACP [Pea01], on peut supposer que la tâche de classification dans cette localité est simple.

Deux **matrices de confusion accompagnées de MCC** sont proposées dans notre méthode pour répondre au besoin Be.4, l'une pour le corpus entier et l'autre pour la localité sélectionnée. Le MCC de la localité entre les prédictions et les classes est également affiché. Les matrices de confusion donnent les sommes marginales et la distribution des classifications et des classes pour chaque modalité. Ces informations sont disponibles par défaut ou au passage de la souris.

La **zone de saisie de texte** donne la possibilité de prédire la classe d'un texte saisi par l'utilisateur pour répondre au besoin Be.4. Ce texte est classé par le réseau de neurones et une ligne jaune représente la position de la prédiction de ce texte dans la vue de la frontière de décision. Nous avons choisi une ligne au lieu d'un point car elle facilite la comparaison entre les données saisies et toute autre donnée dans la localité. Un utilisateur peut récupérer un texte, présent dans la liste des textes, en remplissant à l'aide d'un clic le formulaire de saisie de texte (figure 3.10). Ensuite, il peut modifier ce texte puis le classifier. Cette fonctionnalité permet d'observer l'importance des tokens ou mots ou de leurs ordres dans un texte.

3.6 Évaluation

L'évaluation d'EBBE-Text est divisée en trois parties. La première partie concerne une évaluation effectuée par un public d'experts en visualisation ou en réseaux de neurones. La seconde partie est une étude proposée à des utilisateurs débutants en classification de textes à l'aide de réseaux de neurones. Notre objectif est de montrer que les encodages visuels d'EBBE-Text sont accessibles aux non experts. Dans la troisième partie, nous avons réalisé une étude de cas pour mettre en évidence la manière dont notre méthode fournit de nouvelles connaissances et de nouvelles perspectives sur la classification dichotomique de textes.

3.6.1 Évaluation par des experts

Pour évaluer EBBE-Text, nous avons procédé à une évaluation selon la méthodologie ICE-T ("Insight", "Confidence", "Essence", et "Time") proposée par Wall et al. [Wal+19]. Sept participants (des chercheurs en sciences des données considérés comme des experts) ayant des connaissances sur les tâches de classification, l'apprentissage profond et la visualisation ont rempli le questionnaire ICE-T. Quatre d'entre eux étaient spécialisés dans la visualisation et les trois autres dans les applications des réseaux neuronaux en TAL. Les résultats présentés dans le tableau 3.1 sont globalement satisfaisants. Une description approfondie et les intitulés des questions du questionnaire ICE-T sont disponibles dans les travaux de Wall et al. [Wal+19].

Comme prévu, l'évaluation "Insight" est très positive. La liste arborescente du texte et la visualisation de la position des données dans l'espace de représentation et autour de la frontière de décision permettent aux utilisateurs d'explorer et de faire émerger des connaissances sur les données. L'évaluation "Time" montre l'efficacité d'EBBE-Text pour explorer les données ou même trouver des données particulières. Par exemple, les données incorrectement classifiées sont facilement découvertes en utilisant les options de tri dans la vue globale (figure 3.5.1). L'évaluation de l'"Essence" est moins satisfaisante. Cependant, il est difficile de capturer un sens réel de l'"image globale" concernant les données lorsque le but est de produire des explications locales et donc sur un espace réduit. Toutefois, EBBE-Text propose une matrice de confusion et un MCC qui fournissent des informations sur la qualité globale du classifieur. Enfin, la composante "Confiance" de l'enquête est satisfaisante. EBBE-Text augmente donc la confiance dans les réseaux de neurones. Dans le cas d'une mauvaise conception et d'une faible confiance dans l'outil de visualisation, il aurait été difficile, voire impossible, d'augmenter la confiance des utilisateurs dans les réseaux de neurones alors même qu'il s'agit de l'objectif principal d'EBBE-Text.

3.6.2 Évaluation par des débutants

L'évaluation effectuée sur des utilisateurs débutants en classification de textes à l'aide de réseaux de neurones utilise le jeu de données AmazonReview [HM16]. Nous avons choisi ce jeu en nous éloignant de la thématique initiale de la thèse car les textes et les classes renseignés sont facilement compréhensibles par des non-spécialistes. Il contient des avis sur des produits étiquetés de 1 à 5 (en fonction de leur note) vendus

		Question	μ	σ
Insight $\mu = 5{,}89$	La visualisation permet de répondre aux questions à propos des données	1	6,43	0,24
		2	5,50	2,58
		3	6,00	0,29
	La visualisation fournit une nouvelle ou meilleure compréhension des données	4	6,00	0,57
		5	6,00	0,86
	La visualisation permet de faire des découvertes fortuites	6	5,43	2,24
		7	5,57	2,53
		8	6,17	0,14
Time $\mu = 6{,}09$	La visualisation permet une compréhension rapide pour une navigation efficace	9	6,00	0,57
		10	6,00	0,86
	La visualisation fournit des mécanismes pour rechercher rapidement des informations spécifiques	11	5,86	1,84
		12	6,14	0,41
		13	6,43	0,53
Essence $\mu = 5{,}11$	La visualisation fournit une vue d'ensemble des données	14	4,86	3,27
		15	5,29	1,63
	La visualisation permet de comprendre les données au-delà des cas individuels	16	5,14	0,41
		17	5,14	1,27
Confidence $\mu = 5{,}44$	La visualisation permet d'éviter de faire des déductions erronées	18	6,14	0,41
		19	5,86	0,12
	La visualisation renseigne sur le domaine des données	20	5,00	1,43
	La visualisation aide à comprendre la qualité des données	21	4,67	1,89

TABLE 3.1 – *Résultats obtenus à partir de l'enquête ICE-T sur EBBE-Text. Les réponses possibles s'étendent sur une échelle de un à sept. μ et σ représentent la moyenne et l'écart type des réponses. Chaque partie de l'évaluation est elle-même divisée en plusieurs items ou pour chacun d'entre eux il existe plusieurs questions. Pour simplifier la lecture, nous avons simplement inclu une traduction de l'intitulé de l'item sans les questions. Ces dernières, ainsi que les intitulés originaux sont disponibles dans les travaux de* WALL *et al.* [Wal+19].

par Amazon. Nous avons conservé uniquement les avis correspondant aux meilleures et aux pires notes (1 et 5), ne contenant pas plus de 64 mots pour coller aux spécificités du réseau de neurones (un RNN) utilisé. Le jeu de données d'entraînement final contient 4 millions d'entrées. Pour cette étude, nous avons utilisé les 50 000 premières entrées. Dans les figures qui suivent, les nœuds orange représentent les avis négatifs et les nœuds vert les avis positifs. De la même manière, être à gauche de la frontière de décision signifie que le classifieur prédit le texte négativement et être à droite positivement.

Nous avons évalué nos encodages visuels à l'aide d'un questionnaire à choix multiple (QCM) avec quatre modalités pour chaque question (chapitre 5.2.2). Ce QCM est composé de 23 questions. Chacune d'entre elles fait référence à un ou plusieurs besoins. 21 participants ont accompli cette évaluation. Il s'agit d'étudiants en master de mathématiques et d'informatique appliquée aux sciences humaines et sociales. Ces étudiants ont été considérés comme des utilisateurs débutants car ils ne manipulent pas de réseaux de neurones en TAL quotidiennement mais ont une idée claire de ce qu'est une tâche de classification ou du fonctionnement d'un réseau de neurones.

Les résultats de l'évaluation, présentés dans le tableau 3.2, montrent le nombre de questions par besoins. La majorité des réponses étaient correctes pour tous les encodages visuels liés aux besoins.

	Be.1	Be.2	Be.3	Be.4	Be.5	Be.6	Be.7
Nombre de questions	3	9	6	2	5	2	1
Fréquence de réponses correctes	0,90	0,85	0,70	0,76	0,77	0,64	0,71

Table 3.2 – *Résultats obtenus à partir de l'étude effectuée sur des utilisateurs débutants vis-à-vis des besoins identifiés dans la section 3.3.2.*

Nous avons évalué l'efficacité de l'encodage visuel avec un test du χ^2 [Pea00]. Le niveau de confiance selon les résultats de notre test était de 0,05 et la distribution de toutes les réponses était significativement différente d'une distribution uniforme. Les résultats des questions ont validé l'encodage par rapport aux besoins. En effet, nous avons observé une différence significative pour chacune d'entre elles. Cependant, une des questions, concernant le nombre de voisins directs plus proches de la frontière de décision que celui sélectionné (Be.3), a montré que l'un des encodages n'était pas efficace. Il semble que lorsque la donnée sélectionnée a un voisin plus éloigné de la frontière de décision, les utilisateurs débutants confondent la donnée sélectionnée avec son voisin le plus éloigné dans la vue de la frontière de décision. Nous avons corrigé ce problème en entourant la donnée sélectionnée avec un cercle noir. Nous avons fait de même pour les sous-vues de l'espace de représentation des données issues des techniques de réduction de dimension.

Dans les questions répondant aux besoins Be.6 et Be.7, les utilisateurs devaient apprécier les distances entre les données. La proximité des données pouvait faire l'objet d'appréciations subjectives : cela explique pourquoi la fréquence des bonnes réponses est plus faible que dans les autres questions. Cependant, dans le cas de mauvaises réponses, ces réponses sont les moins mauvaises d'entre toutes les propositions de la question à choix multiples.

3.6.3 Études de cas

Afin de montrer comment l'utilisation d'EBBE-Text se généralise à travers différents jeux de données et réseaux de neurones, nous présentons trois études de cas basées sur trois réseaux de neurones et deux jeux de données. Le premier est un AE qui reconstruit la donnée d'entrée (section2.2.2). Les deux autres classifient uniquement les données. Les deux premiers cas d'étude utilisent le jeu de données Amazon Review [HM16] (présenté dans la section 3.6.2). Le dernier utilise un jeu de données d'informations et d'infoxs [2] [ATS18 ; ATS17].

La première étude de cas montre comment nous pouvons inspecter les classifications et comment nous pouvons extraire des connaissances sur la qualité du réseau de neurones à l'aide d'EBBE-Text. La deuxième étude présente comment nous pouvons inspecter les prédictions erronées et comment nous pouvons trouver ce qui est pertinent

2. https://www.kaggle.com/clmentbisaillon/fake-and-real-news-dataset

dans le texte pour les prédictions. Elle présente également comment essayer un texte saisi par l'utilisateur pour confirmer ou infirmer une hypothèse sur le fonctionnement interne d'un réseau de neurones. La troisième étude de cas montre comment interpréter les distances à la frontière de décision et l'incertitude. Elle montre également que des connaissances sur la qualité de l'entraînement du réseau de neurones et sur le réseau de neurones lui-même peuvent être extraites, en inspectant les distances à la frontière de décision des prédictions.

3.6.3.1 RNN multi-tâche

Dans notre première étude de cas, le jeu de données contient des avis sur les produits d'Amazon [HM16] étiquetés comme "négatifs" (orange, doit se trouver à gauche de la frontière de décision) ou "positifs" (vert, doit se trouver à droite de la frontière de décision). Nous vérifions la prédiction sur un ensemble de données comprenant des données d'entraînement et de test (les 50 000 premières entrées). Le réseaux de neurones entraîné est un RNN AE [Del+21a] utilisé pour la classification de textes et la reconstruction des entrées.

Nous nous concentrons sur une localité et observons que son MCC est plus élevé que celui du corpus entier (0,75 *vs* 0,71), ce qui signifie que le classifieur est meilleur dans cette localité que dans le corpus entier (Qu.3). Ensuite, en explorant la localité, nous observons la capacité du réseau de neurones à capturer les nuances comme suit. Nous découvrons un groupe de phrases ayant un sens très proche car voisines dans l'espace de représentation de UMAP et dans la vue de la frontière de décision (Qu.2). Nous comparons trois de ces phrases (figure 3.13). Ces phrases traitent de l'état d'un livre qui est arrivé rapidement. Toutes ces phrases sont classées correctement. Cependant, les nuances de chaque phrase entraînent des disparités de classification visibles dans la sous-vue de la frontière de décision. Les incertitudes de prédiction sont faibles (de l'ordre de 10^{-2} ou moins) (Qu.1). Lorsque l'on regarde le contenu de ces phrases, être arrivé "dans un court laps de temps" semble meilleur que d'être arrivé en "quelques jours". De même, être en "excellent état" semble meilleur que d'être dans "l'état qui était indiqué". Les nuances capturées par le réseau de neurones semblent correctes et le conduisent à classer ces phrases avec certitude et cohérence en fonction de la présence des mots et de leur sens, de la co-occurrence des mots, *etc.* (Qu.5, Qu.6).

3.6.3.2 RNN avec attention à tâche unique

Dans la deuxième étude de cas, le jeu de données est le même que celui utilisé dans la première étude de cas [HM16] avec les 100 000 premières entrées. Le réseau de neurones de classification de textes entraîné est un RNN auto-attentif structuré [Lin+17a] (SSA).

Dans la figure 3.7, nous observons que la donnée sélectionnée est mal classée (Qu.1) : le RNN l'a classée comme "négative" alors que son étiquette est "positive". En regardant le texte, on constate que l'avis semble en fait être plutôt "négatif". Nous observons ce phénomène dans plusieurs localités : il y a beaucoup de textes mal classés qui sont en fait des données faussement étiquetées (Qu.6). Dans cette localité précise, la moitié des textes mal classés sont faussement étiquetés ou difficiles à étiqueter, même pour des humains (Qu.6). Dans la figure 3.7, le texte le plus proche de la frontière de décision

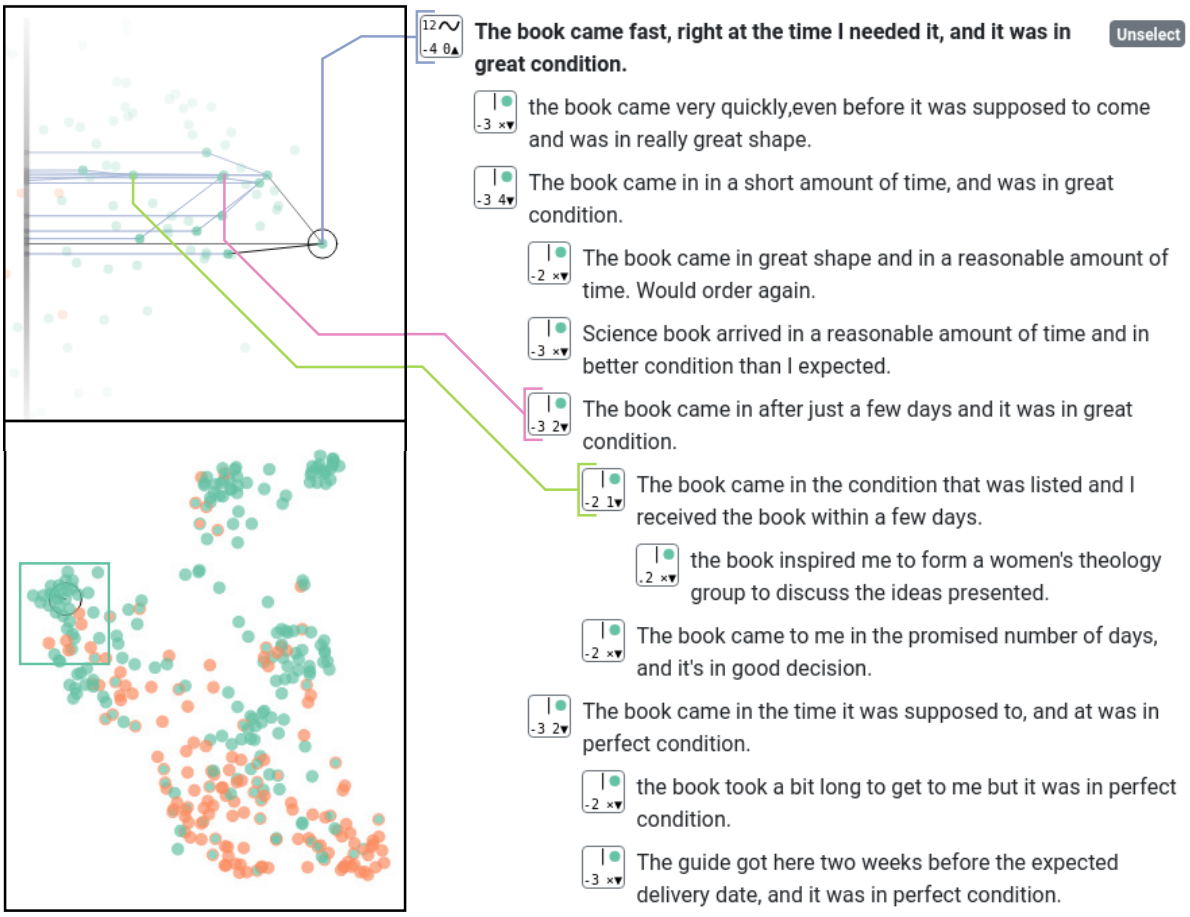


Figure 3.13 – *Différences de classification selon les nuances. En bleu : Le livre est arrivé en peu de temps, et était en excellent état (original en anglais : The book came in a short amount of time, and was in great condition). En rose : Le livre est arrivé après quelques jours seulement et il était en très bon état (original en anglais : The book came after just a few days and it was in great condition). En vert : Le livre est arrivé dans l'état indiqué et j'ai reçu le livre en quelques jours (original en anglais : The book came in the condition that was listed and I received the book within a few days).*

appartenant aux chemins de la donnée sélectionnée jusqu'à la frontière de décision semble positif, à l'exception de la dernière partie, dans laquelle le client déclare qu'il n'a pas reçu le produit (voir la troisième phrase dans la liste de textes). Au cours de notre exploration, nous avons observé que le réseau de neurones était vraiment efficace pour prédire le texte comme négatif lorsque le client n'a pas reçu le produit (Qu.5). Le score d'attention affiché dans la figure 3.7 (voir encadré rouge) montre également que le réseau s'est concentré sur la partie du texte traitant de la non-réception du livre. Pour vérifier si cette partie détermine réellement la classification, nous avons utilisé le formulaire du classifieur du texte d'entrée (section 3.5.2.2) et avons reclassé le texte après avoir supprimé la partie relative à la non-réception. La ligne jaune montre que la phrase est maintenant prédite comme "positive" (Qu.5).

Nous nous concentrons maintenant sur les mots les plus pertinents de la localité. L'un d'entre eux est "disappointed" (déçu). Sans surprise, les phrases contenant ce mot sont classées négativement (Qu.4). Cependant, nous avons trouvé une autre localité où ce mot a été traité très différemment (figure 3.14). Ce qui signifie que la présence

d'un mot n'est pas suffisante et que le réseau traite les textes de manière plus complexe (Qu.6). Nous avons trouvé également deux phrases (figure 3.15) qui partagent deux mots avec un score d'attention élevé : "nowhere" (nulle part, deux phrases dans la localité) et "boring" (ennuyeux, trois phrases dans la localité). Dans ces deux phrases, l'un de ces mots avait un score d'attention plus élevé que l'autre et ce n'était pas deux fois le même qui avait le score d'attention le plus élevé. Le réseau considère donc les mêmes mots différemment selon le contexte global d'un texte (Qu.4, Qu.5 et Qu.6).

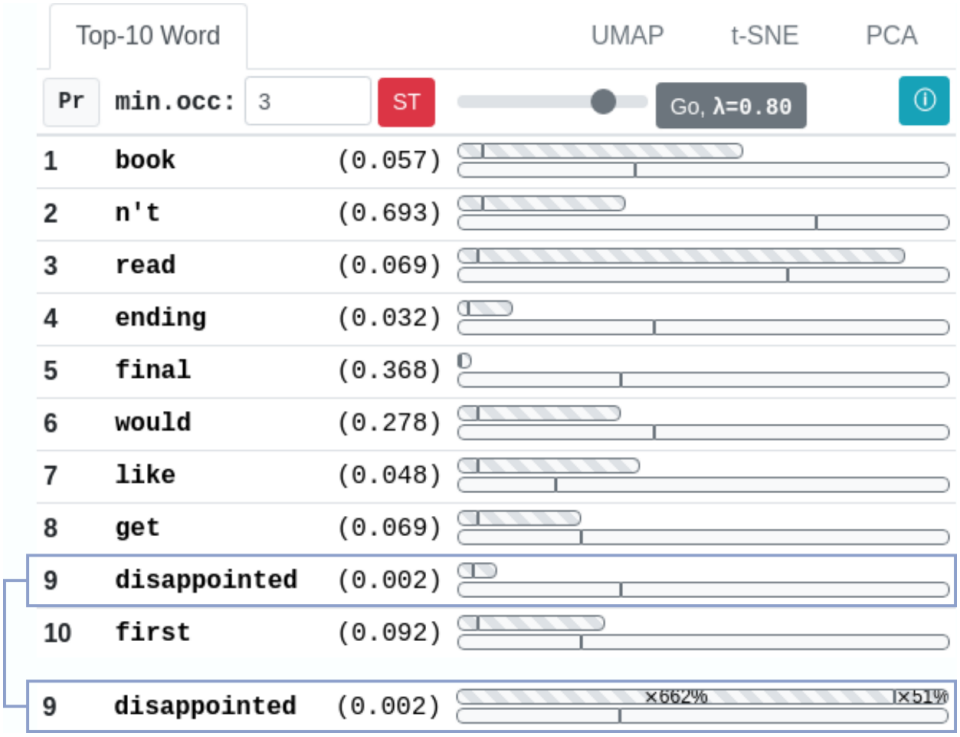


FIGURE 3.14 – *Identification d'un mot traité différemment dans une localité et dans l'ensemble du corpus. En bleu, la modification de la visualisation lorsque l'on clique sur le mot. On observe ici que les textes contenant le mot disappointed sont bien plus souvent classés du côté droit de la frontière de décision dans cette localité que dans l'intégralité du jeu de données.*

3.6.3.3 GPT-2 à tâche unique

Dans le jeu de données [ATS18 ; ATS17] proposé dans la troisième étude de cas, les données sont des titres d'articles étiquetés comme "fake news" ou infoxs en français (fausses informations, oranges, classées correctement du côté gauche de la frontière) ou "real news" ou informations (vraies informations, vertes, classées correctement du côté droit de la frontière). Nous avons inspecté les prédictions uniquement pour l'ensemble de test (8 972 entrées). Le réseau de neurones entraîné était une variante de GPT-2 [Rad+19] utilisée pour la classification de textes [3].

Nous avons d'abord inspecté la matrice de confusion (figure 3.16(a)). Elle montre que deux titres d'articles de "fake news" ont été prédits à tort et que 702 titres d'articles de "real news" ont été prédits à tort. Après avoir ordonné les lignes de la matrice du

3. https://huggingface.co/transformers/model_doc/gpt2.html

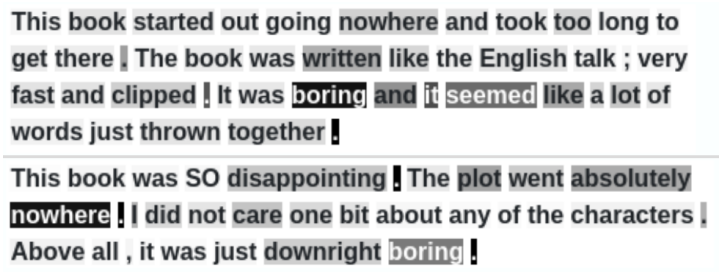


Figure 3.15 – *Score d'attention pour deux textes. On observe ici qu'un même mot peut avoir des scores d'attention différents en fonction du contexte. Un même couple de mots présent dans deux textes peut avoir un ordonnancement du score d'attention différent. Ici, le mot "boring" a un score différent dans les deux textes. Dans un seul des deux textes, il est le mot avec le plus haut score d'attention.*

diagramme de flux de la vue globale pour avoir les localités avec le plus grand nombre d'erreurs de classification (figure 3.16(b)) et des données réparties de manière égale autour de la frontière de décision (figure 3.16(c)), nous avons inspecté la localité N°200 parce qu'elle était en haut de la matrice des localités avec cet ordonnancement (Qu.3). Comme seules des données vertes apparaissent dans cette localité, toutes les entrées devraient être étiquetées comme "real news". Cependant, nous avons observé que près de la moitié des données étaient mal prédites (Qu.3). Nous avons alors choisi d'inspecter les techniques de réduction de dimension (figure 3.16(d)). Dans ce cas, t-SNE est la solution la plus efficace pour trouver une donnée entourée de nombreux voisins, y compris des données faussement et correctement classifiées (Qu.2). Nous avons choisi cette donnée et avons inspecté les textes associés (figure 3.16(f)). Il est difficile ici de comprendre pourquoi le réseau a classé ces textes comme "fake news". Toutefois, nous pouvons remarquer la forte incertitude associée aux prédictions (Qu.1). La plupart des données (*e.g.* l'élément à gauche dans la figure 3.16(e)) ont une incertitude proche de 0,1, ce qui est élevé. Lorsque nous comparons les aperçus de la figure 3.16(b) et de la figure 3.16(c), les distances à la frontière de décision pour les "fake news" prédites comme "fake news" et les "real news" prédites comme "fake news" vont d'au plus $(-)2$ pour cette localité à au moins $(-)12$ pour les localités 640, 648 et 650 (Qu.3). Pour confirmer cette observation pour l'ensemble du jeu de données, nous avons exploré les localités dans la vue globale et observé que les localités contenant des erreurs de classification classent les exemples à une distance d'au plus $(-)4$ de la frontière. Cependant, les localités ne contenant pas d'erreurs de classification descendent rarement en dessous d'une distance de $(-)6$. Cela signifie que la frontière de décision doit être placée plus près du côté des "fake news" pendant l'entraînement du réseau de neurones. L'ensemble de test n'est peut-être pas représentatif de la distribution réelle de l'ensemble de données, ou le réseau devrait s'entraîner davantage pour décaler la frontière de décision et ainsi obtenir de meilleures performances (Qu.6).

3.7 Discussions

Dans cette section, nous présentons les discussions concernant notre méthode. Ces discussions regroupent les interrogations autour des temps d'exécutions (section 3.7.1) et des étapes d'abstraction des données (section 3.7.2). Nous comparons également

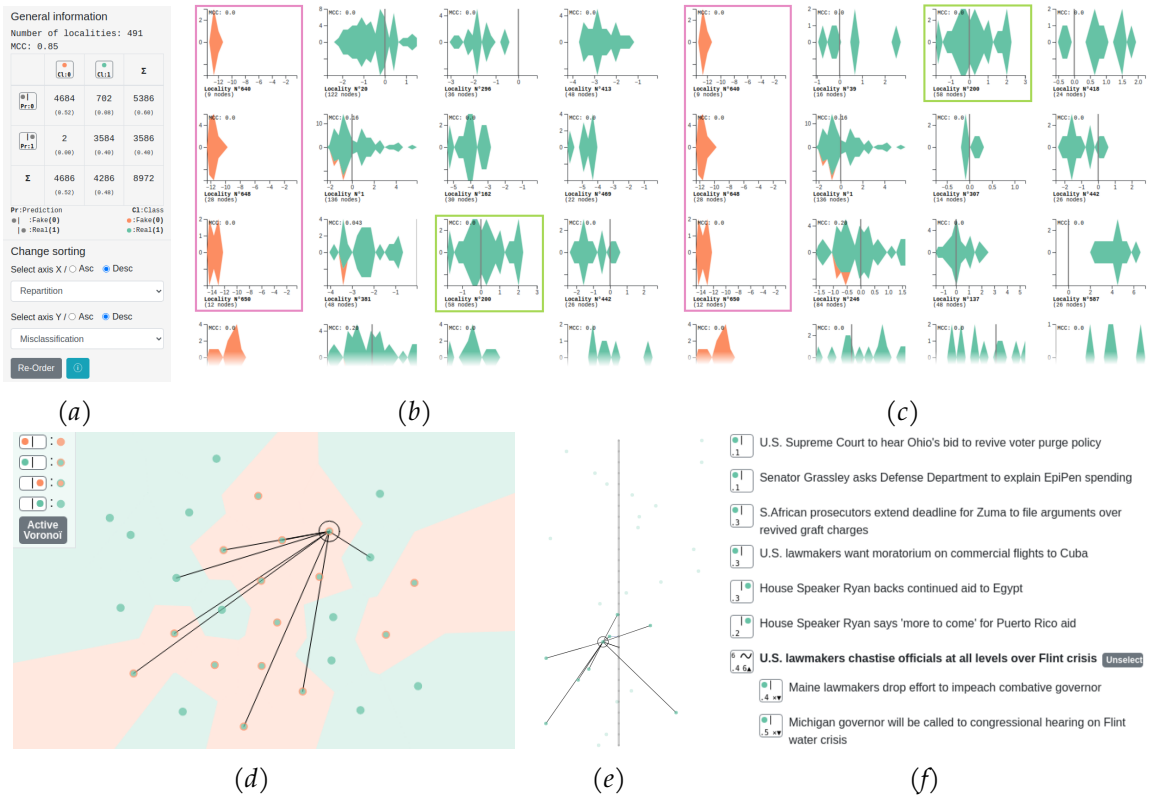


FIGURE 3.16 – *Vues et sous-vues utilisées dans la troisième étude de cas.* (***a***) *Matrice de confusion et ordre choisi pour explorer les localités. Nous observons que le réseau de neurones est meilleur pour prédire correctement les "fake news" (infox).* (***b***) *Vue globale montrant les localités avec le plus grand nombre de mauvaises classifications.* (***c***) *Vue globale montrant les localités ayant les données les plus également réparties autour de la frontière de décision (proche d'une loi normale).* (***d***) *Projection t-SNE dans l'espace 2D des textes appartenant à la localité N°200.* (***e***) *Visualisation de la frontière de décision de la localité N°200.* (***f***) *Listes des textes voisins des données sélectionnées (ici, tous les textes affichés sont des voisins directs du texte sélectionné ou ce dernier). Dans les encadrés roses sont présentées les localités N°640, N°648 et N°650. Dans les encadrés verts est présentée la localité N°200.*

notre méthode aux méthodes similaires (section 3.7.3) avant de justifier notre démarche vis-à-vis de l'interprétabilité (section 3.7.4). Enfin, nous présentons les limites de notre démarche (section 3.7.5) et les leviers qui peuvent conduire à une plus grande interprétabilité (section 3.7.6).

3.7.1 Temps d'exécution

Dans le tableau 3.3, nous présentons les temps d'exécution des différentes étapes de notre approche. L'étape qui prend le plus de temps est la division du graphe de proximité car elle supprime beaucoup d'arêtes. En fixant le nombre de voisins à 3 au lieu de 4, on réduit le temps de calcul. Une autre étape qui prend du temps est la simplification de la frontière, qui calcule beaucoup de plus courts chemins. Enfin, les deux étapes de visualisation peuvent prendre du temps, notamment les arrangements linéaires basés sur l'algorithme TSSA-Φ [RHT08]. Les étapes ont été exécutées sur un processeur ASUS, Intel Xeon avec 40 cœurs à 2,4 GHz, 126 Go de RAMDDR4 à 2400

MHz.

	Étapes	NN dim	SSA 50	GPT-2 1 024	AE RNN 1 024	Section
Abstraction	Projection		$\leq$ 5 sec.	$\leq$ 5 sec.	$\leq$ 5 sec.	3.4.2
	Création du graphe		15 sec.	12 sec.	21 sec.	3.4.3
	Division		31 min.	17 min.	11 min.	3.4.4
	Optimisation		3 min.	2 min.	17 sec.	3.4.5 - 3.4.6
	Connexion		6 sec.	6 sec.	9 sec.	3.4.7
	Simplification		12 min.	25 sec.	$\leq$ 5 sec.	3.4.8
Vis.	Données projetées		54 min.	150 min.	195 min.	3.5.2.1
	Données d'entrée		4 min.	5 min.	52 min.	

Table 3.3 – *Temps d'exécution des différentes étapes avec un ensemble de données de 10 000 entrées.*

3.7.2 Étapes d'abstraction de données

Les étapes d'abstraction des données visent à construire un ensemble significatif et utile de graphes de proximité pour toutes sorties possibles du réseau de neurones. Nous avons choisi de construire notre graphe de proximité avec la méthode UMAP plutôt qu'avec la méthode des k plus proches voisins (KNN) ou d'autres méthodes, car UMAP capture la topologie de la variété sous-jacente aux données et construit ainsi des localités porteuses de sens. L'utilisation du graphe de proximité UMAP garantit également que les poids des arêtes peuvent être utilisés pour modifier le graphe de proximité sans perdre la propriété de connectivité entre les sommets.

En fixant le nombre de voisins à 3 ou plus (4 par défaut dans nos travaux) dans la méthode UMAP, on obtient une composante connexe unique ou quelques grandes composantes connexes (section 3.4.3). La visualisation des localités qui sont très différentes des autres renseigne d'autant plus sur ses liens de voisinage que le nombre de voisins est élevé. Cela signifie que si une localité est très isolée dans l'espace, elle aura plus de voisins liés, car seules quelques arêtes sont effacées lors de l'étape de division (section 3.4.4). Au contraire, fixer un nombre de voisins trop élevé pousse l'étape de division à s'exécuter plus longtemps. Les cinq étapes qui suivent la création du graphe (section 3.4.4-3.4.8) garantissent que même si le graphe de proximité créé avec UMAP a une topologie très particulière, un ensemble de graphes de proximité utilisables est fourni.

La première étape de connexion des données d'entrée aux données de frontière relie uniquement les données initiales à leurs données projetées. Cela signifie que même si des données initiales qui n'appartenaient pas à la même composante y sont désormais rattachées, elles ne seront pas reliées entre elles et n'influenceront donc pas leur visualisation mutuelle dans le placement autour de la frontière (section 3.5.2.1). L'étape de connexion des données à leurs projetés utilise une fonction logarithmique en base 2. Nous pourrions utiliser une base inférieure ou même connecter chaque donnée réelle à sa projection. Notre choix est motivé par le fait que nous voulions garder le

graphe de proximité aussi proche que possible du graphe initial créé avec UMAP. Dans certains cas, le choix d'une base inférieure pourrait être utile, notamment si le nombre maximal de sommets ou d'arêtes est élevé ou si nous avons besoin de plus de liens entre les données d'entrée et les données de frontière dans la sous-vue de la frontière de décision. Enfin, les seuils utilisés dans l'étape de division dans la section 3.4.4 et la section 3.4.6 dépendent de la connaissance de l'ensemble de données. Dans nos expériences, nous utilisons 3 200 arêtes et 800 sommets comme seuils. Ces valeurs peuvent être utilisées par défaut. Des valeurs plus petites donnent des localités plus spécifiques au lieu de grandes localités et donc une explication plus locale et moins générale. Enfin, le jeu de données influence fortement le temps de traitement. Il n'est pas conseillé d'utiliser EBBE-Text avec un jeu de données de plus de 100 000 entrées.

3.7.3 Comparaison aux techniques alternatives

Le tableau 3.4 présente une comparaison, en termes de fonctionnalités, entre EBBE-Text et d'autres techniques alternatives. On constate qu'EBBE-Text offre, dans un seul outil, la plupart des possibilités offertes précédemment par les autres techniques. Dans cette section, nous nous concentrons sur certaines de ces possibilités afin de mettre en évidence les avantages et les limites de notre approche.

	Espace de représentation				Exploration des données				Applications						Contribution à la prédiction		
Travaux	Visualisation deux dimensions	Frontière de décision	Distances cohérentes	Comparaison des plongements	Construction de localités	Chemins à la frontière	Visualisation des données	Classification à la volée	Classification binaire	Classification multi-classe	Tous types de données	Tous types de classifieurs	Tous types de réseaux	Apprentissage actif	Comparaison de modèles	Comparaison des données	Visulisation de la contribution
EBBE-Text [Del+]	●	●	●	●	●	●	●	●	●			●	●			●	●
Ma *et al.* [MM21]	●	●	●	●	●				●	●	●						
Melnik [Mel02]	●	●			●					●	●	●	●		●		
Rodrigues *et al.* [Rod+19]	●	●	●	●					●	●	●	●	●		●		
Migut *et al.* [MWV15]	●	●	●						●		●	●	●		●		
Zhiyong *et al.* [ZC08]	●	●	●						●		●	●	●		●		
Ramamurthy *et al.* [RVM19]	●	●	●						●		●	●	●				
Smilkov *et al.* [Smi+16]	●			●			●				●	●	●				
Seifert *et al.* [SG10]	●		●						●	●	●	●	●	●			
Strobelt *et al.* [Str+19]	●						●		●			●	●			●	
Strobelt *et al.* [Str+17]	●						●			●					●	●	●
Heimerl *et al.* [Hei+12]												●		●			
Zhang *et al.* [Zha+19]							●	●		●					●	●	●
Vig [Vig19]							●		●	●		●					
Vig *et al.* [VB19]							●		●	●							●
Clark *et al.* [Cla+19]							●		●	●							●
Hoover *et al.* [HSG20]							●			●							●
DeRose *et al.* [DWB20]							●			●					●		●
Wang *et al.* [WTC21]							●			●					●		●

TABLE 3.4 – *Comparaison entre les fonctionnalités d'EBBE-Text et celles d'autres techniques alternatives.*

Visualisation de la frontière de décision : Migut, Worring et Veenman [MWV15] proposent la visualisation de la frontière de décision dans l'espace d'entrée à haute dimension, mais les distances à la frontière de décision sont sans signification. Zhiyong et Congfu [ZC08] proposent un algorithme pour trouver les données de la frontière de décision, mais les distances dans leur visualisation sont à nouveau moins significatives que dans la nôtre. Zhang et al. [Zha+19] utilisent des distances significatives pour comparer deux classifieurs, mais toute autre information sur les voisins ou le contenu des données est perdue. Rodrigues et al. [Rod+19] visualisent la frontière de décision des classifieurs en utilisant des techniques de réduction de dimension dans l'espace d'entrée des classifieurs. Deux des cinq techniques de réduction de dimension les plus efficaces qu'ils identifient pour visualiser la frontière de décision d'un réseau de neurones convolutif en classification binaire sont UMAP et t-SNE. Cependant, ils utilisent un ensemble de données linéairement séparable, ce qui n'est pas un ensemble de données réaliste. De plus, leur méthode ne permet pas l'exploration de différentes localités. En plus de notre vue de la frontière de décision, qui corrige cette lacune, notre visualisation inclut les techniques de réduction de dimensions précédemment citées dans nos vues de l'espace à dimensions réduites, fournissant ainsi des vues complémentaires qui permettent l'exploration des données à partir de différents points de vue. Dans notre expérience, nous observons que plus le réseau est efficace (avec un bon score sur les métriques de prédiction), plus la visualisation de la frontière de décision sous forme de ligne est efficace en utilisant les techniques classiques de réduction de dimension (ACP, t-SNE, UMAP). Melnik [Mel02] propose une analyse de connectivité des données dans l'espace d'entrée qui garantit qu'aucune frontière de décision n'existe entre deux données si elles appartiennent à la même région de décision. Différentes régions de décision sont calculées et peuvent être comparées par différents classifieurs tels que les réseau de neurones ou les SVM. Ramamurthy, Varshney et Mody [RVM19] proposent une méthodologie pour comparer les complexités des frontières de décision dans un modèle et donc la capacité de généralisation des modèles pour un ensemble de données donné. Enfin, Ma et Maciejewski [MM21] proposent un outil de visualisation similaire à EBBE-Text qui présente la frontière de décision en utilisant un SVM sur les données proches de la frontière de décision pour construire différents segments linéaires de la frontière de décision, avec mise en évidence de parties précises de la frontière de décision. La plus grande différence entre tous ces travaux et le nôtre est que nous construisons la frontière de décision dans l'espace de représentation des données construit par le réseau de neurones. L'espace ainsi construit est d'un plus grand intérêt pour explorer comment les réseaux de neurones représentent les données et construisent des localités significatives. De plus, nous pouvons calculer la distance réelle à la frontière de décision et explorer des localités porteuses de sens.

Classification multiclasse : EBBE-Text est limité à la classification binaire. L'étendre à la classification multiclasse est un problème intéressant pour des travaux futurs et cela soulèvera des questions nouvelles de recherche. En particulier, il semble que nous devrons recourir à l'utilisation d'une méthode "un contre tous" ou perdre l'information sur la similarité entre les données.

Tous les types de données : EBBE-Text est conçu pour les données textuelles. Les caractéristiques spécifiques aux textes sont la liste des textes et des descripteurs, et la

liste des mots les plus importants (section 3.5.2.2). Les autres caractéristiques ne sont pas spécifiques et peuvent être utilisées pour tous types de données. On peut imaginer utiliser le panneau central différemment pour explorer d'autres données telles que des images, des audios, *etc.*.

Tous les réseaux de neurones et classifieurs : comme mentionné dans la section 2.3.2.4, il existe plusieurs façons d'interpréter le fonctionnement interne d'un réseau de neurones et ses prédictions à l'aide de techniques de visualisations. Seq2seq-vis [Str+19] et LSTMVis [Str+17] se concentrent sur un RNN, tandis qu'EBBE-Text est plus générique et peut être utilisé avec n'importe quel réseau de neurones. Cependant, on peut préférer les informations spécifiques extraites pour les RNN fournies par des outils spécifiques car elles permettent une exploration de l'évolution des états cachés pour chaque mot entré. EBBE-Text peut également être utilisé avec n'importe quel classifieur qui représente les textes dans un espace de représentation en une étape et les classifie linéairement après celle-ci.

Active Learning : EBBE-Text ne propose pas de stratégie d'apprentissage actif à la manière de SEIFERT et GRANITZER [SG10] et HEIMERL et al. [Hei+12]. Il n'est pas encore possible d'étiqueter les données dans EBBE-Text. Il s'agit toutefois d'une fonctionnalité simple à intégrer. Par exemple, nous pourrions ajouter un bouton "changer l'étiquette" sous le descripteur de chaque phrase. Néanmoins, identifier les textes utiles à l'amélioration de la prédiction est difficile car ce ne sont pas forcément ceux qui ont les scores d'incertitude (distance à la frontière de décision) les plus élevés ou ceux qui sont mal classés [Bou16 ; GIG17 ; SS18]. De plus, afin de prendre en compte les nouvelles étiquettes, il serait nécessaire de ré-entraîner le réseau de neurones, de procéder à une nouvelle incorporation des textes et d'exécuter toutes les étapes décrites ci-dessus. Dans de nombreux cas, cela prendrait beaucoup trop de temps pour l'apport probable.

Visualisation de la contribution : dans notre approche, nous visualisons le score d'attention pour mettre en évidence la contribution des tokens à la prédiction. Il convient de noter que notre approche est cohérente avec d'autres scores de contribution. Auparavant, VIG [Vig19], VIG et BELINKOV [VB19] et CLARK et al. [Cla+19] ont proposé une visualisation de l'attention terme à terme. HOOVER, STROBELT et GEHRMANN [HSG20] dans un travail similaire ont ajouté une recherche du plus proche voisin des données sélectionnées dans le corpus textuel. DEROSE, WANG et BERGER [DWB20] proposent un affichage radial interactif dans lequel chaque cercle de tokens correspond à une couche donnée d'un modèle auto-attentif pour afficher les scores d'attention, ce qui permet une comparaison entre deux réseaux de neurones différents. WANG, TURKO et CHAU [WTC21] proposent des dispositions radiales (en grille et de force) et différentes vues pour montrer et comparer les poids d'attention des modèles auto-attentifs pour différentes entrées. Comme présenté dans la section 3.5.2.2, nous proposons d'afficher un score terme à terme du mot sur lequel l'utilisateur passe la souris. Notre méthode nécessite une action de la part de l'utilisateur alors que d'autres n'en nécessitent pas. Cependant, l'utilisation de méthodes alternatives dans EBBE-Text, telles que les graphiques à bosse ou les affichages radiaux, aurait apporté trop de bruit lorsque nous affichons de nombreux textes différents ou de longs textes.

3.7.4 Explications post-hoc ou locales

Les explications post-hoc ou locales visent à expliquer les prédictions, parfois sans suivre le même processus de raisonnement que le réseau de neurones ou sans relier directement les décisions aux entrées. Dans notre travail, nous n'avons pas cherché à savoir si et dans quelle mesure les explications étaient justifiables ou vraies [Lau+19]. Nous avons cherché à savoir si la confiance des utilisateurs dans le réseau de neurones utilisé [Lip18] augmentait à l'aide des explications locales. Le sentiment que le réseau se comporte d'une manière plus humaine augmente la confiance des utilisateurs dans le réseau. Par exemple, l'utilisateur peut voir que les quelques erreurs commises par un réseau correspondent à des données faussement étiquetées, que les scores d'attention sont cohérents pour les mots qui sont les plus importants pour les humains (voir l'étude de cas dans la section 3.6.3.2), ou que l'ordre dans lequel le réseau positionne les données, de la plus proche à la plus éloignée de la frontière de décision, est censé (voir l'étude de cas dans la section 3.6.3.1).

3.7.5 Limites de la visualisation et charge cognitive

En ce qui concerne la visualisation, selon le réseau utilisé et ses prédictions, les données de la sous-vue de la frontière de décision peuvent se chevaucher. Ce chevauchement est dû à la faible quantité de données de frontière liées aux données d'entrée. Il se produit lorsque le réseau de neurones a une grande confiance dans les prédictions et, par conséquent, le graphe de proximité construit par UMAP relie les données d'entrée aux données de frontière en utilisant le poids le plus faible. Cette spécificité fait que les liens entre les données d'entrée et les données de frontière n'existent que par l'intermédiaire de l'étape de connexion de frontière (section 3.4.5) et sont donc peu nombreux.

Dans les localités, certains réseaux de neurones divisent l'espace de représentation de telle sorte que les localités ne contiennent pas de données des deux classes. Les projections des données d'entrée sur la frontière de décision sont donc éloignées les unes des autres. On peut supposer que cela ne permet pas aux utilisateurs de comparer des données avec des prédictions différentes, mais la comparaison de données éloignées les unes des autres dans l'espace de représentation construit par le réseau de neurones n'a aucun sens. De plus, l'information selon laquelle le réseau de neurones sépare les données d'entrée via plusieurs hyperplans (*i.e.* l'hyperplan de la frontière de décision séparant les données d'entrée et l'hyperplan séparant leurs projections) donne un aperçu de la facilité de la tâche de classification pour le réseau de neurones.

En ce qui concerne la charge cognitive associée à l'utilisation d'EBBE-Text, les possibilités de tri dans la vue globale la réduisent pour la recherche de localités d'intérêt. Dans la vue d'une localité, nous pensons que la visualisation proposée réduit la charge cognitive, puisque presque toutes les sous-vues de la vue de localité sont proposées en tout temps. L'utilisateur n'a pas besoin de basculer entre plusieurs sous-vues pour obtenir des informations pertinentes sur la prédiction. Les deux seules fonctionnalités qui ne sont pas accessibles ensemble sont le tableau des mots les plus pertinents et la visualisation de l'espace de réduction des dimensions. Or, cela ne limite pas les utilisateurs dans leur recherche de données pertinentes, comme le montre l'évaluation

ICE-T dans la section 3.6.1. Pour finir, la connaissance et la compréhension du réseau de neurones utilisé sont essentielles pour interpréter correctement la signification des scores d'attention ou la pertinence des mots dans certains cas.

3.7.6 Vers plus d'interprétabilité

La recherche d'une meilleure interprétabilité des classifieurs pourrait conduire les utilisateurs vers des classifieurs transparents (section 2.3.1), *e.g.* les arbres de décision. Même si les réseaux de neurones sont moins transparents, leur structure peut être ajustée pour être plus interprétable. Alors qu'EBBE-Text offre un moyen puissant d'interpréter le fonctionnement interne d'un réseau de neurones, nous avons observé que les réseaux ayant effectué une tâche de reconstruction en plus de la tâche de classification sont plus interprétables par EBBE-Text. En effet, les localités ont tendance à contenir des données étiquetées différemment et se trouvant des deux côtés de la frontière de décision (figures 3.6 et 3.16(b)). Le choix d'un classifieur implique de trouver un bon compromis entre les performances et l'interprétabilité. L'ajout d'une tâche de reconstruction peut diminuer les performances mais augmente considérablement l'interprétabilité. Lorsque les textes sont proches dans la représentation construite par le réseau, l'utilisateur peut les comparer. Un réseau de neurones à tâche unique, en se concentrant uniquement sur la classification, a tendance à attribuer de plus grandes distances dans l'espace de représentation entre des textes étiquetés différemment, même s'ils sont proches les uns des autres. Cependant, on peut imaginer que dans l'espace de représentation d'un humain, des textes presque identiques, mais différents par le sentiment porté, devraient être proches. Dès lors, un réseau de neurones qui inclut une tâche de reconstruction construit un espace de représentation plus interprétable car plus proche de celui des humains.

Notre méthode garantit que dans la vue secondaire de la frontière de décision, les distances à la frontière de décision sont exactes et que les liens entre les données sont significatifs. UMAP produit des voisinages dans le graphe de proximité (section 3.4.3). Au cours des étapes suivantes, nous supprimons les arêtes qui sont moins importantes pour les sommets (petit poids) et les arêtes qui relient les clusters à l'intérieur de grandes composantes (section 3.4.6). Enfin, les liens sont uniquement créés entre les données initiales et leurs projections. La suppression et la création d'arêtes sont nécessaires pour explorer l'espace. À la fin des étapes d'abstraction des données et d'encodage visuel, la visualisation de la frontière de décision représente les liens les plus forts entre les données de UMAP, ainsi que la distance exacte à la frontière de décision, et produit le placement des données autour de la frontière de décision en fonction de ces deux informations (section 3.4.2).

3.8 Conclusions

Dans EBBE-Text, nous proposons une nouvelle approche visuelle pour aider à expliquer les prédictions des réseaux de neurones (et autres classifieurs) pour une tâche de classification binaire de textes. Elle est basée sur une exploration multi-échelle des frontières de décision de différentes localités de l'espace de représentation du

texte. Les méthodes actuelles ne permettent pas de visualiser le degré d'incertitude d'un réseau de neurones quant à ses prédictions, ni les distances et les chemins vers la frontière de décision. Notre méthode d'exploration innovante permet aux utilisateurs de rechercher des explications post-hoc en inspectant les éléments de données proches les uns des autres dans l'espace de représentation du texte et leurs distances à la frontière de décision, ainsi que les scores de contribution.

En ce qui concerne les perspectives, des travaux futurs pourraient porter sur les explications contre-factuelles justifiées [Lau+19] et sur la combinaison de notre méthode avec une démarche d'apprentissage actif [Set09]. Une exploration des différentes techniques de codage-décodage pour permettre la génération de textes sur la frontière de décision [SSB17] est également une perspective prometteuse pour augmenter l'explicabilité des réseaux de neurones. Enfin, permettre aux utilisateurs de comparer différents réseaux en même temps pour des étiquettes identiques proches est une perspective intéressante.

EBBE-Text s'adresse plutôt à des utilisateurs spécialistes ou avec une bonne pratique des réseaux de neurones. Dans les travaux suivants, présentés dans le chapitre 4, nous développons une nouvelle visualisation présentant les résultats d'un réseau de neurones dédié à une classification dichotomique sur plusieurs labels destinée à des utilisateurs non spécialistes.

CHAPITRE 4

CANCER-ANNOT & CANCER-VIS : EXPLORER DES CONCEPTS À PARTIR DES DONNÉES ISSUES DES MÉDIAS SOCIAUX

Sommaire

4.1 Introduction

Dans la contribution précédente, présentée dans le chapitre 3, nous avons introduit l'outil EBBE-Text pour la classification binaire de textes. Ce dernier est dédié à des spécialistes ou utilisateurs réguliers d'algorithmes d'apprentissage automatique. Pour ce type d'utilisateur, il est important de visualiser les prédictions pour contrôler, mieux comprendre les résultats des modèles et s'assurer qu'ils fonctionnent à une échelle globale comme locale. Dans ce nouveau chapitre, nous nous focalisons sur des utilisateurs non spécialistes de ces algorithmes. L'objectif est de proposer de nouvelles métaphores visuelles qui donnent une inuition sur le fonctionnement des classifieurs et augmentent ainsi l'interprétabilité des réseaux et la confiance des utilisateurs dans l'outil. Notre cas d'étude porte sur l'exploration des mentions d'interventions non-médicamenteuses (INM) dans les médias sociaux par des professionnels de santé et des patients souhaitant s'informer sur ces nouvelles pratiques dans le cadre de cancers. Dans ce travail, les choix de conception sont faits au regard de cas d'étude. Néanmoins, des propositions s'inscrivent dans un cadre plus général d'amélioration de l'interprétabilité des réseaux et peuvent donc être utiles pour d'autres applications.

Pour explorer efficacement des données issues des médias sociaux, nous devons être capables de produire des labels pour les catégoriser. Pour cela, nous avons proposé un premier outil de visualisation, nommé Cancer-Annot (figure 4.4), permettant de collecter et d'annoter manuellement des données qui alimentent des réseaux de neurones profonds. Nous avons également proposé des règles métier pour annoter automatiquement des volumes plus importants de données. Une fois les réseaux de neurones entraînés et efficaces sur les labels sélectionnés, l'exploration des données est initiée à l'aide d'un deuxième outil de visualisation : Cancer-Vis. Celui-ci a été conçu pour l'exploration des données issues des médias sociaux concernant les mentions du cancer et des INM associées. Il permet de visualiser les distributions des prédictions pour différents labels issus de la classification des publications. Cette visualisation intègre trois vues principales : (1) une vue des textes classifiés par les réseaux de neurones ; (2) une vue de la distribution des prédictions pour les labels à l'aide d'histogrammes et (3) une vue de la distribution pour deux à quatre labels à l'aide de deux cartes de chaleur. Toutes ces vues sont soumises à la possibilité de filtrer les données selon les résultats des prédictions sur des labels choisis.

Dans la section 4.2, nous motivons l'intérêt de notre cas d'étude. Dans la section 4.3, nous introduisons globalement la démarche poursuivie par chacun des outils développés. Dans la section 4.4, nous nous questionnons sur la démarche d'annotation et proposons l'outil Cancer-Annot pour accomplir cette tâche. Dans la section 4.5, nous présentons comment nos classifieurs ont été entraînés sur les données. Dans la section 4.6, nous listons les besoins des utilisateurs concernant l'exploration des données issues des médias sociaux et l'outil Cancer-Vis construit pour répondre à ce cas d'étude. Dans la section 4.7, nous présentons, à travers deux cas d'études, comment l'outil Cancer-Vis répond à notre problématique et à comment, plus globalement, notre démarche permet d'explorer efficacement les médias sociaux. Dans la section 4.8, nous discutons des choix faits. Enfin dans la section 4.8, nous concluons nos travaux. Ces derniers, initiés en fin de thèse, restent préliminaires.

4.2 Contexte

Dans le domaine de la santé, Internet est devenu la seconde source d'informations après les consultations chez un professionnel de santé. Aujourd'hui, 34,8% des personnes consultent Internet pour rechercher des informations sur leur santé [Bec+13]. Ce chiffre peut monter jusqu'à 61% aux États-Unis [Fox11]. La plupart des personnes utilisent des moteurs de recherche pour initier leurs requêtes. La plupart des liens retournés les dirigent vers des médias sociaux. Les témoignages, sur ces derniers, représentent environ 50% des échanges alors que les partages d'informations scientifiques ne représentent que 20% des échanges [Rom12]. Parmi les sites les plus visités ces dernières années figurent régulièrement les médias sociaux comme Facebook, Instagram, Reddit ou Twitter. Les médias sociaux sont donc des espaces d'échange très utilisés où les patients, sous couvert d'anonymat, relatent librement leurs expériences et s'informent sur leur santé. Ils permettent ainsi une grande liberté d'expression et sont un moyen de communication populaire parmi les patients atteints de cancer pour partager leur vécu de la maladie, rechercher facilement des informations et obtenir du soutien. En France, la lecture de certains forums dédiés au cancer est même recommandée par les principaux organismes impliqués dans la recherche contre le cancer tels que l'INCa et la Ligue Contre le Cancer.

L'objectif de notre démarche dans ce chapitre est d'aider les patients et les professionnels de santé à mieux comprendre comment les patients utilisent ces INM (une activité physique régulière, un recours à un régime nutritionnel approprié, une utilisation d'autres produits, *etc*). Pour cela, nous allons les guider à l'aide d'une collection de données textuelles extraites de médias sociaux en ligne (Reddit, voir figure 4.1, et Twitter, voir figure 4.2). Nous limitons nos travaux à la thématique des cancers. En effet, on estime à 382 000 le nombre de nouveaux cas de cancers (incidence) et à 157 400 le nombre de décès (mortalité) en 2018 en France (INCa [1]). Les cancers de la prostate, du sein, du côlon-rectum et du poumon sont les plus fréquents. Les cancers sont très souvent associés à de nombreux symptômes fonctionnels qui détériorent la qualité de vie et provoquent une grande détresse psychologique pour les patients mais également pour les aidants [Sal+09]. S'intéresser alors aux causes de la détérioration de la qualité de vie comme les symptômes de la maladie ouvre une piste dans l'objectif d'aider les patients à mieux vivre leur maladie. Par ailleurs, les professionnels accompagnant les patients sont de plus en plus enclins à proposer des soins de support et l'utilisation d'INM en complément des traitements biologiques plus classiques du cancer [CN19].

TAPI NZALI [Tap20] a montré l'intérêt d'analyser les textes des médias sociaux relatifs au cancer pour s'informer sur la qualité de vie des patients et sur leur accès aux INM. Ce type d'analyse représente une alternative aux outils déclaratifs classiques (questionnaires, enquêtes). Les enquêtes et les études utilisant des questionnaires validés ou non ou des entretiens par téléphone amènent des biais de réponse connus (*e.g.* désirabilité sociale, réponse pré-établie) qui peuvent sous-estimer des phénomènes émergents. S'intéresser aux données issues des médias sociaux permet la prise en compte des véritables déterminants de la qualité de vie et l'usage réel des INM. La surveillance et la recherche sur les (més)usages de ces nouveaux vecteurs de communication sont en-

1. https://www.oncorif.fr/panorama-des-cancers-en-france-edition-2021/

Figure 4.1 – *Exemple de message issu du média social Reddit.*

Figure 4.2 – *Exemple de message issu du média social Twitter.*

couragées par les agences (*e.g.* INCa), les autorités de santé (*e.g.* HAS) et les fondations (*e.g.* Ligue Contre le Cancer). Le contenu de ces médias et leur volumétrie peuvent dire beaucoup sur les véritables motifs et les usages des patients traités pour un cancer, en particulier, sur leurs symptômes et les effets secondaires post-traitements comme la perte de cheveux après une chimiothérapie. L'intérêt pour ce type d'approche par les soignants témoigne de leur volonté d'impliquer les patients dans l'émergence de nouvelles pratiques directement issues de leurs préoccupations. L'intérêt médical des connaissances présentes dans les médias sociaux est donc désormais avéré. Toutes les informations disséminées dans ces médias peuvent être utilisées comme un vaste réseau de capteurs pour la modélisation de la santé publique à l'échelle de la population [Las+13]. Par exemple, de nombreux travaux ont exploité les médias sociaux pour analyser la propagation ou l'émergence des maladies [SKS12 ; Lop+21 ; Esp+21], pour capturer les liens entre les symptômes et les effets indésirables associés à des traitements [PE11], pour analyser la qualité de vie après un cancer du sein [Tap20] ou pour suivre des maladies en temps réel [Bar17]. Ces travaux fournissent des preuves solides qu'il existe un réel "signal" dans les médias sociaux [Las+13]. Ce dernier peut être exploité pour différentes applications liées à la santé.

Un verrou important lorsque l'on cherche à analyser des données textuelles relatives à la santé est que la plupart des méthodes ont été créées et appliquées sur des types de textes particuliers. Concevoir des méthodes destinées aux textes produits dans les médias sociaux est loin d'être trivial et ceci pour différentes raisons : les messages sont écrits suivant des normes peu contraintes (*e.g.* taille variable des messages, syntaxe et

orthographe libres) relevant de genres textuels émergents encore peu étudiés que ce soit au niveau sémiotique (*e.g.* émoticônes, abréviations, sigles, marqueurs d'emphase, vocabulaire sociolectal) ou au niveau sémantique (positionnements énonciatifs et dialectiques, modalités temporelles et aspectuelles, thématiques, etc.). Par ailleurs, si les éléments médicaux à rechercher sont généralement bien connus comme les facteurs de risque, leur expression symptomatique personnelle et intime dans les médias sociaux est extrêmement variable. De ce fait, impliquer des humains pour annoter ces textes est essentiel. Ceux-ci, au fur et à mesure de l'annotation, développent une expertise vis-à-vis des termes spécifiques utilisés dans les médias sociaux. Pour finir, l'important volume des textes est à considérer pour un passage à l'échelle des méthodes automatiques.

Les outils d'intelligence artificielle sont désormais matures pour raisonner sur les données issues des médias sociaux. Entre autres, les développements récents dans le domaine de l'apprentissage profond [Rav+16] permettent d'envisager différents types d'analyses, comme l'extraction d'information (mots-clés, entité nommées, relations…) à partir des textes [Min+17], l'annotation et la classification puis le raisonnement à partir de ces méta-données. Si ces outils sont très performants, ils souffrent du problème d'explicabilité [Gui+18b ; Zha+21 ; Hua+20] pour assurer la sécurité, limiter les problématiques liées à la discrimination [Mit19 ; Hag20], voire tout simplement éliminer le doute lors de l'interprétation des textes [DK17].

L'objectif de recherche principal de ce chapitre consiste à construire un outil d'exploration des messages issus des médias sociaux. Cet outil propose notamment une boîte à outils permettant de contextualiser en vue d'expliquer les prédictions issues de classifieurs entraînés sur les données de ces médias sociaux. Exploiter les données textuelles de plus en plus massives du web liées à la santé est utile pour étudier les déterminants de la qualité de vie et l'utilisation des INM dans le cas de patients atteints d'un cancer. Nous présentons globalement notre approche dans la section suivante.

4.3 Démarche globale

La figure 4.3 présente la démarche générale de l'approche intégrant (1) la récolte des données et leur étiquetage par des humains, (2) l'étiquetage des données selon des règles automatiques, (3) la classification binaire des données à l'aide de modèles auto-attentifs et (4) l'exploration de données extraites de médias sociaux classifiées par les modèles binaires. Pour les étapes (2) et (4), deux outils ont été produits et sont décrits dans la suite de ce chapitre.

Récolte et annotation des données : Deux stratégies ont été mises en place. Tout d'abord, un premier outil (Cancer-Annot) a été conçu pour collecter et annoter des messages issus des médias sociaux. La consigne donnée aux annotateurs a été d'exporter les données liées au cancer et à l'utilisation d'INM à l'aide de mots-clefs puis de les annoter selon différents labels prédéfinis. Lors de nos explorations préliminaires, nous avions déjà identifié ces labels ainsi qu'une liste d'INM qui a guidé les annotateurs au début de leurs démarches. Ensuite, afin de disposer de plus de données, des règles métier d'annotation, relatives à la provenance et au contenu des messages, ont été

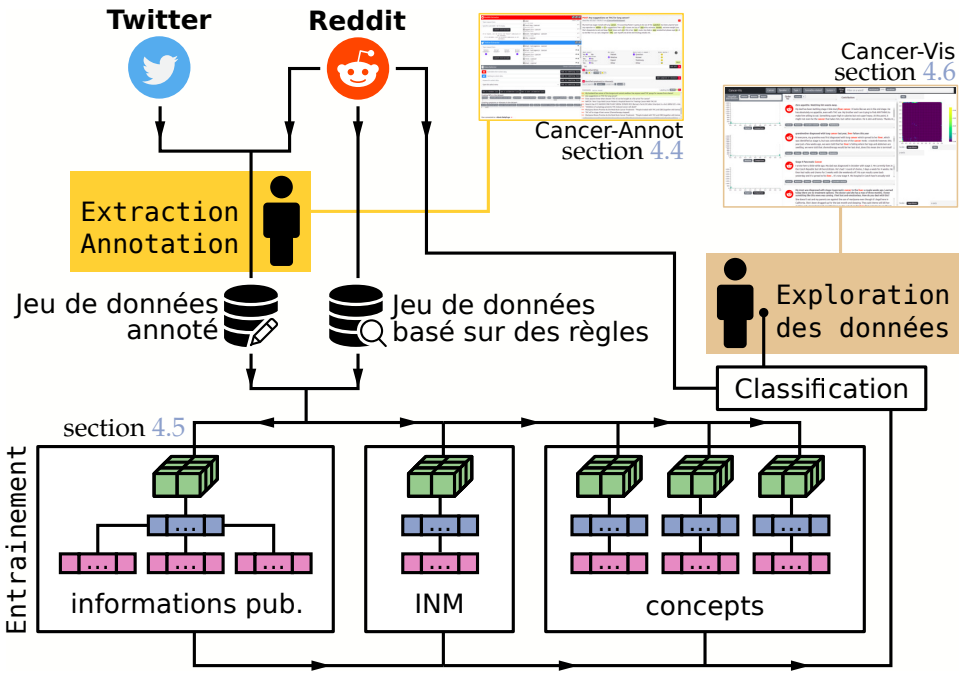


Figure 4.3 – *Étapes menant à l'exploration des données issues des médias sociaux. Les zones colorées en jaune et en doré correspondent aux étapes qui nécessitent l'utilisation d'outils de visualisation de la part d'un ou plusieurs humains. Le premier outil est l'outil d'extraction et d'annotation des données des médias sociaux (Cancer-Annot). Le deuxième outil est l'outil d'exploration des données issues des médias sociaux classifiées par les modèles entraînés (Cancer-Vis).*

utilisées.

Entraînement des réseaux de neurones : une fois les données annotées, l'objectif de l'étape suivante a été d'apprendre un modèle supervisé puis de produire des prédictions sur des données non annotées. Les modèles auto-attentifs, tels que décrits dans la section 2.2.2.4 de l'état de l'art, ont été choisis du fait de leurs performances. L'entraînement est effectué à l'aide des deux jeux de données présentés précédemment. Ces modèles prédisent un ou plusieurs labels. Une procédure d'entraînement utilisant successivement les deux jeux de données a été choisie.

Visualisation : à la suite de cet entraînement, l'utilisateur visualise, filtre et consulte les données à l'aide de l'outil Cancer-Vis de manière à répondre à ses questions concernant les INM. Il peut comparer les prédictions pour différents labels, mettre en lien ces labels et obtenir des informations sur les distributions des prédictions. Il est assisté dans l'exploration et est guidé vers des concepts d'intérêt, à savoir dans notre cas d'étude les INM.

Dans les sections suivantes, nous allons décrire précisément les trois étapes précédentes.

4.4 Cancer-Annot : Outil de récolte des données

Lors d'une première exploration qualitative des données, nous avons observé qu'au-delà de la simple identification des mentions d'INM ou du cancer, d'autres facteurs revêtent une grande importance. Nous adaptons donc, pour l'annotation, la grille proposée par PAGANELLI, CLAVIER et al. [PC+14] caractérisant les types d'informations véhiculés par les messages postés sur les médias sociaux à partir d'une typologie testée précédemment sur des forums de discussion de santé traitant du VIH [Mer+18].

4.4.1 Caractérisation du problème

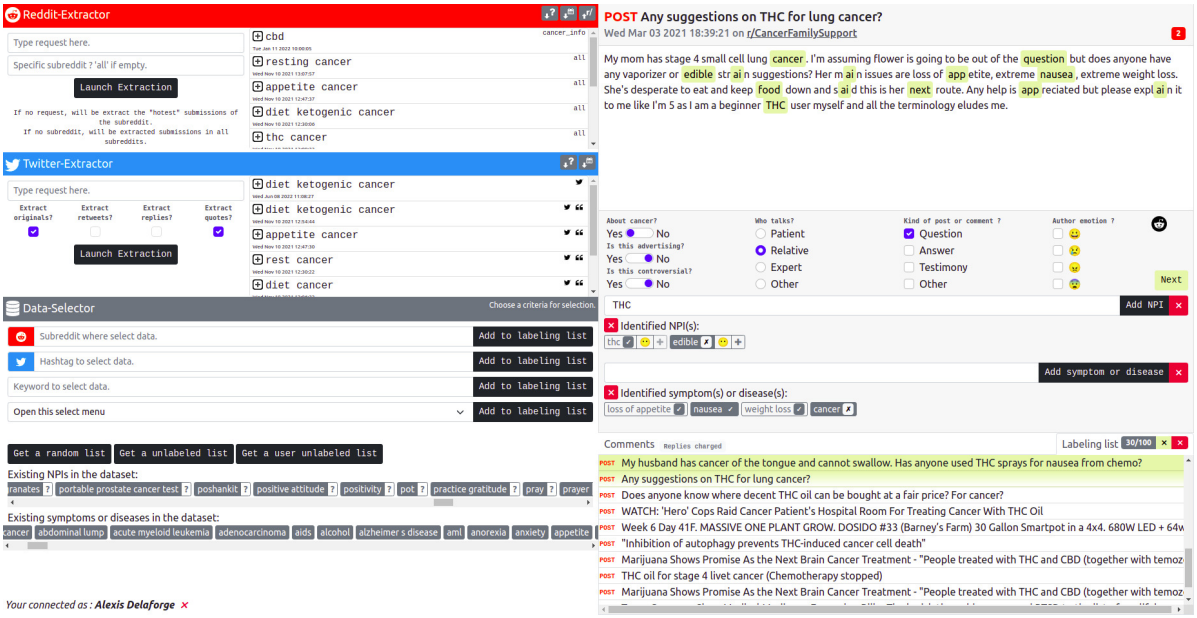


FIGURE 4.4 – *Vue globale de l'outil d'annotation des données issues des médias sociaux.*

Dans la première phase, il est nécessaire d'identifier quels sont les labels utiles à la compréhension des mentions d'INM dans le cadre du cancer en accord avec la méthodologie proposée par PAGANELLI, CLAVIER et al. [PC+14]. Ce choix est réalisé en explorant les données de Reddit et Twitter. Le premier label (CANCER) choisi est celui qui stipule si une publication est relative au cancer. Rapidement, nous observons que beaucoup de publications ont un objectif mercantile. L'objectif de notre démarche étant d'avoir des avis honnêtes de patients ou de proches, nous choisissons d'identifier les publications commerciales via un label dichotomique (COMMERCIAL). Il peut également être pertinent d'identifier le LOCUTEUR [PC+14]. Pour cela, nous créons un label avec quatre modalités : PATIENT, PROCHE, EXPERT et AUTRES. La modalité EXPERT concerne toute personne qui, par son métier, est amenée à fréquenter, traiter ou conseiller les patients et leurs proches. Une autre particularité des publications est qu'elles ont différents objectifs. Certaines sont des questions et obtiennent des réponses, d'autres sont des témoignages. Nous construisons donc un label TYPE avec comme modalités : QUESTION, RÉPONSE, TÉMOIGNAGE, plus une modalité AUTRES dans le cas ou aucune ne correspond [PC+14]. Dans les publications, nous identifions également différents types d'émotions. Par exemple, certaines parlent de l'effet positif des INM ou de

rémissions et évoquent des émotions heureuses. D'autres correspondent à des émotions plus négatives comme la colère, la peur ou la tristesse. Nous construisons un label ÉMOTION contenant ces quatre modalités : HEUREUX, TRISTE, EN COLÈRE, APPEURÉ. Enfin, les publications sur les médias sociaux mènent à de nombreuses discussions. C'est pourquoi, pour traiter les opinions polarisées, nous ajoutons un label dichotomique pour annoter la CONTROVERSE.

Une fois les labels à modalités finies produits, nous identifions les publications évoquant des INM. Nous créons un label INM qui peut prendre n'importe quel type de valeur pour, au cours de l'annotation, produire une liste d'INM. Nous procédons de même avec les symptômes (SYMPTÔME).

4.4.2 Questions des utilisateurs pour l'annotation

Après avoir produit, à l'aide d'une exploration qualitative des données, les labels à renseigner au cours de l'annotation, nous définissons une liste de six questions auxquelles nous devons répondre pour annoter efficacement les données issues des médias sociaux. Ici, les questions sont présentées de manière générale comme si elles ne s'appliquaient pas spécifiquement à notre cas d'étude, mais plus globalement à des données issues des médias sociaux.

Qu.1 Quelles sont les publications contenant des mots choisis par l'utilisateur dans les médias sociaux ?

Qu.2 Comment les autres utilisateurs répondent à une publication et une discussion découle-t-elle de cette dernière ?

Qu.3 Quelles sont les publications déjà annotées ?

Qu.4 Quelles sont les concepts déjà identifiés ?

Qu.5 Quelle est la page web liée à cette publication ?

Qu.6 Existe-t-il des occurrences de concepts déjà annotés dans le texte de la publication ?

En complément de ces questions, nous souhaitons ajouter qu'un concept est une idée générale d'un objet ou d'un ensemble d'objets ayant des caractères communs. Ici les concepts retenus pour notre cas d'étude appartiennent aux INM ou aux symptômes.

4.4.3 Besoins identifiés pour l'outil

Pour répondre à ces questions, nous identifions six besoins auxquels notre outil d'annotation doit répondre :

Be.1 Extraire les données depuis un média social sur critères de mots-clefs pour Qu.1 et Qu.6 ;

Be.2 Apporter le contexte de publication et le contexte de discussion autour des publications pour Qu.2 et Qu.5 ;

Be.3 Afficher le texte, le titre, les mots d'intérêt et les commentaires de la publication pour Qu.2, Qu.5 et Qu.6 :

Be.4 Afficher les annotations déjà effectuées pour une publication et les publications déjà annotées et annoter les données Qu.3 et Qu.6 ;

Be.5 Afficher les textes ayant des INM ou des symptômes identifiés pour Qu.4 ;

Be.6 Construire des listes de données à annoter pour Qu.3.

4.4.4 Outil d'annotation des données

Dans l'outil d'annotation, nous divisons la vue en deux. À gauche se trouve une partie relative à l'extraction des données depuis les médias sociaux et à la construction de listes de données à annoter (figure 4.5). À droite se trouve la zone d'annotation qui concentre l'intégralité des informations sur la publication, les commentaires associés et des boutons pour annoter (figure 4.6).

4.4.4.1 Une extraction ciblée

Le panneau gauche, servant à l'extraction des données et la construction de listes de données à annoter, répond aux besoins Be.1 et Be.6. Il est constitué de deux parties, elles-même divisées en deux parties. La partie haute (en vert dans la figure 4.5) sert à cibler les données à extraire des médias sociaux et répond au besoin Be.1. La partie basse (en rose dans la figure 4.5) est utilisée pour construire des listes de données à annoter selon plusieurs procédures et répond au besoin Be.6.

Exportation des données : la zone d'exportation des données, partie supérieure du panneau de gauche, est dédiée à l'extraction des données Twitter ou Reddit. Reddit est un média social fonctionnant sur le modèle des forums dans lequel il existe des groupes de discussion, appelés subreddit, où les utilisateurs peuvent visionner, poster ou commenter des publications en lien avec le groupe. Il existe par exemple un subreddit r/cancer où les utilisateurs discutent de leurs ressentis, témoignent de leur parcours, demandent des informations à propos du cancer. Twitter est également un média social. Ses utilisateurs ne postent pas dans des fils de discussion dédiés. Twitter se base sur des hashtags et des mentions pour structurer les discussions. Il prend la forme du caractère # suivi du texte du hashtag. Tous les messages le contenant sont regroupés. Enfin, les commentaires sous une publication mentionnent la publication et l'auteur de la publication.

La zone d'extraction des données possède une sous-zone affectée à chaque média social. Dans la première, consacrée à Reddit, l'utilisateur choisit le ou les termes à chercher dans une zone de saisie et le subreddit. Une fois la recherche lancée à l'aide du bouton, une liste apparaît à droite de la zone d'extraction. Cette liste contient les dates et les informations concernant l'exportation des données et peut être triée selon plusieurs critères. Lorsque l'utilisateur lance la recherche, si aucun subreddit n'est précisé, l'exportation s'effectue sur l'intégralité de Reddit indépendamment des subreddits. Une exportation dure quelques secondes si un subreddit est renseigné et peut durer environ une minute si aucun subreddit n'est renseigné. Une limite à

Figure 4.5 – *Panneau gauche de l'outil d'annotation des données servant à l'exportation des données issues de Twitter et Reddit. Dans l'encadré vert se trouvent les outils d'extraction des données et dans l'encadré rose se trouvent les outils construisant les listes à annoter. De haut en bas : le panneau d'extraction des données issues de Reddit puis de Twitter puis le panneau de sélection à annoter.*

l'exportation est fixée à 100 publications. Lorsque l'exportation est terminée, il est possible, à l'aide d'un bouton, d'ajouter l'exportation choisie à la liste de données à annoter (section 4.4.4.2). La sous-partie de Twitter fonctionne de manière similaire à celle de Reddit si ce n'est qu'il n'y a que la zone de termes à renseigner, et que celle-ci peut contenir des hashtags. De plus, on peut choisir de n'inclure que les tweets originaux en laissant de côté les réponses aux tweets (ou commentaires). On peut également choisir d'inclure ou non les tweets citations. Ceux-ci citent un tweet original et ajoutent un texte en contexte. Les tweets citations sont différents des réponses aux tweets. Ces deux zones répondent au besoin Be.1.

Construction des listes à annoter : pour construire la liste des données à annoter, il existe plusieurs méthodes. L'utilisateur peut sélectionner des données aléatoirement ou récupérer les résultats d'une exportation. Il peut sélectionner uniquement les données d'un subreddit, celles contenant des hashtags et des mots-clefs. Il peut préciser qu'il souhaite annoter des messages contenant des mentions de concepts (section 4.4.2) déjà annotés par d'autres utilisateurs. Par exemple, un utilisateur peut chercher toutes les

publications pour lesquelles l'INM "pray" a été identifiée. Cette manière de construire une liste à annoter répond spécifiquement au besoin Be.5. Toutes ces manières de construire des listes répondent au besoin Be.6.

4.4.4.2 Une annotation assistée

Le panneau à droite de l'outil et permettant l'annotation des données contient trois parties (figure 4.6). La partie du dessus, en jaune dans la figure 4.6, contient les informations concernant la publication et répond au besoin Be.3. La partie centrale, en bleu dans la figure 4.6, contient les boutons et zones d'annotation des données et répond au besoin Be.4. La partie basse, en orange dans la figure 4.6, contient la liste des données à annoter et les commentaires concernant la publication et répond au besoin Be.2.

FIGURE 4.6 – *Vue du panneau à droite de l'outil d'annotation des données. De haut en bas : le titre de la publication et le texte de la publication dans l'encadré jaune, la zone d'annotation et ses boutons et zones de saisie des concepts dans l'encadré bleu et la liste des publications à annoter ou des commentaires dans l'encadré orange.*

La publication : cette partie, présentée dans l'encadré jaune de la figure 4.6, renseigne sur toutes les informations relatives à la publication. Elle informe sur le type de

publication : message ou commentaire pour Reddit et tweet, retweet, commentaire ou citation pour Twitter. Ensuite, elle donne le titre ou la publication liée dans le cas de commentaires. La date de la publication et le subreddit (le cas échéant) sont ensuite renseignés, ainsi que le score de la publication, à savoir, le nombre de j'aime dans le cas de Twitter ou la différence entre les votes positifs et négatifs dans le cas de Reddit. Le texte de la publication est présenté dans la partie la plus grande. Ce texte contient des mots balisés relatifs aux termes déjà annotés dans d'autres publications (voir balises jaunes dans l'encadré jaune de la figure 4.6). Cela facilite la sélection du texte car cliquer sur ces balises sélectionne automatiquement le texte pour remplir ensuite, à la manière d'un copier/coller, la zone d'annotation. De plus, cela met en valeur les mots importants au début des annotateurs.

L'annotation : la partie labels est composée de tous les labels à renseigner. Elle se trouve au centre de la partie annotation, dans l'encadré bleu de la figure 4.6. Les labels dichotomiques se renseignent à l'aide de boutons switch. Ces boutons sont en position centrale par défaut et peuvent basculer vers le "Yes" ou le "No" au clic de l'utilisateur. Une fois un label renseigné, il ne peut pas revenir en position centrale. Ensuite, deux zones de saisie sont présentes pour les concepts à identifier. Il suffit de remplir la zone de saisie puis de cliquer sur le bouton "Add" pour ajouter l'INM ou le symptôme. Une fois cet ajout effectué, il est renseigné sous la barre de saisie à l'aide d'un badge gris (petit conteneur avec du texte ou autre à l'intérieur). Un second bouton permet de récupérer le texte sélectionné et de remplir la zone de saisie pour accélérer le processus d'annotation. Le badge d'ajout sous la barre de saisie donne des informations dans le cas où un autre annotateur a déjà identifié l'INM ou le symptôme. Ce badge permet également d'ajouter des informations supplémentaires à l'annotation dans le cas des INM. Ces informations concernent les effets possibles de l'INM ou l'opinion de l'auteur de la publication concernant l'INM (figure 4.7).

Contexte et liste à annoter : dans la partie inférieure de la partie annotation (encadré orange de la figure 4.6) se trouve un encart à deux onglets. Le premier onglet donne la liste des commentaires de la publication ou les publications liées à la publication (retweet, citation) sous une forme hiérarchique. L'annotateur a accès aux discussions découlant de la publication. Par défaut, lors de l'extraction d'une publication, seuls trois commentaires sont exportés. Un bouton permet d'avoir accès à l'intégralité des commentaires après une extraction complémentaire. Le deuxième onglet donne la liste des données à annoter où l'on peut observer les données déjà vues lors de la session d'annotation mais aussi les données déjà annotées. On peut, au choix, afficher le premier ou le deuxième onglet grâce aux boutons supérieurs. Pour le bouton concernant la liste à annoter, il donne la possibilité de vider totalement la liste ou d'y retirer les données annotées au cours de la session.

4.4.5 Résultats

Dans les deux sections suivantes, nous présentons les résultats de l'exportation et de l'annotation et justifions la construction d'un deuxième jeu de données. Nous présentons également les résultats concernant ce dernier.

FIGURE 4.7 – *Confirmation et infirmation de présence d'une INM (`edible` et `THC`) de la part de l'utilisateur avec l'opinion du locuteur et les possibles effets. Le `THC` est une molécule, et `edible` doit faire référence au `THC` consommable sous forme de cachet ou gélule. (a) Infirmation de la présence de `edible` dans la publication. (b) Confirmation de la présence de `edible` avec un avis mitigé du locuteur. (c) Confirmation de la présence de `edible` avec un avis positif du locuteur. (d) Confirmation de la présence de `edible` avec un avis négatif du locuteur et l'ouverture de la saisie de l'effet possible de `edible`. On peut remarquer ici que `edible` n'a pas été identifié par un annotateur alors que `THC` oui. `THC` n'est pas confirmé par l'utilisateur. Un clic sur le bouton "Add NPI" griserait alors tout le badge `THC` signifiant qu'il a été annoté par d'autres annotateurs et par l'annotateur connecté.*

4.4.5.1 Données annotées manuellement

L'exportation des données a permis d'obtenir un jeu de données de 1 635 publications annotées au moins sur un des labels (voir tableau 4.1). Lors de l'annotation, 4 annotateurs anglophones et francophones sur un volume global de 170 heures ont annoté les données. Des disparités de temps d'annotation ont existé entre les annotateurs de plus ou moins 25 heures. Les objectifs d'exportation des données des annotateurs étaient de : (1) découvrir des INM ; (2) concentrer leur annotation sur les données d'intérêt ; (3) exporter des données relatives aux nouvelles INM découvertes dans les publications et les discussions.

Dans le tableau 4.1, on peut observer que les données sont réparties de manière équilibrée entre Twitter et Reddit, avec plus de commentaires annotés issus de Reddit. Lors de l'exploration qualitative, nous avions déjà observé que les publications Reddit étaient plus intéressantes que les discussions Twitter. Il semblerait que les annotateurs ont également préféré ce média qui a été le plus requêté et annoté. Pour chaque commentaire Reddit annoté, il existe entre 3 et 4 publications annotées alors que pour un commentaire Twitter, il existe environ 10 publications annotées. Les commentaires Reddit sont environ trois fois plus souvent annotés que les commentaires Twitter.

	Données labellisées		Données non labellisées	
	Publications	Commentaires	Publications	Commentaires
	669	198	16 894	1.5m
	867			
	698	70	6 782	117
	768			
	1 635			

Table 4.1 – *Résultats de l'annotation par les annotateurs. On observe une part similaire de données annotées issues de Twitter et Reddit. Néanmoins, les annotateurs ont en proportion bien plus annoté les commentaires Reddit que les commentaires Twitter.*

Parmi les données annotées, il existe de grandes disparités dans les résultats de l'annotation pour les différents annotateurs et entre les types de labels.

La figure 4.8 présente les résultats pour les **labels dichotomiques**. On note un accord quasi-systématique entre les annotateurs. On voit également que la plupart des données annotées ne sont pas considérées comme commerciales ou comme générant une controverse. De même, la majorité des données annotées concerne le cancer. Sur 1 635 données annotées sur au moins un label, 1 547, 1 530 et 1 520 ont été annotées sur les labels CANCER, COMMERCIAL et CONTROVERSE. On peut en conclure, que les labels dichotomiques ont été annotés très régulièrement et avec un grand accord entre les annotateurs. Ces résultats sont encourageants pour obtenir de bonnes performances lors de la phase d'apprentissage par un réseau de neurones, bien que les modalités soient grandement déséquilibrées.

La figure 4.9 présente les résultats pour les **labels non-dichotomiques et finis**. On y observe les résultats pour les labels ÉMOTION (figure 4.9(c)), LOCUTEUR (figure 4.9(b)) et TYPE (figure 4.9(a)). On remarque que l'accord entre les annotateurs est très faible. Dans la figure 4.9, à des fins de lisibilité, nous n'avons pas représenté les données dont les scores sont différents de 0. Il est important de comprendre ici, pour lire ces graphiques, qu'une donnée annotée vrai (1) dans les quatre modalités de type par un annotateur et faux (0) dans ces mêmes modalités par un autre annotateur est comprise dans les 579 (AUTRES), 404 (TÉMOIGNAGE), 108 (RÉPONSE) et 263 (QUESTION) données ambiguës sur ces modalités du label LOCUTEUR. En effet, elle aura un score entre 0 et 1 pour chacune de ces modalités. La figure 4.9 montre qu'il y a bien plus de données ambiguës que de données consensuelles.

Du côté des **modalités à labels non finis** comme l'identification des INM ou des symptômes, il n'existe que peu de modalités identifiées un grand nombre de fois (plus de 35 fois sur les 1 635 données annotées). Pour les INM, elles sont au nombre de huit. Les six premières concernent plus ou moins directement le cannabis et ses dérivés. Les deux suivantes sont respectivement un médicament (l'ivermectine) qui n'est donc pas

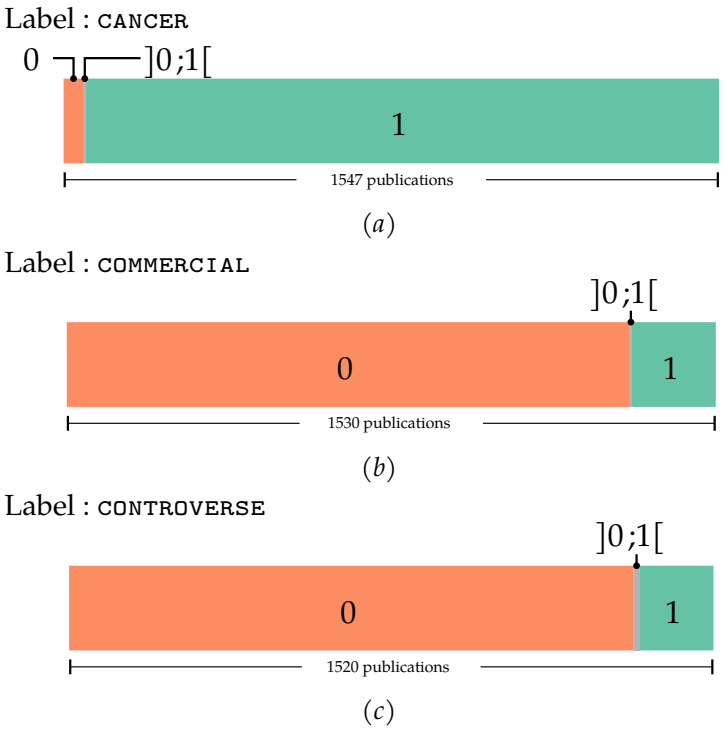


FIGURE 4.8 – *Distribution des scores des publications pour les labels dichotomiques.* (*a*) `cancer`. (*b*) `commercial`. (*c*) `controverse`. *On observe l'accord presque systématique, sur les trois labels, entre annotateurs, car le score des publications annotées est, le plus souvent, égal à 0 (en orange) ou 1 (en vert). Cela signifie que tous les annotateurs ont fait le même choix et que la moyenne des choix est donc égale à 0 ou 1. Un score entre 0 et 1 exclus montre un désaccord entre les annotateurs (en gris).*

une INM, et les régimes cétogènes. Pour les symptômes, ils sont au nombre de deux, c'est l'appétit (concerne le manque d'appétit) et l'anxiété.

Au vu de la très faible proportion de données ayant produit un accord entre les annotateurs sur les labels non dichotomiques, nous produisons un jeu de données supplémentaire décrit dans la section 4.4.5.2.

4.4.5.2 Données annotées par des règles

Nous avons considéré que la structure des subreddits rendait évidente la sélection de labels. Nous avons donc défini un ensemble de règles métier que nous allons décrire dans cette section (tableau 4.2).

Considérons les **labels dichotomiques**. Pour le label CANCER, nous extrayons des données du subreddit r/cancer et nous les annotons automatiquement avec le label CANCER. Pour les données ne concernant pas le cancer, nous extrayons des données indépendamment de leur contenu et de leur subreddit, puis vérifions qu'elles ne contiennent pas de termes relatifs au cancer. La liste des termes considérés est la suivante : cancer, tumour, tumor et oncology. Nous extrayons également des données de r/advertising et r/Controversialopinions de la même manière et conservons celles qui ne contiennent pas de termes relatifs au cancer.

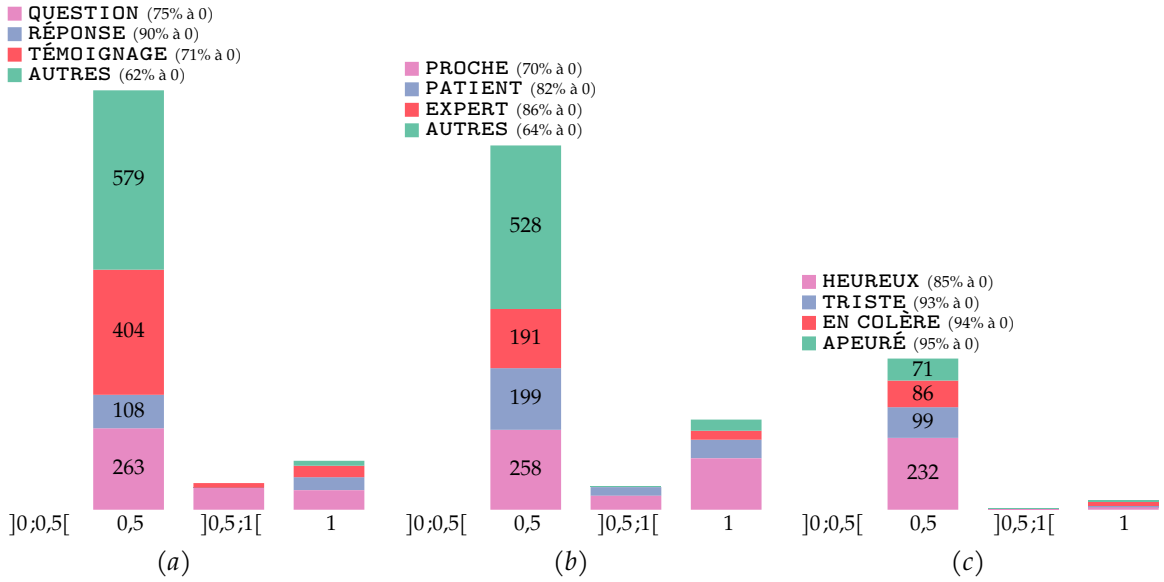


FIGURE 4.9 – *Distribution des labels à plusieurs modalités.* (*a*) `type`. *On remarque que le type de publication est le label le plus renseigné. On observe également un grand nombre de témoignages selon les annotateurs, même si l'accord entre annotateurs est rare.* (*b*) `locuteur`. *On note que le locuteur est renseigné par les annotateurs, avec un consensus plus grand pour ce label aussi bien en proportion qu'en nombre, comparé aux autres labels à plusieurs modalités.* (*c*) `émotions`. *On observe que les émotions sont peu renseignées par les annotateurs. Finalement, sur les trois labels, les annotateurs n'ont pas été en accord lorsque l'un d'entre eux a renseigné un label. Ne figure pas, pour des problématiques de lisibilités la modalité 0. Il est important de lire ici que pour toutes les modalités de tous les labels, le score final est, le plus souvent, égal à zéro et donc qu'un consensus avait lieu sur la réponse "non" à la modalité du label en question. En d'autres termes, le consensus est très fréquent lorsqu'au moins un annotateur choisit le label à zéro, très peu fréquent lorsqu'au moins un annotateur choisit le label à 1.*

Pour le label `COMMERCIAL`, nous extrayons des données du subreddit r/advertising et les annotons comme ayant des objectifs commerciaux. Pour les données étiquetées comme non commerciales, nous sélectionnons les publications du subreddit r/AskDocs et les premiers commentaires des publications des subreddit r/offmychest, r/happy et r/sad.

Pour le label `CONTROVERSE`, nous extrayons les publications et le premier commentaire du subreddit r/Controversialopinions et les annotons comme controversées. Pour les non-controversées, nous choisissons des publications dans tout Reddit, en vérifiant qu'elles ne sont pas commentées ainsi que les publications des subreddits r/offmychest, r/family et r/happy.

Pour les **labels non dichotomiques**, nous choisissons de rendre binaire chacune des modalités et d'entraîner des classifieurs sur ces modalités. La motivation derrière ce choix est de pouvoir ensuite observer des corrélations entre des modalités. Pour la modalité `PATIENT` du label `LOCUTEUR`, nous annotons comme produite par un patient les publications concernant un membre du corps et ne faisant mention d'aucun proche et les publications ayant, dans le titre, une forme anglophone de **"j'ai"** du subreddit r/AskDocs. En effet, lors de l'exploration préliminaire, nous avons remarqué que la formule "j'ai [maladie ou symptômes]" (*e.g.* la formule "j'ai des nausées") est régulière,

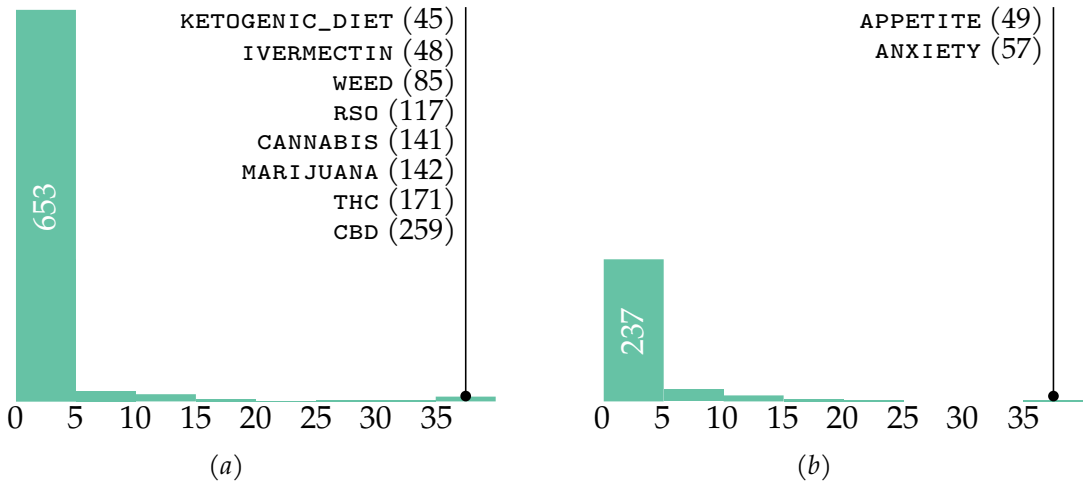


FIGURE 4.10 – *Distribution des labels non-finis.* (*a*) INM. *On note que pour de nombreuses INM identifiées, elles ne le sont qu'un nombre très faible de fois* (*moins de cinq fois*). *Les INM les plus identifiées sont majoritairement en rapport avec le cannabis et ses dérivés. On peut néanmoins citer le régime cétogène qui fait partie des INM les plus citées. L'ivermectine, en revanche, est un médicament et n'est donc pas une INM. On peut néanmoins relever l'attention apportée par les utilisateurs à ce médicament souvent cité.* (*b*) symptôme. *On remarque que les symptômes sont bien moins renseignés que les INM, que ce soit en nombre d'occurrences pour un symptôme ou en nombre de symptômes. Les symptômes les plus identifiés sont ceux qui font référence à l'anxiété et à l'appétit.*

tout comme celle consistant à citer la partie du corps en douleur ou malade (*e.g.* "j'ai le bras qui enfle"). Pour s'assurer que le locuteur parle de sa propre maladie, nous excluons toutes les publications faisant mention de membres de la famille ou d'amis. Pour la modalité PROCHE du label LOCUTEUR, nous sélectionnons les publications issues de r/family. Les données provenant de r/AskDocs ayant une forme anglophone de "j'ai" ou les premiers commentaires des publications annotés comme issus d'un professionnel de santé ont servi comme données ne provenant pas de proches. Certains subreddits produisent un label spécifique pour identifier les publications des professionnels de santé. C'est le cas du subreddit r/AskDocs que nous avons utilisé pour la modalité EXPERT. Pour le label TYPE, nous utilisons trois subreddits. r/AskDocs nous a permis de construire les modalités QUESTION et RÉPONSE dans les cas où la publication était, respectivement, une publication d'origine ou un commentaire. Nous utilisons le subreddit r/NoStupidQuestions de la même manière. Le subreddit r/offmychest regroupe des publications d'utilisateurs ayant besoin de faire part de quelque-chose. Ce subreddit est utilisé pour la modalité TÉMOIGNAGE. À l'inverse, les publications issues de r/NoStupidQuestions ont servi de contre-exemples. Pour le label ÉMOTION, les publications issues du subreddit r/happy ont été annotées positivement pour la modalité HEUREUX et négativement pour les autres. Les publications issues des subreddits r/sad, r/angry et r/worried sont annotées positivement pour les labels TRISTE, EN COLÈRE et APEURÉ, respectivement. Ces mêmes publications ont été annotées négativement pour la modalité HEUREUX.

Pour les **labels non finis**, nous sélectionnons simplement dans les données déjà exportées celles contenant le terme identifié.

label	modalité	all	r/cancer	r/advertising	r/ControversialOpinions	r/AskDocs	r/NoStupidQuestions	r/offmychest	r/family	r/happy	r/sad	r/angry	r/worried	total
CANCER	VRAI		■											29 907
	FAUX	■		■	■									35 640
COMMERCIAL	VRAI			■										7 082
	FAUX					■		■		■	■			68 413
CONTROVERSE	VRAI				■■									12 941
	FAUX	■						■	■	■				96 486
LOCUTEUR	PATIENT					■■								17 073
	PATIENT					■			■					29 989
	PROCHE								■					29 989
	PROCHE					■■								7 257
	EXPERT					■								5 807
	EXPERT					■			■					31 439
TYPE	QUESTION					■	■							50 427
	QUESTION					■	■							25 431
	RÉPONSE					■	■							25 431
	RÉPONSE					■	■							50 427
	TÉMOIGNAGE							■						29 879
	TÉMOIGNAGE						■							20 539
ÉMOTION	HEUREUX									■				26 969
	HEUREUX										■	■	■	36 246
	TRISTE										■			29 143
	TRISTE									■				26 969
	EN COLÈRE											■		7 022
	EN COLÈRE									■				26 969
	APPEURÉ												■	81
	APPEURÉ									■				26 969

■ La publication
■ Le premier commentaire
■ La publication si elle n'a aucun commentaire
■ La publication si elle contient une forme du type "j'ai"
■ Le premier commentaire si l'auteur est un professionnel de santé selon Reddit
■ La publication si elle concerne un membre du corps et aucun proche
■ La publication si elle ne contient pas de termes relatifs au cancer

vrai faux

Table 4.2 – *Données exportées de Reddit.*

Ces deux jeux de données, annotés manuellement et basés sur des règles ont été utilisés pour entraîner des modèles d'apprentissage automatique décrits dans la section suivante.

4.5 Entraînement des réseaux de neurones

Dans cette section, nous présentons les réseaux de neurones utilisés et sur quelles données ils sont entraînés (section 4.5.1). Ensuite, nous présentons la procédure d'entraînement (section 4.5.2). Enfin, nous présentons les résultats de la classification des données et les labels qui ont été retenus pour l'exploration (section 4.5.3).

4.5.1 Différents réseaux de neurones pour plusieurs labels

Pour classer les données issues des médias sociaux, nous utilisons différents modèles auto-attentifs ayant tous la même architecture : AlBERT [Lan+19]. Les textes Reddit peuvent être très longs et AlBERT est limité sur le nombre de tokens en entrée. Dans le cas de textes trop longs, ceux-ci sont tronqués avant la classification.

Trois couples "données d'entrée-classifieur" ont été sélectionnés pour classifier les textes (figure 4.3). Un premier couple prend en entrée les informations sur le LOCUTEUR, le TYPE de publication, le rapport avec le CANCER. Ce classifieur construit un seul espace de représentation des textes pour de multiples tâches de classifications. Le deuxième couple, prend en entrée tous les mots relatifs à une famille d'INM, dans notre cas d'étude, le cannabis et ses dérivés (RELATIF AU CANNABIS, en rose dans le tableau 4.3). Enfin, le dernier couple associe un seul concept à un seul classifieur et donc un seul mot présent dans les textes pour un seul classifieur (en vert dans le tableau 4.3). Cette structure a notamment été utilisée pour les modalités du label SYMPTÔME.

4.5.2 Procédure d'entraînement

Pour l'entraînement, la grande différence de volume entre les données annotées manuellement, moins volumineuses, et celles annotées par des règles nous a conduit à mettre en place une stratégie d'entraînement spécifique pour ne pas négliger les premières, moins volumineuses. Les réseaux sont donc entraînés à l'aide d'un batch de données annotées par l'humain puis d'un batch de données annotées à l'aide de règles. Pour le réseau de neurones, chaque batch ne concerne qu'un seul label. Les données annotées par l'humain ont donc été vues bien plus de fois que les données annotées à l'aide de règles. Le volume de données utilisées pour l'entraînement est présenté dans le tableau 4.3.

4.5.3 Résultats et labels sélectionnés

Nous avons conservé uniquement les classifieurs avec un score F1 moyen (sur les modalités VRAI et FAUX) supérieur à 0.75. Malgré des scores plus faibles, nous avons conservé les labels relatif au locuteur, au type et aux dérivés du cannabis car ils sont importants pour le cas applicatif.

		données annotées				scores classe vrai		
		manuellement		par règles				
catégorie	label	VRAI	FAUX	VRAI	FAUX	P	R	F1
	RELATIF AU CANNABIS	531	645	3 761	3 761	0.36	0.52	0.42
	CANCER	1 205	40	29 279	34 929	0.97	0.99	0.98
	COMMERCIAL	158	1 076	6 957	67 132	0.0	0.0	0.0
	CONTROVERSE	154	1073	12 682	94 460	0.0	0.0	0.0
LOCUT.	PATIENT	229	1 002	16 746	29 346	0.64	0.86	0.74
LOCUT.	PROCHE	371	860	29 346	7 111	0.99	0.27	0.43
LOCUT.	EXPERT	176	1 055	5 683	30 774	1.0	0.02	0.03
TYPE	QUESTION	365	876	49 434	24 932	0.64	0.92	0.75
TYPE	RÉPONSE	153	1 088	24 932	49 434	0.23	0.33	0.27
TYPE	TÉMOIGNAGE	352	889	29 269	20 118	0.2	0.01	0.02
SYMPTÔME (anglais)	PAIN	124	124	22 345	22 387	0.87	0.96	0.91
SYMPTÔME (anglais)	SICK	3	3	7 233	7 256	0.83	0.81	0.82
SYMPTÔME (anglais)	WEAK	4	4	3 398	3 418	0.68	0.87	0.76
SYMPTÔME (anglais)	DEPRESSION	16	17	6 014	6 030	0.80	0.89	0.84
SYMPTÔME (anglais)	LYMPHOMA	17	17	1 033	1 037	0.47	1.0	0.64
SYMPTÔME (anglais)	APPETITE	36	37	1 142	1 145	0.96	1.0	0.98
SYMPTÔME (anglais)	ANXIETY	44	44	9 469	9 481	0.91	0.91	0.91
SYMPTÔME (anglais)	VOMITING	13	13	849	853	0.06	1.0	0.11
SYMPTÔME (anglais)	PTSD	1	2	969	977	0.0	0.0	0.0

TABLE 4.3 – *Jeux de données d'entraînement avec une sélection de labels. On remarque la grande disparité dans le volume des données entre les données annotées à l'aide de règles ou à l'aide des annotateurs. Parmi cette sélection de labels, certains n'ont pas obtenu de résultats satisfaisants lors de l'entraînement. C'est par exemple le cas pour les labels* `commercial`, `controverse`, `votimiting` *et* `ptsd`. *Le label* `relatif au cannabis` *(ici en rose) a été entraîné sur la présence de plusieurs mots relatifs au cannabis. Les labels en orange sont issus d'un unique réseau de neurones à plusieurs tâches de classification. Enfin, chacun des labels en vert a été utilisé pour entraîner un réseau de neurones spécifique à un terme uniquement. Les labels sélectionnés pour être utilisés dans notre cas d'étude sont marqués d'un carré jaune. Les scores de classification dans les trois dernières colonnes sont ceux de la classe* `vrai` *sur le jeu de données exporté. Nous utilisons la classe* `vrai` *uniquement car, en pratique, c'est la présence d'un concept qui intéresse l'utilisateur, pas son absence. Le jeu de données test est déséquilibré pour de nombreux labels. C'est notamment le cas pour tous les symptômes où dans les données de test il y a beaucoup plus de données labélisées à* `faux` *qu'à* `vrai`. *Certaines des valeurs présentes ici diffèrent de celles dans le tableau 4.2 du fait que les publications et les commentaires, ayant été supprimés de la plateforme Reddit, ont été écartés pour l'entraînement.*

Les labels CONTROVERSE, COMMERCIAL et ÉMOTION n'obtiennent pas de bons résultats. Il s'avère que les données exportées et annotées ne sont pas suffisamment nombreuses pour ces labels. Ils sont mis de côté pour la suite de cette étude.

En ce qui concerne les symptômes, nous obtenons des résultats satisfaisants pour huit d'entres eux : PAIN, SICK, WEAK, TUMOUR, CHEMO, DEPRESSION, APPETITE et ANXIETY. Généralement, il s'agit des labels pour lesquels le volume de données disponibles dans les deux jeux de données est le plus grand. Tous ne concernent pas spécifiquement un

symptôme comme SICK et CHEMO mais ont été identifiés comme pertinents par l'expert.

Les labels finalement conservés sont donc le rapport avec le cancer (CANCER), le locuteur (PATIENT, EXPERT, RELATIVE), le type de publication (QUESTION, ANSWER, TESTIMONY), le cannabis et ses dérivés (RELATIF AU CANNABIS) et les symptômes (SYMPTÔME). Améliorer les résultats pour les autres labels aurait nécessité de récolter de nouvelles données.

4.6 Cancer-Vis : Outil d'exploration des données.

Dans cette section, nous présentons l'outil d'exploration des données issues des médias sociaux, le cœur de nos travaux sur ce chapitre. Nous caractérisons le problème relatif à cette exploration dans la section 4.6.1. Ensuite, nous présentons les questions auxquelles nous souhaitons répondre et les besoins identifiés dans la section 4.6.2 et dans la section 4.6.3, respectivement. Enfin, nous présentons la solution proposée : Cancer-Vis dans la section 4.6.4.

4.6.1 Caractérisation du problème

L'outil de visualisation des données est destiné à des personnes du corps médical ou des patients cherchant des informations à propos des INM et leur utilisation dans le cas d'un cancer. L'outil doit permettre une exploration facilitée des médias sociaux une fois qu'une publication d'intérêt a été identifiée. Nous listons différentes questions auxquelles les utilisateurs souhaitent répondre et les besoins de l'outil associés à ces questions dans les deux sections suivantes.

4.6.2 Questions des utilisateurs pour l'exploration

Nous énonçons six questions dans un cadre général non spécifique à notre cas d'étude.

Qu.1 Quelles sont les publications concernant un des concepts identifiés ?

Qu.2 Quelles sont les publications concernant un des labels entraînés ?

Qu.3 Quels sont les liens existants entre les différents labels ?

Qu.4 Quels sont les textes respectant les filtres voulus ?

Qu.5 Quel sont les commentaires d'une publication et d'où provient-elle ?

Qu.6 Quels labels peut-on associer à une publication et comment le ou les réseau(x) traitent le texte pour ces labels ?

4.6.3 Besoins identifiés pour l'outil

Pour répondre à ces questions, nous identifions cinq besoins auxquels l'outil d'annotation doit répondre :

Be.1 Filtrer les données sur tous les labels en fonction des prédictions du réseau ou des réseaux de neurones et de la présence de mots pour Qu.1 et Qu.2 ;

Be.2 Visualiser les corrélations entre plusieurs variables, mais aussi la distribution des données sur un label ou sur des couples de labels pour Qu.2 et Qu.3 ;

Be.3 Afficher les textes correspondant aux filtres, les ordonner sur plusieurs labels et fournir les labels les concernant Qu.1, Qu.2 et Qu.4 ;

Be.4 Accéder au contenu original sur le média social de provenance d'un texte pour Qu.5 ;

Be.5 Accéder à un score de contribution issu ou non du fonctionnement du réseau de neurones pour Qu.6.

4.6.4 Exploration des données issues des médias sociaux

Nous présentons, dans cette section comment l'outil Cancer-Vis répond aux besoins énoncés dans la section 4.6.3.

4.6.4.1 Attributs des données

Les données sont associées à des prédictions pour chaque label. Elles possèdent aussi différents scores de contribution (section 2.3.2.4). Dans notre cas d'étude, ces valeurs sont issues des différents réseaux de neurones. Les titres et les textes des publications sont bien évidemment essentiels, mais il est également nécessaire d'avoir la liste des tokens traités par les réseaux de manière à pouvoir produire la visualisation de la contribution, c'est-à-dire le score d'attention issu du modèle auto-attentif pour chaque token. Toutes ces informations sont utilisées dans l'outil de visualisation Cancer-Vis présenté dans la figure 4.11.

4.6.4.2 Filtres et exploration

La visualisation explorant les données issues des médias sociaux est composée d'une partie supérieure où l'on trouve les labels et la zone de filtre et d'une partie inférieure informant sur les données (figure 4.11).

Filtres et badges des labels : la partie supérieure de la visualisation, encadrée en bleu dans la figure 4.11, permet de filtrer les données pour répondre au besoin Be.1. Pour utiliser les labels dans le reste de la visualisation (partie inférieure), il suffit de glisser le badge d'un label sur la zone de dépôt associée. Les badges se trouvent soit dans des listes déroulantes ouvrables dans la partie supérieure, soit directement dans la partie supérieure. Dans notre cas d'étude, seuls les symptômes se trouvent dans une liste déroulante. La zone dépôt pour filtrer peut contenir zéro, un ou plusieurs badges. Il est, par exemple, possible de filtrer les données qui ont une prédiction supérieure ou égale à une valeur saisie sur le label concerné. Pour un filtrage plus fin sur les données, il est possible d'ajouter des règles de disjonction logique (AND ou OR).

Les prédictions issues des réseaux de neurones sont utiles lorsque l'on a pu établir, à l'avance, les labels d'intérêt. Néanmoins, au cours d'une exploration, un utilisateur

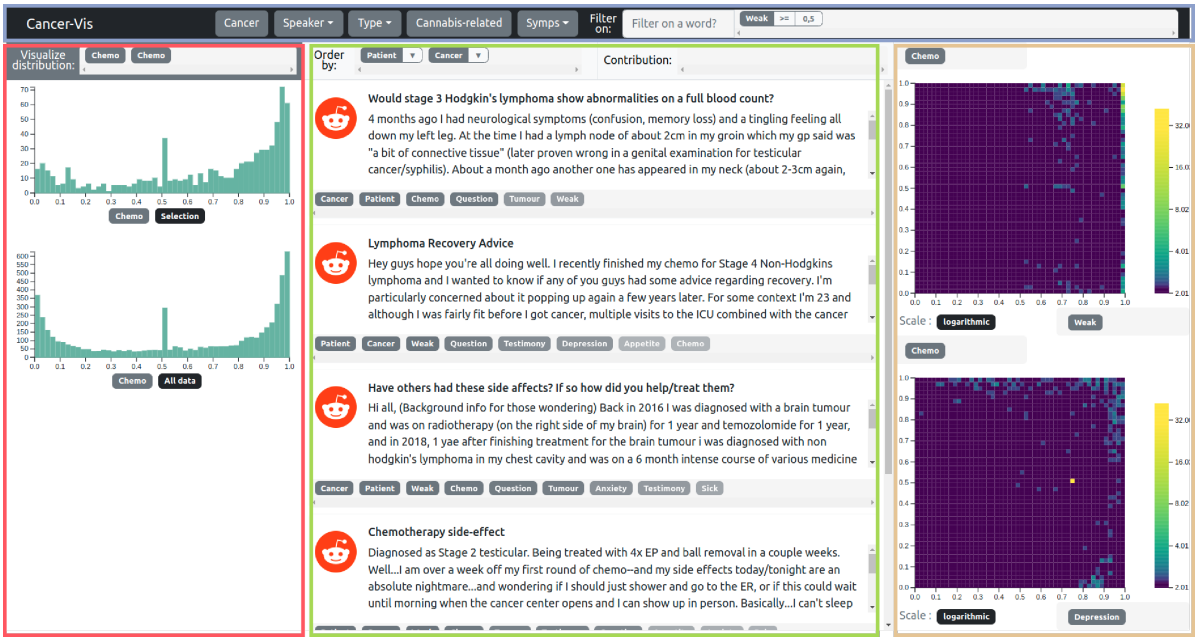


FIGURE 4.11 – *Vue globale de l'outil Cancer-Vis. La partie contenant les labels et permettant de filtrer les données est dans l'encadré bleu. L'encadré rouge contient les distributions des labels. La partie centrale, dans l'encadré vert, présente les textes. Enfin, dans l'encadré doré, deux cartes de chaleur présentent les liens entre plusieurs labels.*

peut être intéressé par de nouveaux concepts sur lesquels aucun réseau de neurones n'a été entraîné (Be.1). Ainsi, nous proposons également un filtre sur les occurrences des mots. Ce filtre se combine au précédent avec une disjonction logique (AND ou OR).

Exploration des données : la partie inférieure de la visualisation permet l'exploration. Elle est composée de trois panneaux.

Le panneau central, encadré en vert dans la figure 4.11, assure une visualisation des textes. Par défaut, cinq textes correspondant au filtrage et suivant l'ordonnancement sont affichés. Deux zones de dépôts de badges servent, respectivement, à ordonner les textes sur un label ou un groupe de labels (Be.3) et à afficher le score de contribution pour un label choisi (Be.5). Dans le cas où l'utilisateur voudrait ordonner sur plusieurs labels, le score d'ordonnancement est calculé en multipliant les scores d'ordonnancement sur les labels demandés. La liste des textes peut être étendue de cinq textes supplémentaires à l'aide d'un clic sur le bouton associé en bas de la liste des textes.

Par défaut, les textes sont présentés dans des encadrés. Ils contiennent le lien vers la page associée à la publication, le titre et le texte de la publication si il y en a. Sur la partie inférieure des encadrés sont visibles différents badges dont l'opacité dépend de la prédiction. Ils sont ordonnés du plus probable au moins probable en n'apparaissent que si la prédiction excède la valeur 0.5. Cette fonctionnalité répond au besoin Be.3.

Il est possible de prendre en considération une information supplémentaire, le score de contribution issu du fonctionnement du réseau, pour repérer la contribution des mots dans les textes. Nous utilisons, dans notre cas d'étude, le score d'attention

pour chaque couple de tokens. Une fois le badge déposé dans la zone de dépôt, la zone de texte est modifiée pour proposer une vue token par token avec le score de contribution au survol des tokens issu de la dernière couche d'attention du modèle auto-attentif. Elle passe alors de la vue par défaut (figure 4.12(a)), où le titre et le texte de la publication sont clairement identifiables comme différents, à la vue contribution (figure 4.12(b)), où seuls les tokens apparaissent de la manière dont ils ont été utilisés par le réseau.

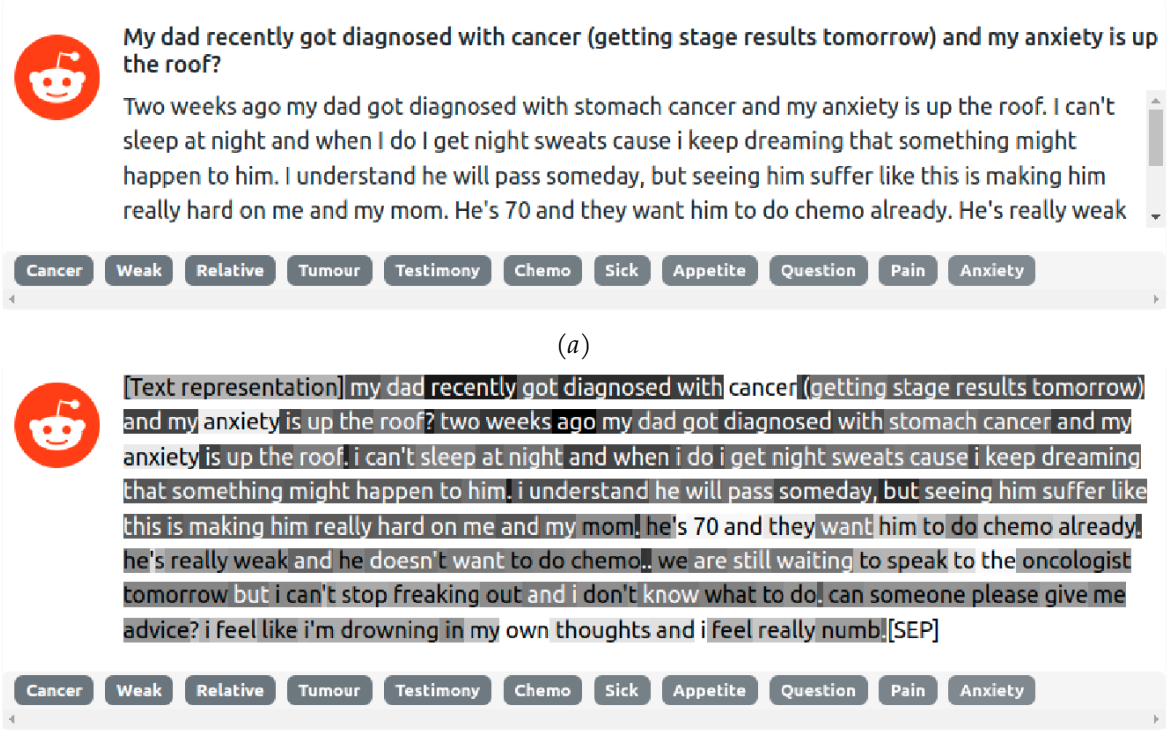


Figure 4.12 – *Affichage des textes dans l'outil Cancer-Vis. (a) Texte sans score d'attention. (b) Texte avec le score d'attention au survol du token de représentation du texte pour la classifieur* `anxiety`.

Le panneau de gauche, encadré en rouge dans la figure 4.11, produit une distribution des prédictions pour un label lorsque l'on glisse dans la zone de dépôt un badge. Cette distribution est proposée sous forme d'histogramme qui donne trois informations : la distribution sur l'intégralité des données, la distribution sur les données filtrées et la comparaison entre les deux distributions précédemment mentionnées. Ces trois informations aident à savoir si un filtre impacte la distribution des données. Cela permet aussi de savoir si le réseau a été sensible à la présence de mots ou de tokens, ou si des labels sont corrélés (une étude approfondie de la corrélation entre les labels est disponible sur le panneau de droite). Plusieurs distributions de labels différents peuvent être visualisées simultanément. Elles sont alors présentées les unes sous les autres. Ces distributions répondent au besoin Be.2.

Dans le panneau de droite, encadré en doré dans la figure 4.11, nous affichons deux cartes de chaleur à deux dimensions produisant une comparaison des distributions des prédictions entre deux labels. Ces cartes apparaissent lorsque les deux zones de dépôt sont renseignées. Elles ont la même échelle et sont donc comparables. L'outil intègre

deux échelles, une échelle linéaire et une échelle logarithmique. Cette dernière est essentielle lorsque l'on s'intéresse à des classifieurs certains de leur prédictions. Enfin, les cartes de chaleur ne traitent que les données correspondant au filtrage. Cela affine la vue des données en excluant les zones denses des distributions à l'aide du filtrage si cela est nécessaire. Ces cartes de chaleur et leurs fonctionnalités répondent au besoin Be.2. Elles montrent l'incertitude des réseaux et dans quelle mesure cette incertitude peut-être commune à deux classifieurs. Cela montre aussi si des corrélations entre ces derniers existent. Il serait intéressant de vérifier qu'elles sont plus faciles à interpréter par un utilisateur non spécialiste que les matrices de confusion.

4.7 Études de cas

Dans cette section, nous présentons deux cas d'études répondant aux questions des utilisateurs. Le premier présente l'univers des mentions d'INM relatives au cannabis sur le média social Reddit. Son objectif est de valider la démarche présentée dans ce chapitre. Le deuxième cas d'étude montre comment les classifieurs traitent les publications et le lien qui existe entre dépression et anxiété.

4.7.1 Dérivés du cannabis

Dans la section 4.4.5.1, nous avons montré que les INM les plus mentionnées concernent en majorité les dérivés du cannabis. Ainsi, dans ce premier cas d'étude, nous nous plaçons dans une situation où un patient ou un proche aimerait avoir des informations sur les possibles effets des dérivés du cannabis. Nous choisissons d'ordonner les publications RELATIVES AU CANNABIS et au CANCER (voir encadré jaune dans la figure 4.13). Ensuite, nous souhaitons isoler des symptômes et filtrons les publications relatives à PAIN, WEAK, DEPRESSION, APPETITE et ANXIETY (voir encadré bleu dans la figure 4.13). L'outil de visualisation inclut les disjonctions logiques. Ainsi n'importe quelle publication ayant une prédiction d'au moins 0.9 sur un de ces labels est affichée.

Sur les cinq publications affichées, quatre concernent spécifiquement le cancer et au moins un de ces symptômes. Les deux premières publications concernent le cancer du pancréas. L'auteur de la première demande si l'utilisation de "marijuana" peut avoir des effets négatifs dans le cas d'un poids faible et d'une pression artérielle faible. L'auteur aimerait utiliser de la "marijuana" de manière à améliorer l'appétit de son père, atteint du cancer. Ce qui ressort des réponses à sa question, consultées sur Reddit, est qu'il faut en parler au professionnel de santé en premier lieu. Des patients ou anciens patients, affirment que la "marijuana" aide pour les problèmes liés à l'appétit. Une personne précise même que dans son cas, en utilisant la "majijuana" avec une pression artérielle faible, il allait bien (en anglais : "I was fine."). D'autres mentionnent qu'il est préférable d'ingurgiter les dérivés de cannabis plutôt que de les fumer. Les terme utilisés sont "edibles", "oral thc" ou encore "CBD oil".

Pour préciser la recherche sur les possibles effets des dérivés du cannabis sur l'appétit, nous nous focalisons sur les publications où la prédiction pour le label APPETITE est supérieure à 0.9 et où un patient s'exprime et non un proche comme dans

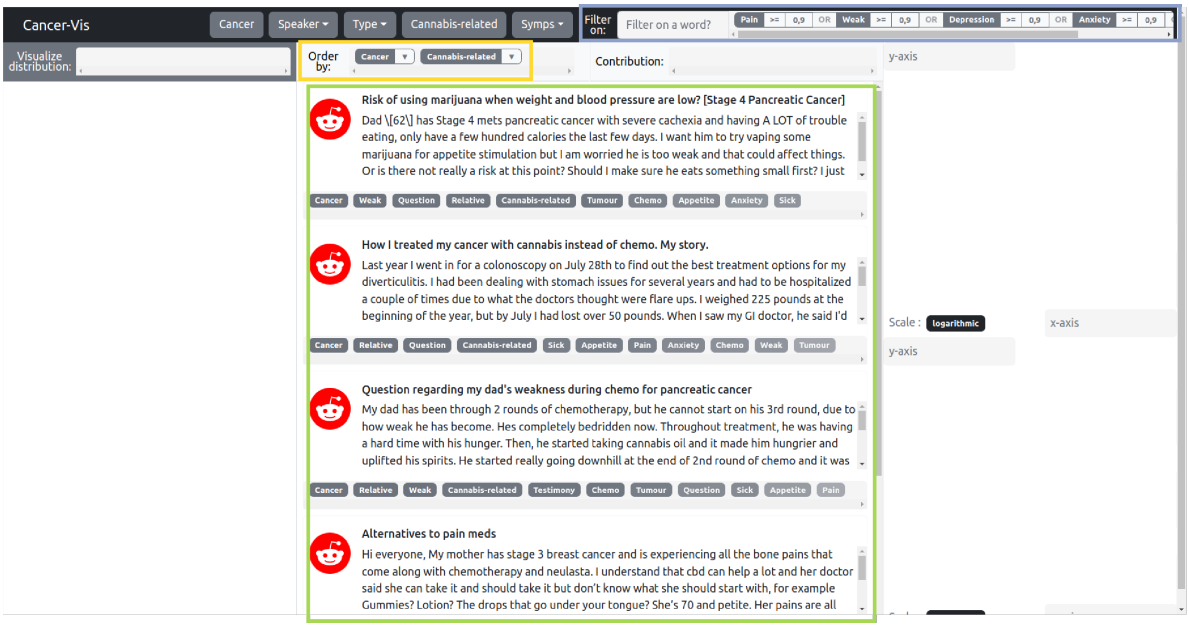


FIGURE 4.13 – *Exploration des textes mentionnant un* `symptôme` *en lien avec le* `cancer` *et l'utilisation d'une INM* `relative du cannabis`. *On observe, ici, dans l'encadré jaune l'ordonnancement choisi utilisant les labels* `relatif du cannabis` *et* `cannabis`. *Dans l'encadré bleu, les filtres choisis impactent les textes affichés dans l'encadré vert.*

les publications précédentes. Le score de prédiction pour le label PATIENT doit donc lui aussi être supérieur à 0.9. Beaucoup de publications ne sont pas réellement issues de patients compte tenu des performances du classifieur (voir tableau 4.3). Cependant, nous avons pu isoler deux publications pertinentes. La première est produite par un patient en rémission. Il s'est servi des dérivés du cannabis pour contrer les effets indésirables du cancer. Il souhaite recueillir des témoignages de patients ayant eux aussi utilisé des dérivés du cannabis. Parmi ces témoignages, certains relatent des effets positifs sur la douleur ou l'anorexie mais aussi un manque d'effet (figure 4.14). Dans les produits essayés, on trouve : "suckers / gummi bears" et "Simpson oil". Dans la seconde publication, une femme présente sa situation, dans laquelle elle a essayé de nombreux traitements alternatifs pour des problèmes de nausées malheuresement sans succès. Elle cite notamment la "CBD oil" (figure 4.15).

Beaucoup de publications ont été supprimées de Reddit par des modérateurs ou par les utilisateurs eux-mêmes mais les commentaires sont toujours disponibles. Sur les règles de publications, dans le subreddit, figure cette règle : "No homeopathy / nature / quack medicine". Cela montre la controverse que peut soulever l'utilisation de dérivés du cannabis dans le cadre du cancer. Quand des commentaires de patients ou de proches avancent que des effets peuvent exister pour ces produits, les modérateurs évitent que ces messages soient visibles de manière à ce que ces produits ne se substituent pas aux traitements et soient toujours utilisés après consultation d'un spécialiste.

Pour conclure ce cas d'étude, les avis présents sur Reddit appartiennent au domaine du déclaratif et doivent être considérés ainsi. Néanmoins, après une exploration

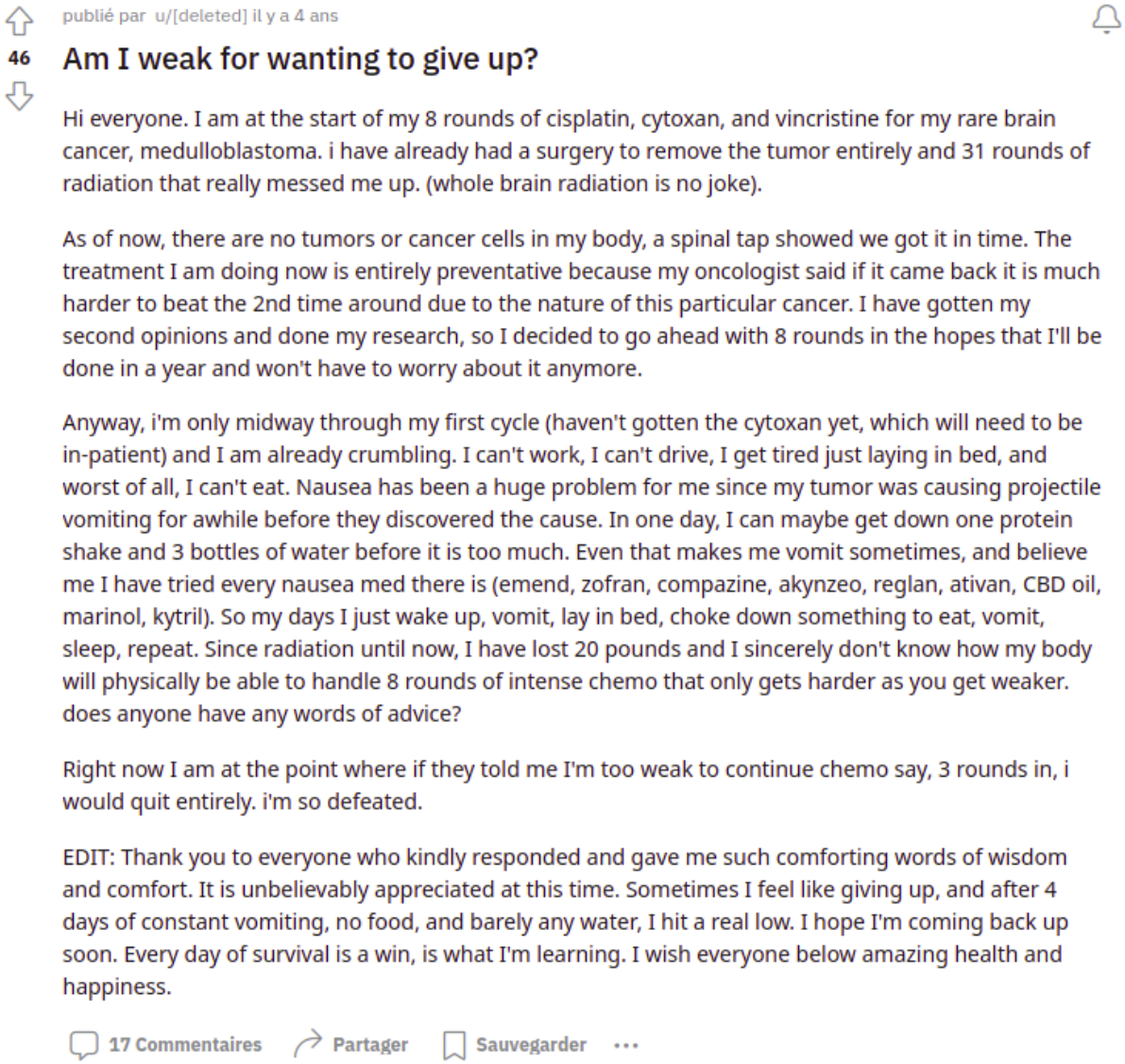


FIGURE 4.14 – *Publication sur Reddit.*

préliminaire via notre outil, on peut observer que pour les patients atteints d'un cancer, l'utilisation des dérivés du cannabis peut être une solution pour améliorer la qualité de vie vis-à-vis de certains des symptômes. Il existe donc de nombreux effets positifs mentionnées par les utilisateurs, bien qu'une méfiance existe et qu'il n'existe pas de consensus sur ces effets.

4.7.2 Anxiété et dépression

Dans ce second cas d'étude, nous souhaitons nous concentrer sur les symptômes de dépression et d'anxiété (figure 4.16) qui sont souvent mentionnés dans le cadre du cancer [Rin+21 ; Mat+21 ; Sal+22 ; Bre+22].

De manière à étudier les liens possibles entre ces deux symptômes, nous souhaitons tout d'abord voir si ces classifieurs s'appuient sur des termes similaires. Pour le classifieur associé au label DEPRESSION par exemple, nous constatons que l'attention se

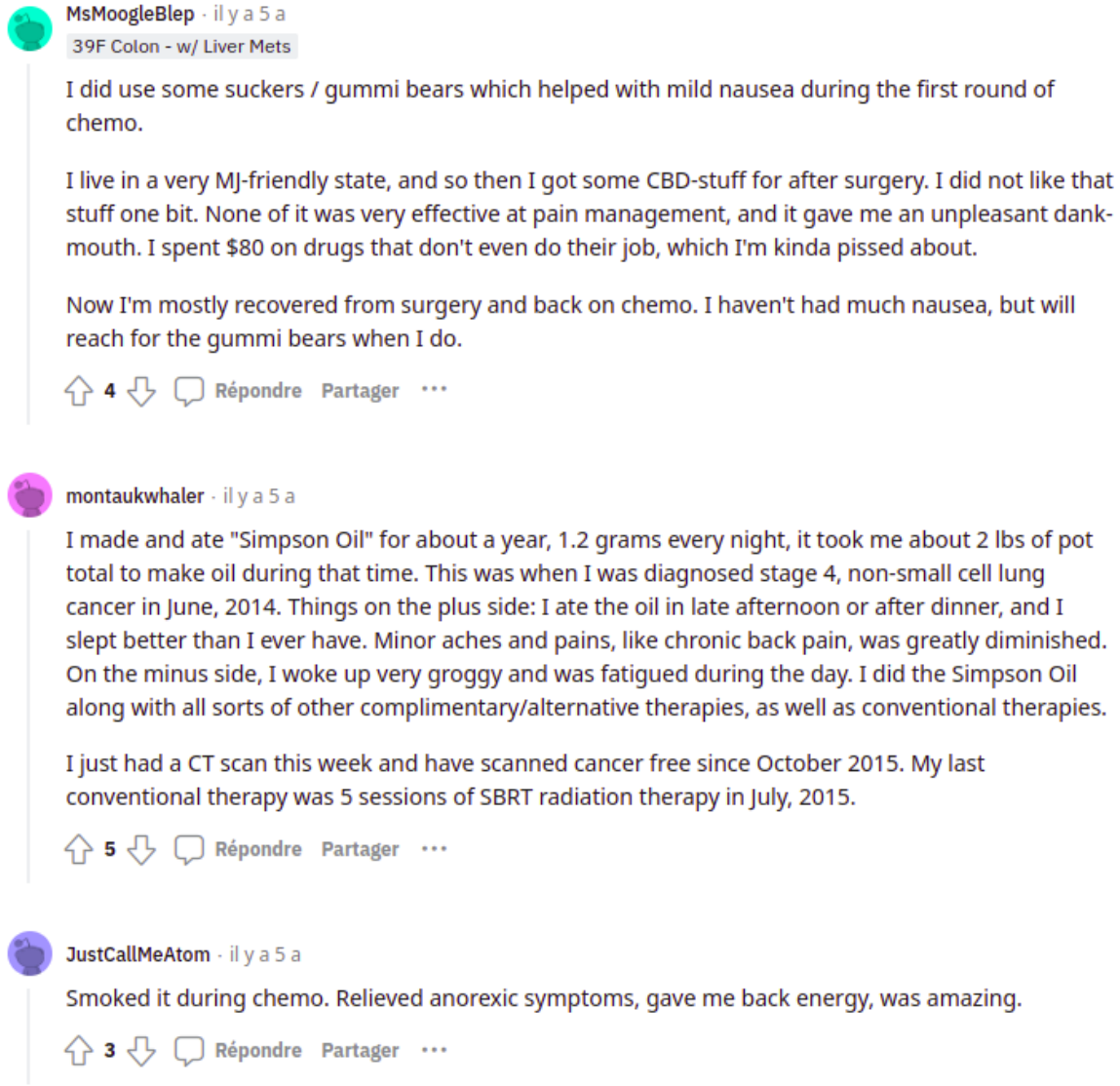


Figure 4.15 – *Commentaires sur Reddit.*

concentre sur le terme ANXIETY dans un des textes (figure 4.17) puisque ce mot est celui avec le fond le plus foncé lorsque l'on affiche le score d'attention. Cette observation n'est pas vraie pour tous les autres textes. Cela signifie que le classifieur s'appuie sur des concepts plus complexes que la simple présence d'un mot. Néanmoins, nous observons, grâce à la carte de chaleur où les axes présentent les scores de classification des textes, qu'il existe une corrélation entre les deux classifieurs. Cette interprétation ne signifie pas que les textes mentionnant la dépression mentionnent l'anxiété et inversement malgré les performances acceptables de ces classifieurs (tableau 4.3). Néanmoins, cela peut signifier qu'il y a des similitudes entre ces deux concepts selon leur apparition dans les textes ou leur plongement lexical.

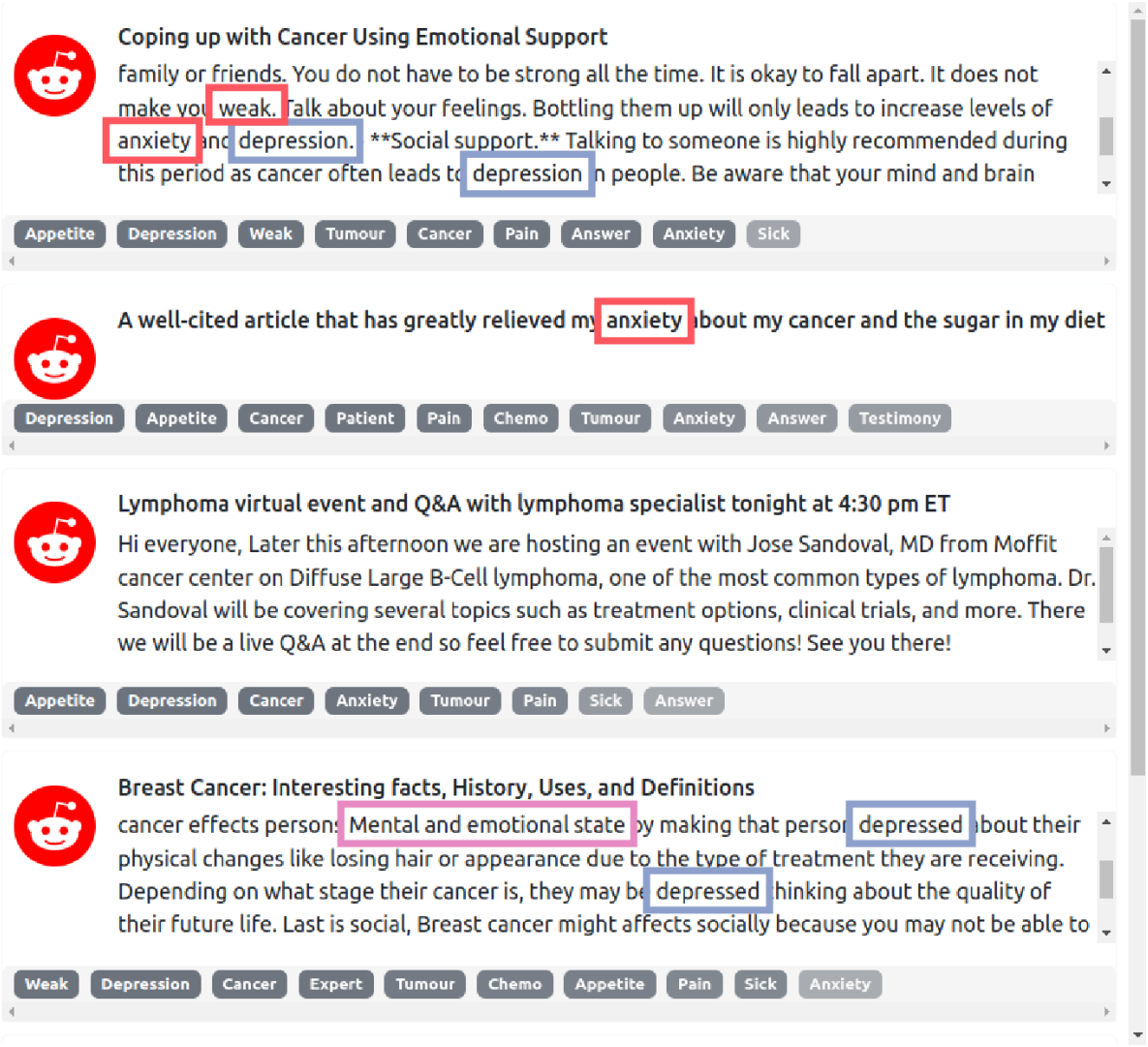


FIGURE 4.16 – *Textes concernant le label* `depression`. *On observe, ici, dans les encadrés bleus, les mentions explicites de la dépression, dans les encadrés roses les termes implicites de la dépression et dans les encadrés rouges les mentions explicites d'autres symptômes. On constate que le label* `anxiety` *pour lequel nous avons entraîné un classifieur apparaît deux fois dans ces textes.*

4.8 Discussions

Dans cette section, nous discutons en premier lieu des résultats concernant notre sujet d'étude à savoir : les mentions d'INM dans les médias sociaux (section 4.8.1). Ensuite, nous discutons des objectifs et les limites de l'outil d'exploration des données (section 4.8.2). Enfin, nous discutons des limites de la méthode dans le cas général et dans le cas spécifique de notre cas d'étude (section 4.8.3).

4.8.1 Interventions non-médicamenteuses et cancer

Au cours des explorations des publications mentionnant les INM dans le cadre du cancer, nous constatons qu'elles ne font ressortir que peu de résultats à la suite de l'annotation manuelle (section 4.4.5). Les mentions d'INM concernent très majoritaire-

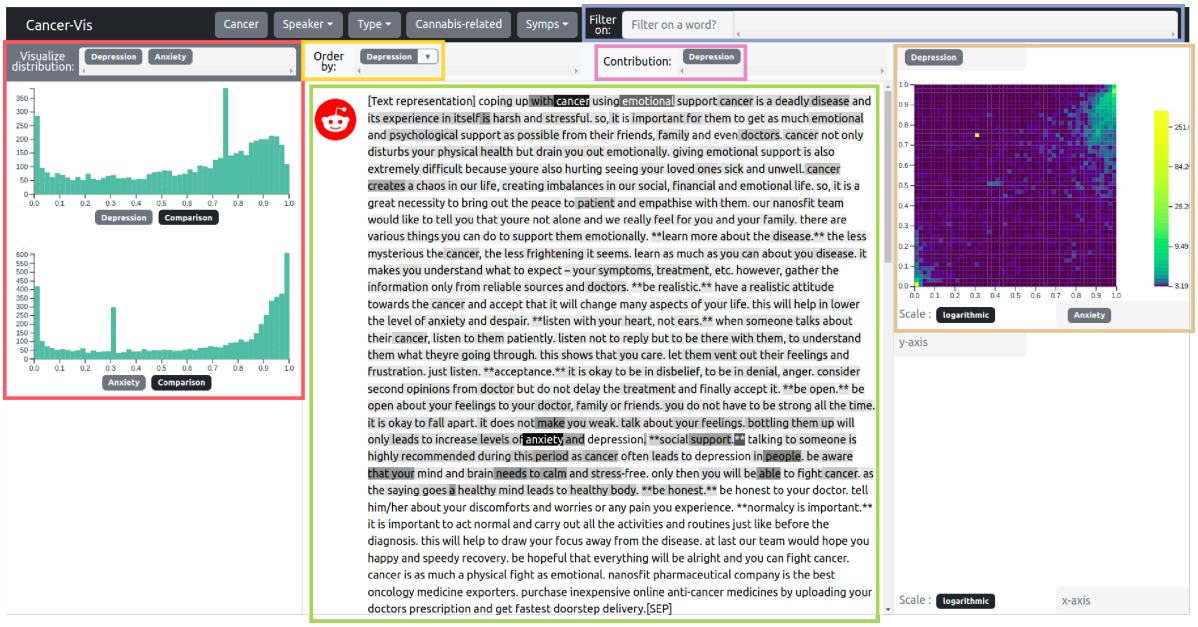


Figure 4.17 – *Vue globale de l'outil d'exploration des données issues des médias sociaux. Sur cette vue, les données ne sont pas filtrées (encadré bleu) mais elles sont ordonnées (encadré jaune).Sur les trois parties de la visualisation, nous observons à droite la distribution des prédictions pour les labels* `depression` *et* `anxiety` *dans l'encadré rouge. Dans la partie centrale, dans l'encadré vert, le score de contribution (ici, l'attention) pour le label* `depression` *(encadré rose) sur les textes (un seul est visible ici, les autres le sont à l'aide du défilement de la page). Enfin, sur la partie gauche, dans l'encadré doré, une carte de chaleur avec en ordonnée les prédictions pour la label* `depression` *et en abscisse les prédictions pour le label* `anxiety`.

ment les produits dérivés du cannabis sous de nombreuses appellations. Les pratiques autour des dérivés du cannabis sont variables. Certains patients les consomment en huile, en produits comestibles ou en suppositoire, quand d'autres les fument. Les molécules ciblées sont également différentes. On y trouve le CBD et le THC. Les effets attribuables à ces INM sont variables également. Elles sont mentionnées comme aidant pour la perte d'appétit, l'anxiété, la dépression, la douleur et les vomissements le plus souvent. Bien que les avis évoqués autour de ces pratiques soient en majorité positifs, il existe des patients pour lesquels elles n'ont pas eu d'effets ou eu des effets négatifs. Il est important de prendre en compte que les avis présents dans les publications ne correspondent qu'aux personnes ayant pris la parole et ne sont pas représentatifs de tous les patients. Au delà des avis, les patients ou proches de patients appuient, dans la quasi-totalité des cas, que la priorité est toujours de consulter son professionnel de santé. Il est aussi important de noter qu'il existe une part de publications à objectifs commerciaux. Enfin, les INM autres que les dérivés du cannabis ne sont mentionnées que trop rarement pour être explorées dans Cancer-Vis.

4.8.2 Outil d'exploration des données

L'outil d'exploration des données repose sur la qualité des données annotées et des classifieurs. Lors de la mise au point de notre cas d'étude, nous avons sélectionné les labels pour lesquels les performances étaient acceptables. Pour intégrer d'autres

labels, il faudrait revoir le processus de collecte et d'annotation. De plus, nous avons fait le choix, contrairement au chapitre précédent, de ne pas intégrer une exploration fine des performances du réseau de neurones car cet outil cible des "non spécialistes" de ces méthodes. Bien que des informations sur la distribution des prédictions issues des classifieurs ou une carte de chaleur comparant celles-ci est disponible, l'outil ne donne pas des scores de classification directement. Néanmoins, il met en lumière l'incertitude du réseau ou une ambiguïté dans les données. Les badges sous les textes permettent aussi d'identifier les erreurs de classification. Dans notre cas d'étude, nous avons constaté que le score d'attention est très souvent difficilement interprétable. Cet outil d'exploration permet donc de donner une intuition aux non spécialistes du fonctionnement des réseaux sans expliquer complètement les prédictions, ce qui nécessiterait le développement de fonctionnalités supplémentaires.

4.8.3 Méthodologie

La méthodologie proposée dans ce chapitre comporte certaines limites. Nous allons énoncer celles-ci et les traiter dans le cas général puis, plus spécifiquement, dans notre cas d'étude.

Données disponibles et entraînement : la première limite concerne les données disponibles issues des médias sociaux. L'annotation manuelle sur des labels dichotomiques est efficace. Pour les autres labels, la tâche est difficile. Dans un cadre général, le choix des concepts à identifier et des labels concernant les publications doit être mesuré au regard des volumes de données qu'il est possible de récolter. Des concepts trop peu renseignés ne permettent pas d'entraîner des classifieurs. Des labels trop ambiguës sont difficilement annotables par des humains. Dans notre cas d'étude par exemple, le concept d'INM ne donnait pas accès à assez de données pour les traiter toutes indépendamment. Les labels concernant l'émotion, le locuteur ou la controverse étaient ambiguës pour les annotateurs. Pour régler ces problématiques, une solution pourrait être, pour les labels difficiles, de ne pas entraîner de classifieur supervisé mais de se limiter à des recherches par mots clés ou se baser sur des règles métier comme proposé.

Exportation et labélisation des données : une deuxième limite réside dans le processus d'exportation des données. Généralement, elles sont exportées en amont de la labélisation par les annotateurs. Dans notre méthodologie, nous souhaitions découvrir de nouvelles INM et avons laissé les annotateurs exporter les données au fil de leurs lectures. Nous constatons qu'à l'issu de l'annotation, aucune nouvelle INM n'a été identifiée en quantité suffisante pour être étudiée plus précisément. Ainsi, on peut se questionner sur la nécessité de demander aux annotateurs d'être responsables de l'exportation.

Une autre limite réside dans la quantité des labels que les annotateurs ont du renseigner. La tâche était trop difficile et rébarbative pour que les annotateurs ne commettent pas d'erreur, notamment en oubliant des labels. La construction d'un second jeu de données annoté par des règles métier a donc été essentiel. Celui-ci aurait pu être utilisé avant l'annotation manuelle pour limiter le travail des annotateurs qui n'auraient eu qu'à valider les annotations.

Durée de vie des données : comme mentionné dans la section 4.7.1, certaines données peuvent disparaître suite à la modération ou à la suppression par les auteurs. Dans un intervalle de deux mois entre l'exportation des données et leur analyse, nous avons pu observer qu'un nombre considérable de messages n'étaient plus disponibles. Par exemple, il sera difficile d'étudier des phénomènes comme les "fake news", à moins d'avoir accès aux messages supprimés. Dans notre cas d'étude, nous avons pu observer mais pas quantifier l'importance de la modération existant autour des INM. La sensibilité du concept étudié doit donc être prise en compte lors de la mise en place de la méthodologie.

4.9 Conclusions

En conclusion, concernant l'étude des INM sur les médias sociaux, nous avons présenté deux outils. Le premier (Cancer-Annot) permet l'extraction des données et l'annotation des données issues des médias sociaux. Cet outil a permis de produire un premier jeu de données qui a été complété par un deuxième construit à partir de règles. Le second outil (Cancer-Vis) concerne l'exploration des données issues des médias sociaux. Cet outil est basé sur la visualisation des prédictions de réseaux de neurones appliqués à la classification de textes selon différents labels. Nous avons montré son intérêt dans le cadre de deux cas d'étude portant sur l'utilisation des dérivés du canabis et sur leur lien avec l'anxiété et la dépression.

Ces travaux préliminaires ont ouvert des perspectives prometteuses. Des travaux futurs devraient porter sur l'amélioration des performances de classification, sur la combinaison des deux outils pour mettre en place une démarche d'apprentissage actif [Set09]. Profiter de l'espace de représentation des messages est également une perspective intéressante en utilisant par exemple les graphes de relations que l'on peut extraire des réseaux sociaux. Enfin, la prise en compte de la temporalité est importante pour identifier de nouveaux concepts en temps réel, par exemple dans une perspective de veille.

CHAPITRE 5

CONCLUSIONS ET PERSPECTIVES

Sommaire

5.1 Conclusions

Dans cette section, nous résumons les contributions et soulignons les apports des travaux de cette thèse. La section 5.1.1, concerne l'état de l'art des techniques de visualisation utilisées pour augmenter l'interprétabilité des réseaux de neurones en TAL. La section 5.1.2 concerne la visualisation de la frontière de décision pour la production d'explications des prédictions. La section 5.1.3 concerne l'exploration des données issues des médias sociaux.

5.1.1 Visualiser et interpréter les réseaux de neurones en classification de textes

Dans l'état de l'art présenté dans le chapitre 2, nous nous sommes focalisés sur le foisonnement récent des techniques de visualisation de données appliquées à l'interprétation des réseaux de neurones en classification de textes.

Tout d'abord, nous nous sommes intéressés aux différentes méthodes d'apprentissage profond utilisées pour le TAL. Nous avons été attentifs à décrire les architectures de base. Nous avons produit des figures originales et uniformisées en présentant leur chronologie d'apparition.

Dans une seconde partie, nous avons mis en avant les problématiques liées à l'interprétabilité, la transparence et l'explicabilité, qui ont été très étudiées récemment. Nous avons souligné le manque de consensus dans la communauté, notamment dans la présentation des concepts et des définitions. Nous avons donné nos propres interprétations de ces concepts, qui ont servi de fil conducteur aux travaux des chapitres 3 et 4. Nous avons ensuite listé plusieurs familles de méthodes d'explication des prédictions afin de positionner la suite de nos contributions.

Pour finir, nous nous sommes focalisés sur les techniques de visualisation de données conçues pour augmenter l'interprétabilité des réseaux de neurones en TAL et qui sont au cœur de cette thèse. Nous avons lié ces méthodes aux architectures auxquelles elles s'appliquent, avant de présenter des méthodes agnostiques aux architectures. Les travaux présentés dans le chapitre 3 se situent dans cette dernière famille de méthodes.

5.1.2 EBBE-Text : Expliquer les prédictions d'un classifieur neuronal pour les données textuelles

Le chapitre 3 présente l'outil de visualisation de la frontière de décision EBBE-Text.

Nous avons défini une nouvelle approche pour visualiser la frontière de décision. Plus spécifiquement, nous détaillons différentes méthodes mises en place pour construire des localités dans un espace de représentation et produire une visualisation d'un graphe de proximité possédant des données de frontière. Ces méthodes peuvent être utilisées indépendamment l'une de l'autre.

Nous avons développé l'outil de visualisation EBBE-Text qui s'adresse à des spécialistes des méthodes de classification automatique. En agrégeant différentes informations issues ou non du fonctionnement du modèle, cet outil permet de comprendre

les prédictions d'un réseau de neurones en TAL, grâce notamment à la possibilité de classifier des données à la volée. Il se compose de deux vues : (1) une vue globale d'un ensemble de localités et (2) une vue par localité permettant d'explorer finement l'espace de représentation des données et de visualiser la frontière de décision comme une ligne. Cet outil a donné lieu à une évaluation poussée, quantitative et qualitative sur des données réelles, avec des utilisateurs experts et débutants en classification. Deux études de cas soulignent aussi l'intérêt métier.

Ces travaux ont donné lieu à trois publications, deux dans une conférence française [Del+21a ; Del+21b] et une dans une revue internationale [Del+].

5.1.3 Cancer-Annot et Cancer-Vis : Explorer des concepts à partir des données issues des médias sociaux

Ces travaux ont été abordés lors de la dernière année de la thèse et restent préliminaires. Les apports du chapitre 4 concernant l'étude des INM dans les médias sociaux sont divisés en deux parties.

Tout d'abord, nous avons décrit une nouvelle méthodologie pour extraire et annoter les données issues des médias sociaux à l'aide de l'outil Cancer-Annot et d'un ensemble de règles métier. Les données annotées servent ensuite à entraîner des classifieurs.

Nous avons également développé l'outil Cancer-Vis qui s'adresse à des non spécialistes des méthodes de classification automatique. Cet outil permet une exploration des données issues des médias sociaux à l'aide de labels produits par des classifieurs. L'évaluation de cet outil reste préliminaire mais il a déjà permis une première étude des INM dérivées du cannabis lors de leur utilisation dans le cadre du cancer, à partir des données issues de Reddit. Cette étude met en avant qu'il existe une proportion considérable de patients relatant des effets positifs liés à l'utilisation de ces INM.

5.2 Perspectives

Dans cette section, nous présentons les perspectives de nos travaux selon deux axes : la visualisation de la frontière de décision pour la production d'explications des prédictions est présentée dans la section 5.2.1 ; l'exploration des données issues des médias sociaux est abordée dans la section 5.2.2.

5.2.1 Frontière de décision et explications

La visualisation de la frontière de décision des réseaux de neurones assiste un utilisateur dans la compréhension des réseaux, leur débogage et la construction d'explications des prédictions. Dans ce cadre, la conservation des distances à la frontière dans la visualisation de l'espace de représentation, lui-même de grande dimension, est un atout. L'exploration des localités ajoute à cela la possibilité de comparer entre eux des exemples similaires. Nous prévoyons de traiter le cas de la classification multiclasse, alors qu'EBBE-Text ne traite que le cas binaire (section 5.2.1.1), de créer de nouvelles métriques d'explicabilité (section 5.2.1.2) et de générer de nouveaux exemples pour

une plus grande explicabilité des réseaux de neurones (section 5.2.1.3) à l'aide de nouvelles visualisations (section 5.2.1.4).

5.2.1.1 Classification multiclasse

EBBE-Text s'intéresse seulement à la visualisation de la frontière de décision pour des classifications binaires. La conservation des distances dans une visualisation en deux dimensions est possible car le cas binaire garantit l'existence d'une unique direction optimale rapprochant ou éloignant une prédiction de la frontière de décision. Ce n'est plus vrai dans le cas multiclasse (cf. voir Fig 5.1). EBBE-Text propose également un placement cohérent des données dans la visualisation de la frontière de décision car il utilise le graphe de proximité issu de UMAP. Ce graphe capture la topologie fondamentale de l'ensemble de données.

Une première perspectives consiste à représenter, dans un espace à deux dimensions, plusieurs frontières de décision, chacune relative à une classe. La méthode naïve consiste à utiliser, comme dans EBBE-Text, UMAP pour positionner les données, puis à placer les frontières de décisions afin qu'elles séparent au mieux les données en fonction de leurs labels. Cependant, cette approche soulève deux problématiques. La première concerne les informations erronées qui pourraient être extraites de la représentation. Par exemple, certaines données ne seront pas positionnées correctement vis-à-vis d'une ou plusieurs frontières (figure 5.1). Il faudra donc gérer ce type d'ambiguïté à l'aide de techniques algorithmiques (*e.g.* en déformant le positionnement issu de UMAP ou en utilisant une autre technique), visuelles (*e.g.* en trouvant d'autres modes de représentation) et/ou interactives (*e.g.* en proposant une approche de type "détail à la demande" [Shn03]). La seconde problématique concerne l'encombrement visuel que peut produire la représentation d'un grand nombre de données et de frontières. Ici aussi, diverses solutions pourront être envisagées, comme un positionnement plus adéquat (*e.g.* en positionnant certains objets par rapport à d'autres), des méthodes intelligentes d'interaction (permettant d'obtenir rapidement des informations non disponibles), des modes d'agrégation [Sev+21 ; Str+17 ; Min+17 ; Del+].

Une première piste que nous explorerons pour résoudre le **problème des informations erronées** sera de mettre en place un système de "détail à la demande" [Shn03] permettant d'ouvrir une vue détaillée dans laquelle les distances réelles entre certains objets sont présentées. Il est par exemple possible de sélectionner une donnée et de produire une visualisation dans laquelle toutes les distances entre cette donnée et les autres sont exactes, ainsi que les distances entre les données et une frontière de décision (figure 5.2(b)). D'autres exemples de conservation de distances possibles sont présentés dans la figure 5.2. Pour toutes ces méthodes certaines distances sont correctes et d'autres non. Il faudra donc trouver des techniques visuelles permettant de distinguer les informations valides des artefacts visuels issus de la construction des diagrammes.

Pour ces vues détaillées, la **problématique d'encombrement visuel** devra aussi être abordée. D'un côté, la construction de localités dans l'espace de représentation est essentielle pour éviter l'encombrement visuel qu'un trop grand nombre de données et de frontières apportent. D'un autre côté, ne pas construire de localité permet d'avoir

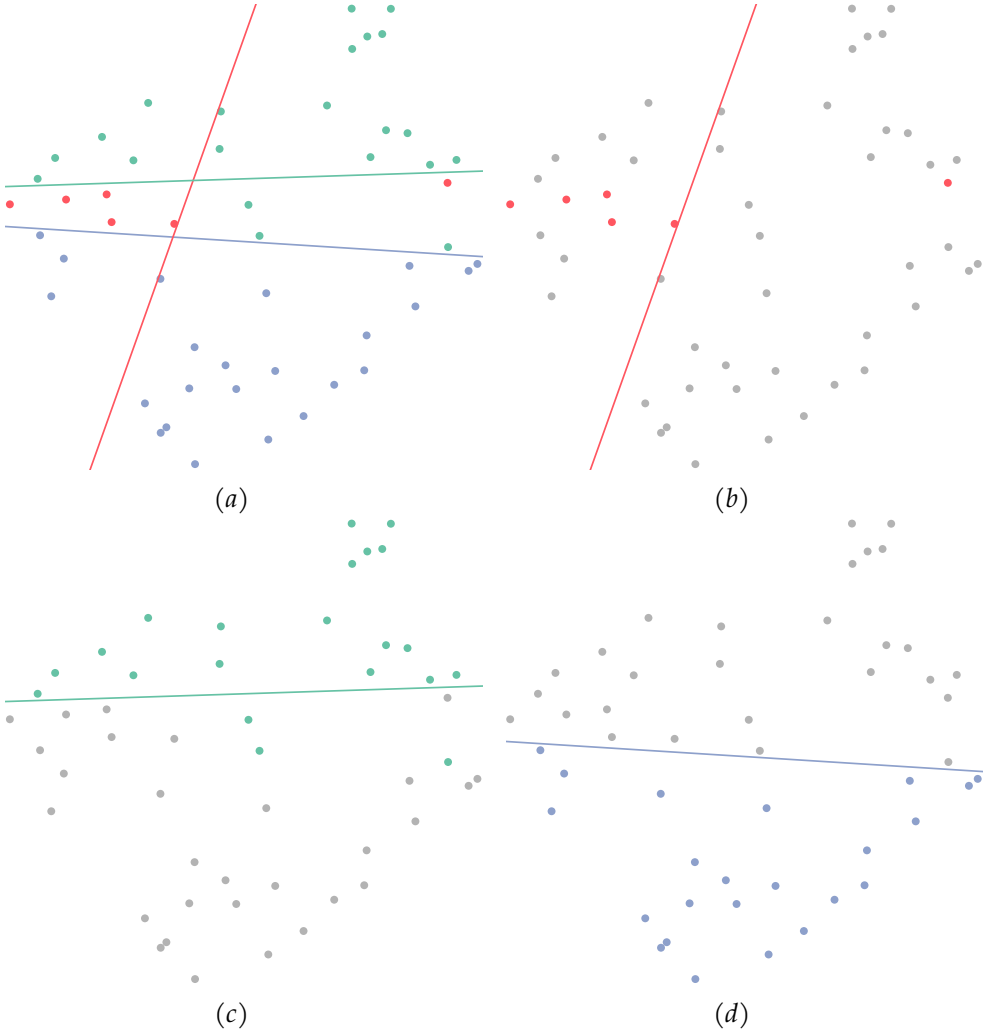


Figure 5.1 – *Projection dans un espace à deux dimensions d'un jeu de données et placement de frontières de décision linéaires d'un classifieur à trois modalités.* (*a*) *Les couleurs des points donnent la classe de la donnée les concernant.* (*b*) *La frontière (**droite rouge**) ne sépare les données correctement dans l'espace de dimension réduite puisque, des deux côtés de la frontière, des données sont mal positionnées.* (*c*) *La frontière (**droite verte**) ne sépare pas les données correctement dans l'espace de dimension réduite puisque, pour un côté de la frontière, des données sont mal positionnées.* (*d*) *La frontière (**droite bleue**) sépare correctement les données dans l'espace de dimension réduite mais cela ne garantit pas que les distances sont fidèles à celles de l'espace de représentation d'origine.*

une vue plus globale des données. Dans EBBE-Text, nous construisons des localités pour visualiser la frontière de décision, à l'aide de la topologie des données. Nous réduisons ainsi l'encombrement visuel. Si l'on ne veut pas s'appuyer sur cette topologie mais plutôt sur l'environnement de la donnée sélectionnée dans une vue détaillée, un système de filtres permet de sélectionner le nombre de données à afficher en fonction de divers critères (*e.g.* les k plus proches voisins de la donnée sélectionnée dans l'espace de représentation ou les données contenant certains mots). De la même manière, si les frontières de décisions sont trop nombreuses, on peut afficher uniquement celles apportant le plus d'informations.

Les pistes envisagées dans cette section se basent sur la sélection d'une donnée

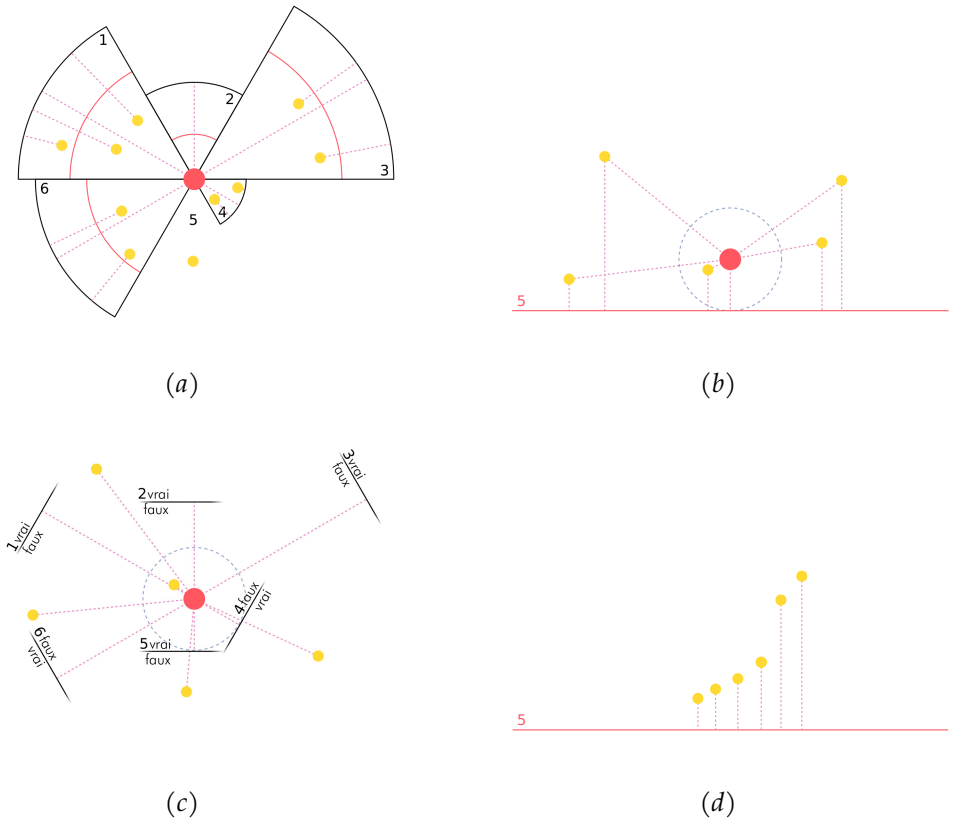


FIGURE 5.2 – *Construction de visualisations expliquant les prédictions pour la classification multiclasse. Les* ***lignes discontinues roses*** *correspondent aux distances qui peuvent être préservées par rapport aux* ***objets rouges*** *de référence sélectionnés (donnée ou frontière). Hormis ces dernières, les distances entre les objets sont erronées. Pallier ce problème est l'un des challenges majeurs des perspectives présentées. Les* ***cercles bleus*** *sont une visualisation des sphères d'explications justifiées de sortie. Dans les quatre figures, toutes les données sont prédites dans la classe 5. (a) Pour une donnée de référence, on obtient sa distance à chaque frontière de décision ("un contre tous") à l'aide des secteurs circulaires représentant les classes. On observe ici que la classe la plus probable, après la classe 5, est la classe 4. Dans chaque secteur circulaire, l'arc de cercle noir est la frontière de décision. On observe dans ces secteurs combien de données sont plus proches, de la frontière concernée, que la donnée de référence ne l'est (e.g. deux voisins da la donnée référence sont plus proches de changer de la classe 5 vers 6 que cette donnée ne l'est). Les* ***arcs de cercle rouges****, dans les sections circulaires, représentent la distance de la donnée de référence à la frontière la plus proche. Cela montre quelles sont les données qui nécessitent moins de modification, pour changer de prédiction, que la donnée de référence n'en nécessite. On observe qu'une donnée ne se trouve dans aucun secteur circulaire. Pour cette donnée, les probabilités d'appartenir aux autres classes sont toujours inférieures à celles de la donnée de référence. Dans cette figure, les distances entre les données sont toutes erronées. (b) On observe qu'il est possible de conserver pour une donnée de référence sa distance à toutes les frontières et à toutes les données. Il n'existe qu'une seule donnée plus proche de la donnée de référence que de son projeté sur la frontière de décision de la classe 5. On observe également, comme dans la figure (a) que les classes les plus probables pour la donnée de référence sont les classes 5, 4 et 2. (c) On observe qu'il est possible de conserver pour chaque donnée, sa distance à une donnée de référence et sa distance à la frontière de décision de la classe 5 (5 contre tous). Toutes les positions des données (hormis celle de référence) vis-à-vis des frontières sont erronées. Dans la figure (d), on observe les distances des données à une frontière.*

pour la construction d'une représentation correcte de certaines distances liées à cette donnée. À l'opposé d'EBBE-Text qui construit des localités cohérentes dans l'espace de représentation et n'a pas besoin de sélection, ce type d'approche devrait pouvoir traiter le problème de classification multiclasse.

5.2.1.2 Explications justifiées

Dans nos travaux, les données visualisées sont issues de l'espace de représentation construit par le réseau de neurones. Dans EBBE-Text, nous signifions à l'aide d'un code couleur que deux voisins dans l'espace de représentation peuvent ne pas être prédits de la même manière. Nous ne vérifions pas si un segment reliant deux données depuis l'espace d'entrée du réseau de neurones est sécant à la frontière de décision. Dit autrement, est-ce que l'interpolation linéaire de deux points d'une même classe peut donner une autre classe ? Laugel et al. [Lau+19] proposent une méthode pour créer des explications justifiées (section 2.3.2.3) pour résoudre ce problème. Cette méthode peut être appliquée à la projection en deux dimensions de l'espace de représentation associée à une ou plusieurs frontières de décision. Boggust, Carter et Satyanarayan [BCS19], pour comparer des plongements lexicaux, cherchent les k plus proches voisins d'une donnée dans les deux espaces de représentation. Comme indiqué dans la section 5.2.1.1, cela peut servir à réduire l'encombrement visuel et à créer des localités. Une des perspectives pour visualiser la frontière de décision dans un espace à deux dimensions est donc d'y inclure la visualisation d'explications justifiées. Cette visualisation peut inclure la projection d'un chemin de l'espace d'origine dans l'espace construit par le réseau de neurones.

En combinant les travaux de [Lau+19] et [BCS19], on peut fixer un seuil k et produire pour les k plus proches voisins d'une donnée le caractère justifié des explications. Ce caractère peut prendre plusieurs formes. Soit deux données A et B. Selon Laugel et al. [Lau+19], si A et B ne sont pas classées de la même manière, elles ne produisent pas d'explication justifiée. De même, si A et B sont classées de la même manière et que le segment de l'une à l'autre dans l'espace d'entrée traverse la frontière de décision, alors elles ne produisent pas d'explication justifiée. Si ce segment ne traverse pas la frontière de décision alors elles produisent une explication justifiée. B justifie A et inversement. Laugel et al. [Lau+19] introduisent également une notion d'explication en chaîne. Ces explications fonctionnent de proche en proche et expliquent une prédiction en utilisant des chemins liant deux données.

Les notions précédentes s'appliquent à l'espace d'entrée mais de nouvelles métriques peuvent aussi être définies dans l'espace construit par le réseau de neurones. Nous appelons ces explications proposées par Laugel et al. [Lau+19] des explications justifiées d'entrée. Nous proposons une nouvelle perspective : les explications justifiées de sortie dans l'espace de représentation construit par le réseau de neurones. Par exemple, pour deux données A et B, si A est plus proche de B que de sa projection orthogonale A' sur le plan de la frontière de décision, l'explication de A par B est justifiée en sortie. Chaque donnée existante ou générée dans la n-sphère de centre A et de rayon $d(A, A')$ est également justifiée en sortie, où n est la dimension de l'espace de représentation et $d(A, A')$ la distance entre A et A' dans cet espace. On ajoute donc des modalités au caractère justifiable d'une explication.

Une autre métrique pour expliquer les prédictions est la notion de score de direction à la frontière dans l'espace construit par le réseau de neurones. Soit A une donnée, A' sa projection orthogonale sur une frontière entre deux classes, $v(A)$ l'ensemble des voisins de A et $B \in v(A)$, on calculer pour tout B l'angle $\widehat{AA'B}$. À partir de ces angles, on peut déterminer quelles sont les données qui sont positionnée au plus près du chemin de A vers la frontière de décision. Cet angle représente le score de direction à la frontière et montre à quel point une donnée B explique la prédiction de A. Une approche topologique pourra à nouveau nous permettre de considérer les angles les plus pertinents malgré la grande dimension.

Ces métriques peuvent être intégrées dans une visualisation de la frontière de décision avec un encodage spécifique pour identifier, dans l'espace à deux dimensions, les données permettant d'expliquer la donnée sélectionnée depuis l'espace d'entrée ou de sortie.

Pour répondre à la problématique de l'encombrement visuel, on peut aussi construire le seuil k de voisins à afficher en fonction des données expliquant A. k peut alors être fixé en fonction d'un rapport entre les explications justifiées et non justifiées. Plus k augmente, moins la proportion d'explications justifiées sera grande. On peut également fixer k à l'explication non justifiée D la plus proche de A. D est alors le k-ième voisin de A et le seul à ne pas produire d'explication justifiée. Il peut cependant exister des explications justifiées d'entrée plus éloignées de A que D ne l'est.

5.2.1.3 Espace unique, génération d'exemples et auto-encodage

Comme présenté dans la section 3.8, une des perspectives concerne la génération de données. Les réseaux AE, grâce à un espace de données cohérent, accroissent leur interprétabilité. Une perspective est de générer, au sein des n-sphères ou sur les chemins entre les données, de nouvelles données explicatives. Ces dernières, artificielles, peuvent augmenter le nombre d'explications disponibles pour une donnée et la nature de l'explication. Les réseaux AE permettent la génération de données et donc la visualisation de données artificielles qu'elles soient textuelles ou non. L'approche que nous proposons dans cette section s'applique à tous types de données.

5.2.1.4 Localité d'une donnée et ses explications justifiées de sortie

En suivant les pistes évoquées dans les sections 5.2.1.1 et 5.2.1.2, on peut produire, dans la visualisation des frontières de décision (figures 5.2(a)-5.2(d)), des sphères d'explications justifiées de sortie (figure 5.2(b) et 5.2(c)), des chemins d'explications justifiées d'entrée ainsi que les métriques d'explications justifiées pour une localité de k voisins d'une donnée A (seule ou dans un couple frontière-donnée). Cela permet d'un coup d'œil de voir quelles sont les données qui justifient A pour quelles frontières et dans quels espaces tout en offrant la possibilité de générer des données dans ces sphères.

5.2.2 Exploration des données issues des médias sociaux

La principale limitation des travaux présentés dans le chapitre 4 est le manque de données disponibles pour l'analyse. Évidemment, pour d'autres thématiques que les INM utilisées dans le cas d'un cancer, on peut espérer en récolter beaucoup plus.

En outre, nous avons largement commenté la difficulté de catégoriser les données manuellement avant de réaliser l'étape d'apprentissage. Cette tâche est très coûteuse et il convient de la réduire avec des systèmes semi-automatiques. Dans ce contexte, des méthodes d'apprentissage actif, pendant lesquelles l'oracle intervient pour choisir les meilleurs exemples à étiqueter, s'avèrent prometteuses [Mer+18]. L'intuition est la suivante : en choisissant les exemples intelligemment et non aléatoirement, les modèles devraient s'améliorer avec moins d'efforts pour l'oracle et donc à moindre coût (*i.e.* avec moins d'exemples annotés). Il pourrait être intéressant d'intégrer un tel processus dans l'outil EBBE-Text décrit dans le chapitre 3.

Une perspective intéressante consiste à assembler l'outil EBBE-Text et Cancer-Viz pour profiter pleinement de l'espace de représentation construit afin de naviguer entre les textes. Par exemple, pour un texte sélectionné, un utilisateur pourrait retrouver ses plus proches voisins et se déplacer de proche en proche. Dans notre cas d'étude, un professionnel de santé pourrait plus facilement creuser l'émergence d'un nouveau concept comme une nouvelle INM peu citée par exemple en retrouvant les messages similaires.

Une autre perspective concerne la quantification des mentions. Actuellement, notre outil nous permet d'explorer les mentions en naviguant dans l'ensemble des messages récoltés mais nous ne proposons pas de visualisation qui donne des statistiques sur les labels et les concepts. Cela permettrait de produire une vision globale des données issues des réseaux sociaux. Par exemple, dans notre cas d'étude, il pourrait être intéressant de connaître la proportion des messages mentionnant une INM en particulier, et combien relatent un ressenti positif vis-à-vis de cette INM.

Une dernière perspective concerne la prise en compte des aspects temporels. Il pourrait être intéressant de monitorer les médias sociaux et détecter l'émergence de nouveaux concepts au cours du temps, en labellisant les données en continu pour alerter les utilisateurs sur l'augmentation ou la diminution d'un concept précis et en détectant des données s'éloignant des classes prédéfinies et pouvant représenter de nouveaux concepts. Dans notre cas d'étude, les professionnels de santé pourraient par exemple détecter l'utilisation de nouvelles INM, suivre l'adhésion des patients à la recommandation de nouvelles INM, *etc.*

Plus généralement, il est nécessaire de développer de nouveaux outils pour extraire de l'information médicale des médias sociaux de manière (semi-)automatique. Nous pourrions par exemple étudier les influences mises en jeu dans le processus de décision du patient et qui sont souvent mentionnées dans ces médias [LSC22]. Nous pourrions également mesurer les conséquences de l'apparition des médias sociaux dans la relation Patient/Médecin [For+21]. Un autre aspect important que nous avons également identifié lors de nos explorations des mentions d'INM est le risque de désinformation ou de propagation d'informations erronées [Ska+22], qu'il faut détecter automatiquement

pour aider les modérateurs. Il s'agit d'un pré-requis pour que les médias sociaux deviennent un support pertinent à une médecine ouverte, participative et collaborative.

BIBLIOGRAPHIE

[Alb+20] Emanuele ALBINI, Piyawat LERTVITTAYAKUMJORN, Antonio RAGO et Francesca TONI. « Dax : Deep argumentative explanation for neural networks ». In : *arXiv preprint arXiv :2012.05766* (2020) (cf. pages 31, 34, 36, 41, 43).

[Arr+17] Leila ARRAS, Grégoire MONTAVON, Klaus-Robert MÜLLER et Wojciech SAMEK. « Explaining recurrent neural network predictions in sentiment analysis ». In : *arXiv preprint arXiv :1706.07206* (2017) (cf. pages 30, 31, 36, 43, 44, 47).

[Arr+19] Leila ARRAS, Ahmed OSMAN, Klaus-Robert MÜLLER et Wojciech SAMEK. « Evaluating recurrent neural network explanations ». In : *arXiv preprint arXiv :1904.11829* (2019) (cf. page 34).

[ATS17] Hadeer AHMED, Issa TRAORE et Sherif SAAD. « Detection of online fake news using n-gram analysis and machine learning techniques ». In : *International Conference on Intelligent, Secure, and Dependable Systems in Distributed and Cloud Environments (ISDDC)*. 2017, p. 127-138. DOI : https://doi.org/10.1007/978-3-319-69155-8_9 (cf. pages 77, 80).

[ATS18] Hadeer AHMED, Issa TRAORE et Sherif SAAD. « Detecting opinion spams and fake news using text classification ». In : *Security and Privacy* 1.1 (2018). DOI : https://doi.org/10.1002/spy2.9 (cf. pages 77, 80).

[Bac+15] Sebastian BACH, Alexander BINDER, Grégoire MONTAVON, Frederick KLAUSCHEN, Klaus-Robert MÜLLER et Wojciech SAMEK. « On Pixel-Wise Explanations for Non-Linear Classifier Decisions by Layer-Wise Relevance Propagation ». In : *PLOS ONE* 10.7 (juill. 2015), p. 1-46. DOI : 10.1371/journal.pone.0130140. URL : https://doi.org/10.1371/journal.pone.0130140 (cf. page 30).

[Bar+20] Alejandro BARREDO ARRIETA, Natalia DÍAZ-RODRÍGUEZ, Javier DEL SER, Adrien BENNETOT, Siham TABIK, Alberto BARBADO, Salvador GARCIA, Sergio GIL-LOPEZ, Daniel MOLINA, Richard BENJAMINS, Raja CHATILA et Francisco HERRERA. « Explainable Artificial Intelligence (XAI) : Concepts, taxonomies, opportunities and challenges toward responsible AI ». In : *Information Fusion* 58 (2020), p. 82-115. ISSN : 1566-2535. DOI : https://doi.org/10.1016/j.inffus.2019.12.012. URL : https://www.sciencedirect.com/science/article/pii/S1566253519308103 (cf. page 34).

[Bar17] Joana M BARROS. « Text Mining from Social Media for Public Health Applications ». In : *Proceedings of the 2017 International Conference on Digital Health*. 2017, p. 223-224 (cf. page 95).

[BCS19] Angie BOGGUST, Brandon CARTER et Arvind SATYANARAYAN. « Embedding comparator : Visualizing differences in global structure and local neighborhoods via small multiples ». In : *arXiv preprint arXiv :1912.04853* (2019) (cf. pages 36, 40, 41, 131).

[Bea+20] Valérie BEAUDOUIN, Isabelle BLOCH, David BOUNIE, Stéphan CLÉMENÇON, Florence D'ALCHÉ-BUC, James EAGAN, Winston MAXWELL, Pavlo MOZHAROVSKYI et Jayneel PAREKH. « Flexible and context-specific AI explainability : a multidisciplinary approach ». In : *SSRN* (2020). DOI : 10.2139/ssrn.3559477 (cf. pages 23, 27).

[Bec+13] Francois BECK, Viêt NGUYEN-THANH, Jean-Baptiste RICHARD et Émilie RENAHY. « Usage d'internet : les jeunes, acteurs de leur santé ? » In : *Agora débats/jeunesses* 1 (2013), p. 102-112 (cf. page 94).

[Bén+21] Clément BÉNARD, Gérard BIAU, Sébastien DA VEIGA et Erwan SCORNET. « SHAFF : Fast and consistent SHApley eFfect estimates via random Forests ». In : *arXiv preprint arXiv :2105.11724* (2021) (cf. pages 31, 32, 36).

[Ber20] Matthew BERGER. « Visually Analyzing Contextualized Embeddings ». In : *2020 IEEE Visualization Conference (VIS)*. IEEE. 2020, p. 276-280 (cf. pages 36, 40).

[Boj+17] Piotr BOJANOWSKI, Edouard GRAVE, Armand JOULIN et Tomas MIKOLOV. « Enriching Word Vectors with Subword Information ». In : *Transactions of the Association for Computational Linguistics* 5 (2017), p. 135-146. DOI : 10.1162/tacl_a_00051. URL : https://aclanthology.org/Q17-1010 (cf. page 13).

[Bou16] Djallel BOUNEFFOUF. « Exponentiated Gradient Exploration for Active Learning ». In : *Computers* 5.1 (2016). ISSN : 2073-431X. DOI : 10.3390/computers5010001 (cf. page 86).

[Bow+15] Samuel R BOWMAN, Luke VILNIS, Oriol VINYALS, Andrew M DAI, Rafal JOZEFOWICZ et Samy BENGIO. « Generating sentences from a continuous space ». In : *arXiv preprint arXiv :1511.06349* (2015) (cf. page 22).

[BR20] Ravali BOORUGU et G. RAMESH. « A Survey on NLP based Text Summarization for Summarizing Product Reviews ». In : *2020 Second International Conference on Inventive Research in Computing Applications (ICIRCA)*. 2020, p. 352-356. DOI : 10.1109/ICIRCA48905.2020.9183355 (cf. pages 3, 11).

[Bre+22] Clara BREIDENBACH, Paula HEIDKAMP, Kati HILTROP, Holger PFAFF, Anna ENDERS, Nicole ERNSTMANN et Christoph KOWALSKI. « Prevalence and determinants of anxiety and depression in long-term breast cancer survivors ». In : *BMC psychiatry* 22.1 (2022), p. 1-10 (cf. page 118).

[Bro+20] Tom BROWN, Benjamin MANN, Nick RYDER, Melanie SUBBIAH, Jared D KAPLAN, Prafulla DHARIWAL, Arvind NEELAKANTAN, Pranav SHYAM, Girish SASTRY, Amanda ASKELL, Sandhini AGARWAL, Ariel HERBERT-VOSS, Gretchen KRUEGER, Tom HENIGHAN, Rewon CHILD, Aditya RAMESH, Daniel ZIEGLER, Jeffrey WU, Clemens WINTER, Chris HESSE, Mark CHEN, Eric SIGLER, Mateusz LITWIN, Scott GRAY, Benjamin CHESS, Jack CLARK, Christopher BERNER, Sam MCCANDLISH, Alec RADFORD, Ilya SUTSKEVER et Dario AMODEI. « Language Models are Few-Shot Learners ». In : *Advances in Neural Information Processing Systems*. Sous la dir. de H. LAROCHELLE, M.

Ranzato, R. Hadsell, M. F. Balcan et H. Lin. T. 33. Curran Associates, Inc., 2020, p. 1877-1901 (cf. page 23).

[Bru+18] Gino Brunner, Yuyi Wang, Roger Wattenhofer et Michael Weigelt. « Natural language multitasking : analyzing and improving syntactic saliency of hidden representations ». In : *arXiv preprint arXiv :1801.06024* (2018) (cf. pages 36, 40).

[CDL16] Jianpeng Cheng, Li Dong et Mirella Lapata. « Long short-term memory-networks for machine reading ». In : *arXiv preprint arXiv :1601.06733* (2016) (cf. page 19).

[Cha+20] Angelos Chatzimparmpas, Rafael M. Martins, Ilir Jusufi et Andreas Kerren. « A survey of surveys on the use of visualization for interpreting machine learning models ». In : *Information Visualization* 19.3 (2020), p. 207-233. doi : 10.1177/1473871620904671 (cf. page 23).

[Che+18] Jianbo Chen, Le Song, Martin J Wainwright et Michael I Jordan. « L-shapley and c-shapley : Efficient model interpretation for structured data ». In : *arXiv preprint arXiv :1808.02610* (2018) (cf. pages 31, 32, 36).

[Cho+14] Kyunghyun Cho, Bart van Merrienboer, Dzmitry Bahdanau et Yoshua Bengio. « On the Properties of Neural Machine Translation : Encoder–Decoder Approaches ». In : *SSST@EMNLP*. 2014 (cf. pages 16, 18).

[Chu+14] Junyoung Chung, Çaglar Gülçehre, Kyunghyun Cho et Yoshua Bengio. « Empirical Evaluation of Gated Recurrent Neural Networks on Sequence Modeling ». In : *ArXiv* abs/1412.3555 (2014) (cf. page 18).

[CL20] Ian Covert et Su-In Lee. « Improving kernelshap : Practical shapley value estimation via linear regression ». In : *arXiv preprint arXiv :2012.01536* (2020) (cf. pages 31, 32, 36).

[Cla+19] Kevin Clark, Urvashi Khandelwal, Omer Levy et Christopher D Manning. « What does bert look at ? an analysis of bert's attention ». In : *arXiv preprint arXiv :1906.04341* (2019) (cf. pages 36, 45, 46, 84, 86).

[Cla+20] Kevin Clark, Minh-Thang Luong, Quoc V Le et Christopher D Manning. « Electra : Pre-training text encoders as discriminators rather than generators ». In : *arXiv preprint arXiv :2003.10555* (2020) (cf. page 23).

[CN19] François Carbonnel et Gregory Ninot. « Identifying Frameworks for Validation and Monitoring of Consensual Behavioral Intervention Technologies : Narrative Review ». In : *J Med Internet Res* 21.10 (oct. 2019), e13606. issn : 1438-8871. doi : 10.2196/13606. url : http://www.ncbi.nlm.nih.gov/pubmed/31621638 (cf. page 94).

[Con+17] Alexis Conneau, Holger Schwenk, Loic Barrault et Yann Lecun. « Very Deep Convolutional Networks for Text Classification ». In : *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics : Volume 1, Long Papers*. Valencia, Spain : Association for Computational Linguistics, avr. 2017, p. 1107-1116. url : https://aclanthology.org/E17-1104 (cf. page 16).

[CZJ20] Hanjie Chen, Guangtao Zheng et Yangfeng Ji. « Generating hierarchical explanations on text classification via feature interaction detection ». In : *arXiv preprint arXiv :2004.02015* (2020) (cf. pages 31, 32, 34, 36, 48).

[DAS19] Kedar DHAMDHERE, Ashish AGARWAL et Mukund SUNDARARAJAN. « The shapley taylor interaction index ». In : *arXiv preprint arXiv :1902.05622* (2019) (cf. pages 31, 32, 36).

[Del+] A. DELAFORGE, J. AZÉ, S. BRINGAY, C. MOLLEVI, A. SALLABERRY et M. SERVAJEAN. « EBBE-Text : Explaining Neural Networks by Exploring Text Classification Decision Boundaries ». In : *IEEE Transactions on Visualization and Computer Graphics* (*early access*) 01 (), to appear. ISSN : 1941-0506. DOI : 10.1109/TVCG.2022.3184247 (cf. pages 5, 7, 84, 127, 128).

[Del+21a] Alexis DELAFORGE, Jérôme AZÉ, Arnaud SALLABERRY, Maximilien SERVAJEAN, Sandra BRINGAY et Caroline MOLLEVI. « EBBE-Text : Visualisation de la frontière de décision des réseaux de neurones en classification automatique de textes ». In : *Revue des Nouvelles Technologies de l'Information* EGC, RNTI-E-37 (2021), p. 169-180 (cf. pages 5, 7, 78, 127).

[Del+21b] Alexis DELAFORGE, Jérôme AZÉ, Arnaud SALLABERRY, Maximilien SERVAJEAN, Sandra BRINGAY et Caroline MOLLEVI. « Expliquer les prédictions des réseaux de neurones par l'exploration de l'espace de représentation et de la frontière de décision à l'aide d'EBBE-Text ». In : *Revue des Nouvelles Technologies de l'Information* Extraction et Gestion des Connaissances, RNTI-E-37 (2021), p. 485-492 (cf. pages 5, 7, 127).

[Dev+18] Jacob DEVLIN, Ming-Wei CHANG, Kenton LEE et Kristina TOUTANOVA. « Bert : Pre-training of deep bidirectional transformers for language understanding ». In : *arXiv preprint arXiv :1810.04805* (2018) (cf. pages 14, 20, 35, 45).

[Die+18] Dennis DIEFENBACH, Vanessa LOPEZ, Kamal SINGH et Pierre MARET. « Core techniques of question answering systems over knowledge bases : a survey ». In : *Knowledge and Information Systems* 55.3 (juin 2018), p. 529-569. ISSN : 0219-3116. DOI : 10.1007/s10115-017-1100-y. URL : https://doi.org/10.1007/s10115-017-1100-y (cf. page 11).

[DK17] Finale DOSHI-VELEZ et Been KIM. « A roadmap for a rigorous science of interpretability ». In : *arXiv preprint arXiv :1702.08608* 2 (2017), p. 1 (cf. page 96).

[DM04] Lynda C DOWARD et Stephen P MCKENNA. « Defining patient-reported outcomes ». In : *Value in health* 7 (2004), S4-S8 (cf. page 4).

[DSZ16] Anupam DATTA, Shayak SEN et Yair ZICK. « Algorithmic Transparency via Quantitative Input Influence : Theory and Experiments with Learning Systems ». In : *2016 IEEE Symposium on Security and Privacy* (*SP*). 2016, p. 598-617. DOI : 10.1109/SP.2016.42 (cf. pages 31, 32, 36).

[DVC21] Jie DU, Chi-Man VONG et C. L. Philip CHEN. « Novel Efficient RNN and LSTM-Like Architectures : Recurrent and Gated Broad Learning Systems and Their Applications for Text Classification ». In : *IEEE Transactions on Cybernetics* 51.3 (2021), p. 1586-1597. DOI : 10.1109/TCYB.2020.2969705 (cf. page 23).

[DWB20] Joseph F DEROSE, Jiayao WANG et Matthew BERGER. « Attention flows : Analyzing and comparing attention mechanisms in language models ». In : *IEEE Transactions on Visualization and Computer Graphics* 27.2 (2020), p. 1160-1170 (cf. pages 36, 45, 46, 84, 86).

[El-+20] Wafaa El-Kassas, Cherif Salama, Ahmed Rafea et Hoda Mohamed. « Automatic Text Summarization : A Comprehensive Survey ». In : *Expert Systems with Applications* 165 (juill. 2020), p. 113679. doi : 10.1016/j.eswa.2020.113679 (cf. pages 3, 11).

[Eri+21] Gabriel Erion, Joseph D Janizek, Pascal Sturmfels, Scott M Lundberg et Su-In Lee. « Improving performance of deep learning models with axiomatic attribution priors and expected gradients ». In : *Nature machine intelligence* 3.7 (2021), p. 620-631 (cf. pages 31, 32, 36).

[Esp+21] Laura Espinosa, Ariana Wijermans, Francisco Orchard, Michael Höhle, Thomas Czernichow, Pietro Coletti, Lisa Hermans, Christel Faes, Esther Kissling et Thomas Mollet. « Epitweetr : Early warning of public health threats using Twitter data ». In : *MedRxiv* (2021) (cf. page 95).

[Fan+21] Feng-Lei Fan, Jinjun Xiong, Mengzhou Li et Ge Wang. « On interpretability of artificial neural networks : A survey ». In : *IEEE Transactions on Radiation and Plasma Medical Sciences* 5.6 (2021), p. 741-760 (cf. page 34).

[FM13] Peter M Fayers et David Machin. *Quality of life : the assessment, analysis and interpretation of patient-reported outcomes*. John Wiley & Sons, 2013 (cf. page 4).

[For+21] Ella M E Forgie, Hollis Lai, Bo Cao, Eleni Stroulia, Andrew J Greenshaw et Helly Goez. « Social Media and the Transformation of the Physician-Patient Relationship : Viewpoint ». In : *J Med Internet Res* 23.12 (déc. 2021), e25230. issn : 1438-8871. doi : 10.2196/25230. url : http://www.ncbi.nlm.nih.gov/pubmed/34951596 (cf. page 133).

[For87] Steven Fortune. « A sweepline algorithm for Voronoi diagrams ». In : *Algorithmica* 2.1 (1987), p. 153. doi : 10.1007/BF01840357 (cf. page 74).

[Fox11] Susannah Fox. « The Social Life of Health Information, 2011 ». In : (2011) (cf. page 94).

[Gar91] David G Garson. « Interpreting neural network connection weights ». In : (1991) (cf. page 30).

[GIG17] Yarin Gal, Riashat Islam et Zoubin Ghahramani. « Deep Bayesian Active Learning with Image Data ». In : *34th International Conference on Machine Learning (ICML)*. 2017, p. 1183-1192 (cf. page 86).

[Gil+18] Leilani H. Gilpin, David Bau, Ben Z. Yuan, Ayesha Bajwa, Michael Specter et Lalana Kagal. « Explaining Explanations : An Overview of Interpretability of Machine Learning ». In : *IEEE 5th International Conference on Data Science and Advanced Analytics (DSAA)*. 2018, p. 80-89. doi : 10.1109/DSAA.2018.00018 (cf. page 24).

[Gui+18a] Riccardo Guidotti, Anna Monreale, Salvatore Ruggieri, Dino Pedreschi, Franco Turini et Fosca Giannotti. « Local Rule-Based Explanations of Black Box Decision Systems ». In : *ArXiv* abs/1805.10820 (2018) (cf. pages 28, 29, 31, 36, 48).

[Gui+18b] Riccardo Guidotti, Anna Monreale, Salvatore Ruggieri, Franco Turini, Fosca Giannotti et Dino Pedreschi. « A survey of methods for explaining black box models ». In : *ACM computing surveys (CSUR)* 51.5 (2018), p. 1-42 (cf. pages 2, 34, 96).

[Hag20] Thilo HAGENDORFF. « The ethics of AI ethics : An evaluation of guidelines ». In : *Minds and Machines* 30.1 (2020), p. 99-120 (cf. pages 2, 96).

[Hao+19] Yaru HAO, Li DONG, Furu WEI et Ke XU. « Visualizing and understanding the effectiveness of BERT ». In : *arXiv preprint arXiv :1908.05620* (2019) (cf. pages 36, 45, 46).

[Har15] Adam W HARLEY. « An Interactive Node-Link Visualization of Convolutional Neural Networks ». In : *ISVC*. 2015, p. 867-877 (cf. pages 26, 36, 41).

[HB03] Mark HARROWER et Cynthia A. BREWER. « ColorBrewer.org : An Online Tool for Selecting Colour Schemes for Maps ». In : *The Cartographic Journal* 40.1 (2003), p. 27-37. DOI : 10.1179/000870403235002042 (cf. page 65).

[He+20] Yingzhe HE, Guozhu MENG, Kai CHEN, Xingbo HU et Jinwen HE. « Towards security threats of deep learning systems : A survey ». In : *IEEE Transactions on Software Engineering* (2020) (cf. page 2).

[Hei+12] Florian HEIMERL, Steffen KOCH, Harald BOSCH et Thomas ERTL. « Visual classifier training for text document retrieval ». In : *IEEE Transactions on Visualization and Computer Graphics* 18.12 (2012), p. 2839-2848. DOI : 10.1109/TVCG.2012.277 (cf. pages 84, 86).

[HG18] Florian HEIMERL et Michael GLEICHER. « Interactive analysis of word vector embeddings ». In : *Computer Graphics Forum*. T. 37. 3. Wiley Online Library. 2018, p. 253-265 (cf. pages 36, 40).

[HM16] Ruining HE et Julian MCAULEY. « Ups and Downs : Modeling the Visual Evolution of Fashion Trends with One-Class Collaborative Filtering ». In : *25th International Conference on World Wide Web (WWW)*. 2016, p. 507-517. DOI : 10.1145/2872427.2883037 (cf. pages 75, 77, 78).

[Hoc98] Sepp HOCHREITER. « The Vanishing Gradient Problem During Learning Recurrent Neural Nets and Problem Solutions ». In : *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems* 06.02 (1998), p. 107-116. DOI : 10.1142/S0218488598000094. eprint : https://doi.org/10.1142/S0218488598000094. URL : https://doi.org/10.1142/S0218488598000094 (cf. page 17).

[Hoh+19] Fred HOHMAN, Minsuk KAHNG, Robert PIENTA et Duen Horng CHAU. « Visual analytics in deep learning : An interrogative survey for the next frontiers ». In : *IEEE Transactions on Visualization and Computer Graphics* 25.8 (2019), p. 2674-2693. DOI : 10.1109/TVCG.2018.2843369 (cf. pages 3, 30, 35, 36, 46, 55).

[HR02] Geoffrey HINTON et Sam ROWEIS. « Stochastic Neighbor Embedding ». In : *Proceedings of the 15th International Conference on Neural Information Processing Systems*. NIPS'02. Cambridge, MA, USA : MIT Press, 2002, p. 857-864. URL : http://dl.acm.org/citation.cfm?id=2968618.2968725 (cf. pages 36, 37, 39).

[HR03] Geoffrey E HINTON et Sam ROWEIS. « Stochastic Neighbor Embedding ». In : *Advances in Neural Information Processing Systems (NIPS)*. T. 15. MIT Press, 2003, p. 857-864 (cf. page 74).

[HR18] Jeremy Howard et Sebastian Ruder. « Universal language model fine-tuning for text classification ». In : *arXiv preprint arXiv :1801.06146* (2018) (cf. page 20).

[HS97] Sepp Hochreiter et Jürgen Schmidhuber. « Long Short-Term Memory ». In : *Neural Comput.* 9.8 (nov. 1997), p. 1735-1780. issn : 0899-7667. doi : 10.1162/neco.1997.9.8.1735. url : https://doi.org/10.1162/neco.1997.9.8.1735 (cf. pages 16, 17, 41).

[HSG20] Benjamin Hoover, Hendrik Strobelt et Sebastian Gehrmann. « exBERT : A Visual Analysis Tool to Explore Learned Representations in Transformer Models ». In : *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics : System Demonstrations*. Online : Association for Computational Linguistics, juill. 2020, p. 187-196. doi : 10.18653/v1/2020.acl-demos.22. url : https://aclanthology.org/2020.acl-demos.22 (cf. pages 36, 45, 46, 84, 86).

[Hua+20] Xiaowei Huang, Daniel Kroening, Wenjie Ruan, James Sharp, Youcheng Sun, Emese Thamo, Min Wu et Xinping Yi. « A survey of safety and trustworthiness of deep neural networks : Verification, testing, adversarial attack and defence, and interpretability ». In : *Computer Science Review* 37 (2020), p. 100270 (cf. pages 2, 96).

[IQ20] Touseef Iqbal et Shaima Qureshi. « The survey : Text generation models in deep learning ». In : *Journal of King Saud University - Computer and Information Sciences* (2020). issn : 1319-1578. doi : https://doi.org/10.1016/j.jksuci.2020.04.001. url : https://www.sciencedirect.com/science/article/pii/S1319157820303360 (cf. pages 3, 11).

[Ish+21] Abid Ishaq, Muhammad Umer, Muhammad Faheem Mushtaq, Carlo Medaglia, Hafeez Ur Rehman Siddiqui, Arif Mehmood et Gyu Sang Choi. « Extensive hotel reviews classification using long short term memory ». In : *Journal of Ambient Intelligence and Humanized Computing* 12.10 (oct. 2021), p. 9375-9385. issn : 1868-5145. doi : 10.1007/s12652-020-02654-z. url : https://doi.org/10.1007/s12652-020-02654-z (cf. page 23).

[Iyy+15] Mohit Iyyer, Varun Manjunatha, Jordan Boyd-Graber et Hal Daumé III. « Deep Unordered Composition Rivals Syntactic Methods for Text Classification ». In : *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1 : Long Papers)*. Beijing, China : Association for Computational Linguistics, juill. 2015, p. 1681-1691. doi : 10.3115/v1/P15-1162. url : https://aclanthology.org/P15-1162 (cf. page 15).

[Jin+19] Xisen Jin, Zhongyu Wei, Junyi Du, Xiangyang Xue et Xiang Ren. « Towards hierarchical importance attribution : Explaining compositional semantics for neural sequence models ». In : *arXiv preprint arXiv :1911.06194* (2019) (cf. pages 31, 34, 36, 48).

[JSL21] Joseph D Janizek, Pascal Sturmfels et Su-In Lee. « Explaining explanations : Axiomatic feature interactions for deep networks ». In : *Journal of Machine Learning Research* 22.104 (2021), p. 1-54 (cf. pages 31, 32, 36).

[Kar+19] Shigeki KARITA, Nanxin CHEN, Tomoki HAYASHI, Takaaki HORI, Hirofumi INAGUMA, Ziyan JIANG, Masao SOMEKI, Nelson Enrique Yalta SOPLIN, Ryuichi YAMAMOTO, Xiaofei WANG, Shinji WATANABE, Takenori YOSHIMURA et Wangyou ZHANG. « A Comparative Study on Transformer vs RNN in Speech Applications ». In : *2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*. 2019, p. 449-456. DOI : 10.1109/ASRU46091.2019.9003750 (cf. page 42).

[KCA15] Akos KÁDÁR, Grzegorz CHRUPAŁA et Afra ALISHAHI. « Lingusitic Analysis of Multi-Modal Recurrent Neural Networks ». In : *Proceedings of the Fourth Workshop on Vision and Language*. Lisbon, Portugal : Association for Computational Linguistics, sept. 2015, p. 8-9. DOI : 10.18653/v1/W15-2804. URL : https://aclanthology.org/W15-2804 (cf. page 43).

[KCA17] Akos KÁDÁR, Grzegorz CHRUPAŁA et Afra ALISHAHI. « Representation of linguistic form and function in recurrent neural networks ». In : *Computational Linguistics* 43.4 (2017), p. 761-780 (cf. pages 36, 43).

[KGB14] Nal KALCHBRENNER, Edward GREFENSTETTE et Phil BLUNSOM. « A Convolutional Neural Network for Modelling Sentences ». In : *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1 : Long Papers)*. Baltimore, Maryland : Association for Computational Linguistics, juin 2014, p. 655-665. DOI : 10.3115/v1/P14-1062. URL : https://aclanthology.org/P14-1062 (cf. pages 15, 16).

[Kim+16] Yoon KIM, Yacine JERNITE, David SONTAG et Alexander M RUSH. « Character-aware neural language models ». In : *Thirtieth AAAI conference on artificial intelligence*. 2016 (cf. page 16).

[Kim14] Yoon KIM. « Convolutional Neural Networks for Sentence Classification ». In : *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Doha, Qatar : Association for Computational Linguistics, oct. 2014, p. 1746-1751. DOI : 10.3115/v1/D14-1181. URL : https://aclanthology.org/D14-1181 (cf. page 16).

[KJF15] Andrej KARPATHY, Justin JOHNSON et Li FEI-FEI. « Visualizing and understanding recurrent networks ». In : *arXiv preprint arXiv :1506.02078* (2015) (cf. pages 36, 42, 43).

[Kow+17] Kamran KOWSARI, Donald E. BROWN, Mojtaba HEIDARYSAFA, K. MEIMANDI, Matthew S. GERBER et Laura E. BARNES. « HDLTex : Hierarchical Deep Learning for Text Classification ». In : *2017 16th IEEE International Conference on Machine Learning and Applications (ICMLA)* (2017), p. 364-371 (cf. page 23).

[Kow+19] Kamran KOWSARI, Kiana JAFARI MEIMANDI, Mojtaba HEIDARYSAFA, Sanjana MENDU, Laura BARNES et Donald BROWN. « Text classification algorithms : A survey ». In : *Information* 10.4 (2019), p. 150 (cf. pages 3, 11).

[Lan+19] Zhenzhong LAN, Mingda CHEN, Sebastian GOODMAN, Kevin GIMPEL, Piyush SHARMA et Radu SORICUT. « Albert : A lite bert for self-supervised learning of language representations ». In : *arXiv preprint arXiv :1909.11942* (2019) (cf. pages 23, 110).

[Las+13] Walter S. LASECKI, Young Chol SONG, Henry KAUTZ et Jeffrey P. BIGHAM. « Real-Time Crowd Labeling for Deployable Activity Recognition ». In :

CSCW '13. San Antonio, Texas, USA : Association for Computing Machinery, 2013, p. 1203-1212. ISBN : 9781450313315. DOI : 10.1145/2441776.2441912. URL : https://doi.org/10.1145/2441776.2441912 (cf. page 95).

[Lau+19] Thibault Laugel, Marie-Jeanne Lesot, Christophe Marsala, Xavier Renard et Marcin Detyniecki. « The Dangers of Post-hoc Interpretability : Unjustified Counterfactual Explanations ». In : *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI-19*. International Joint Conferences on Artificial Intelligence Organization, juill. 2019, p. 2801-2807. DOI : 10.24963/ijcai.2019/388. URL : https://doi.org/10.24963/ijcai.2019/388 (cf. pages 29, 87, 89, 131).

[LDG16] Tal Linzen, Emmanuel Dupoux et Yoav Goldberg. « Assessing the Ability of LSTMs to Learn Syntax-Sensitive Dependencies ». In : *Transactions of the Association for Computational Linguistics* 4 (déc. 2016), p. 521-535. ISSN : 2307-387X. DOI : 10.1162/tacl_a_00115. eprint : https://direct.mit.edu/tacl/article-pdf/doi/10.1162/tacl_a_00115/1567418/tacl_a_00115.pdf. URL : https://doi.org/10.1162/tacl%5C_a%5C_00115 (cf. pages 36, 43, 44).

[Lec+98] Y. Lecun, L. Bottou, Y. Bengio et P. Haffner. « Gradient-based learning applied to document recognition ». In : *Proceedings of the IEEE* 86.11 (1998), p. 2278-2324. DOI : 10.1109/5.726791 (cf. page 15).

[LEL18] Scott M Lundberg, Gabriel G Erion et Su-In Lee. « Consistent individualized feature attribution for tree ensembles ». In : *arXiv preprint arXiv :1802.03888* (2018) (cf. pages 31, 32, 36, 47).

[LeN19] Alexander LeNail. « NN-SVG : Publication-Ready Neural Network Architecture Schematics ». In : *Journal of Open Source Software* 4.33 (2019), p. 747. DOI : 10.21105/joss.00747. URL : https://doi.org/10.21105/joss.00747 (cf. pages 26, 36).

[Li+15] Jiwei Li, Xinlei Chen, Eduard Hovy et Dan Jurafsky. « Visualizing and understanding neural models in nlp ». In : *arXiv preprint arXiv :1506.01066* (2015) (cf. pages 30, 31, 36, 37, 43-45, 47).

[Li+18] Quan Li, Kristanto Sean Njotoprawiro, Hammad Haleem, Qiaoan Chen, Chris Yi et Xiaojuan Ma. « Embeddingvis : A visual analytics approach to comparative network embedding inspection ». In : *2018 IEEE Conference on Visual Analytics Science and Technology (VAST)*. IEEE. 2018, p. 48-59 (cf. pages 36, 39).

[Li+20] Qian Li, Hao Peng, Jianxin Li, Congying Xia, Renyu Yang, Lichao Sun, Philip S Yu et Lifang He. « A survey on text classification : From shallow to deep learning ». In : *arXiv preprint arXiv :2008.00364* (2020) (cf. pages 3, 11).

[Lin+17a] Zhouhan Lin, Minwei Feng, Cicero Nogueira dos Santos, Mo Yu, Bing Xiang, Bowen Zhou et Yoshua Bengio. « A structured self-attentive sentence embedding ». In : *International Conference on Learning Representations (ICLR)*. 2017 (cf. page 78).

[Lin+17b] Zhouhan Lin, Minwei Feng, Cicero Nogueira dos Santos, Mo Yu, Bing Xiang, Bowen Zhou et Yoshua Bengio. « A Structured Self-Attentive

Sentence Embedding ». In : *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*. OpenReview.net, 2017. URL : https://openreview.net/forum?id=BJC%5C_jUqxe (cf. page 22).

[Lip18] Zachary C. LIPTON. « The Mythos of Model Interpretability : In Machine Learning, the Concept of Interpretability is Both Important and Slippery. » In : *Queue* 16.3 (2018), p. 31-57. ISSN : 1542-7730. DOI : 10.1145/3236386.3241340 (cf. pages 23, 25, 27, 34-36, 55, 87).

[Liu+18a] Shusen LIU, Peer-Timo BREMER, Jayaraman J. THIAGARAJAN, Vivek SRIKUMAR, Bei WANG, Yarden LIVNAT et Valerio PASCUCCI. « Visual Exploration of Semantic Relationships in Neural Word Embeddings ». In : *IEEE Transactions on Visualization and Computer Graphics* 24.1 (2018), p. 553-562. DOI : 10.1109/TVCG.2017.2745141 (cf. pages 36, 39).

[Liu+18b] Shusen LIU, Zhimin LI, Tao LI, Vivek SRIKUMAR, Valerio PASCUCCI et Peer-Timo BREMER. « Nlize : A perturbation-driven visual interrogation tool for analyzing and interpreting natural language inference models ». In : *IEEE transactions on visualization and computer graphics* 25.1 (2018), p. 651-660 (cf. pages 36, 45).

[Liu+19a] Yang LIU, Eunice JUN, Qisheng LI et Jeffrey HEER. « Latent space cartography : Visual analysis of vector space embeddings ». In : *Computer Graphics Forum*. T. 38. 3. Wiley Online Library. 2019, p. 67-78 (cf. page 39).

[Liu+19b] Yinhan LIU, Myle OTT, Naman GOYAL, Jingfei DU, Mandar JOSHI, Danqi CHEN, Omer LEVY, Mike LEWIS, Luke ZETTLEMOYER et Veselin STOYANOV. « Roberta : A robustly optimized bert pretraining approach ». In : *arXiv preprint arXiv :1907.11692* (2019) (cf. page 23).

[LL17] Scott M LUNDBERG et Su-In LEE. « A unified approach to interpreting model predictions ». In : *Advances in neural information processing systems* 30 (2017) (cf. pages 31, 32, 34, 36, 47).

[LM14] Quoc LE et Tomas MIKOLOV. « Distributed Representations of Sentences and Documents ». In : *Proceedings of the 31st International Conference on Machine Learning*. Sous la dir. d'Eric P. XING et Tony JEBARA. T. 32. Proceedings of Machine Learning Research 2. Bejing, China : PMLR, juin 2014, p. 1188-1196. URL : https://proceedings.mlr.press/v32/le14.html (cf. page 15).

[LMJ16] Jiwei LI, Will MONROE et Dan JURAFSKY. « Understanding neural networks through representation erasure ». In : *arXiv preprint arXiv :1612.08220* (2016) (cf. pages 31, 33, 34, 36, 43, 44, 47).

[Lop+21] Milena LOPREITE, Pietro PANZARASA, Michelangelo PULIGA et Massimo RICCABONI. « Early warnings of COVID-19 outbreaks across Europe from social media ». In : *Scientific Reports* 11.1 (jan. 2021), p. 2147. ISSN : 2045-2322. DOI : 10.1038/s41598-021-81333-1. URL : https://doi.org/10.1038/s41598-021-81333-1 (cf. page 95).

[LSC22] Mingda LI, Jinhe SHI et Yi CHEN. « Identifying Influences in Patient Decision-making Processes in Online Health Communities : Data Science Approach ». In : *J Med Internet Res* 24.8 (août 2022), e30634. ISSN : 1438-

8871. DOI : 10.2196/30634. URL : http://www.ncbi.nlm.nih.gov/pubmed/36044266 (cf. page 133).

[LT19] Piyawat Lertvittayakumjorn et Francesca Toni. « Human-grounded evaluations of explanation methods for text classification ». In : *arXiv preprint arXiv :1908.11355* (2019) (cf. page 34).

[Mar+15] Martín Abadi, Ashish Agarwal, Paul Barham, Eugene Brevdo, Zhifeng Chen, Craig Citro, Greg S. Corrado, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Ian Goodfellow, Andrew Harp, Geoffrey Irving, Michael Isard, Yangqing Jia, Rafal Jozefowicz, Lukasz Kaiser, Manjunath Kudlur, Josh Levenberg, Dandelion Mané, Rajat Monga, Sherry Moore, Derek Murray, Chris Olah, Mike Schuster, Jonathon Shlens, Benoit Steiner, Ilya Sutskever, Kunal Talwar, Paul Tucker, Vincent Vanhoucke, Vijay Vasudevan, Fernanda Viégas, Oriol Vinyals, Pete Warden, Martin Wattenberg, Martin Wicke, Yuan Yu et Xiaoqiang Zheng. *TensorFlow : Large-Scale Machine Learning on Heterogeneous Systems*. Software available from tensorflow.org. 2015. URL : https://www.tensorflow.org/ (cf. pages 36, 39-43, 45, 49).

[Mat+21] Asha Mathew, Ardith Z Doorenbos, Min Kyeong Jang et Patricia E Hershberger. « Acceptance and commitment therapy in adult cancer survivors : a systematic review and conceptual model ». In : *Journal of Cancer Survivorship* 15.3 (2021), p. 427-451 (cf. page 118).

[Mat75] B.W. Matthews. « Comparison of the predicted and observed secondary structure of T4 phage lysozyme ». In : *Biochimica et Biophysica Acta - Protein Structure* 405.2 (1975), p. 442-451. ISSN : 0005-2795. DOI : 10.1016/0005-2795(75)90109-9 (cf. page 65).

[Mca99] Andrew J. Mcallister. *A New Heuristic Algorithm For The Linear Arrangement Problem*. Rapp. tech. Faculty of Computer Science, University of New Brunswick, 1999 (cf. page 68).

[Mel02] Ofer Melnik. « Decision Region Connectivity Analysis : A Method for Analyzing High-Dimensional Classifiers ». In : *Machine Learning* 48.1–3 (2002), p. 321-351. ISSN : 0885-6125. DOI : 10.1023/A:1013968124284 (cf. pages 36, 49, 84, 85).

[Mer+18] Yves Mercadier, Jérôme Azé, Sandra Bringay, Viviane Clavier, Erick Cuenca Pauta, Céline Paganelli, Pascal Poncelet et Arnaud Sallaberry. « # AIDS Analyse Information Dangers Sexualité : caractériser les discours à propos du VIH dans les forums de santé ». In : *IC : Ingénierie des Connaissances*. 2018, p. 71-86 (cf. pages 98, 133).

[MFC19] Marzieh Mozafari, Reza Farahbakhsh et Noel Crespi. « A BERT-based transfer learning approach for hate speech detection in online social media ». In : *International Conference on Complex Networks and Their Applications*. Springer. 2019, p. 928-940 (cf. page 20).

[MH08] L.J.P. van der Maaten et G.E. Hinton. « Visualizing High-Dimensional Data Using t-SNE ». In : (2008) (cf. pages 36, 37, 39, 74).

[MHM18] Leland McInnes, John Healy et James Melville. « Umap : Uniform manifold approximation and projection for dimension reduction ». In : *arXiv preprint arXiv :1802.03426* (2018) (cf. pages 31, 36, 37, 39, 60, 74).

[Mik+13a] Tomas MIKOLOV, Kai CHEN, Gregory S. CORRADO et Jeffrey DEAN. « Efficient Estimation of Word Representations in Vector Space ». In : *ICLR*. 2013 (cf. pages 12, 15, 28, 39).

[Mik+13b] Tomas MIKOLOV, Ilya SUTSKEVER, Kai CHEN, Greg S CORRADO et Jeff DEAN. « Distributed Representations of Words and Phrases and their Compositionality ». In : *Advances in Neural Information Processing Systems*. Sous la dir. de C. J. C. BURGES, L. BOTTOU, M. WELLING, Z. GHAHRAMANI et K. Q. WEINBERGER. T. 26. Curran Associates, Inc., 2013 (cf. pages 28, 29, 39, 40, 49).

[Mik+17] Tomas MIKOLOV, Edouard GRAVE, Piotr BOJANOWSKI, Christian PUHRSCH et Armand JOULIN. « Advances in pre-training distributed word representations ». In : *arXiv preprint arXiv :1712.09405* (2017) (cf. page 41).

[Mil95] Linda MILNE. « Feature selection using neural networks with contribution measures ». In : *AI-CONFERENCE-*. Citeseer. 1995, p. 571-571 (cf. pages 30, 36).

[Min+17] Yao MING, Shaozu CAO, Ruixiang ZHANG, Zhen LI, Yuanzhe CHEN, Yangqiu SONG et Huamin QU. « Understanding hidden memories of recurrent neural networks ». In : *2017 IEEE Conference on Visual Analytics Science and Technology (VAST)*. IEEE. 2017, p. 13-24 (cf. pages 36, 43, 44, 96, 128).

[Min+21] Shervin MINAEE, Nal KALCHBRENNER, E. CAMBRIA, Narjes NIKZAD, Meysam Asgari CHENAGHLU et Jianfeng GAO. « Deep Learning–based Text Classification ». In : *ACM Computing Surveys (CSUR)* 54 (2021), p. 1-40 (cf. pages 3, 11, 14, 22).

[Mit19] Brent MITTELSTADT. « Principles alone cannot guarantee ethical AI ». In : *Nature Machine Intelligence* 1.11 (2019), p. 501-507 (cf. pages 2, 96).

[MLY18] W James MURDOCH, Peter J LIU et Bin YU. « Beyond word importance : Contextual decomposition to extract interactions from lstms ». In : *arXiv preprint arXiv :1801.05453* (2018) (cf. pages 31, 33, 34, 36, 43, 44).

[MM21] Yuxin MA et Ross MACIEJEWSKI. « Visual Analysis of Class Separations With Locally Linear Segments ». In : *IEEE Transactions on Visualization and Computer Graphics* 27.1 (2021), p. 241-253. DOI : 10.1109/TVCG.2020.3011155 (cf. pages 36, 49, 84, 85).

[MS17] W James MURDOCH et Arthur SZLAM. « Automatic rule extraction from long short term memory networks ». In : *arXiv preprint arXiv :1702.02540* (2017) (cf. pages 31, 33, 36, 43, 44).

[Mun14] Tamara MUNZNER. *Visualization analysis and design*. CRC press, 2014 (cf. page 35).

[MWV15] M. A. MIGUT, M. WORRING et C. J. VEENMAN. « Visualizing Multi-Dimensional Decision Boundaries in 2D ». In : *21st ACM International Conference on Knowledge Discovery & Data Mining (SIGKDD)* 29.1 (2015), p. 273-295. DOI : 10.1007/s10618-013-0342-x (cf. pages 36, 49, 74, 84, 85).

[OQW20] Ray OSHIKAWA, Jing QIAN et William Yang WANG. « A Survey on Natural Language Processing for Fake News Detection ». In : *LREC*. 2020 (cf. page 11).

[Ouy+15] Xi OUYANG, Pan ZHOU, Cheng Hua LI et Lijun LIU. « Sentiment Analysis Using Convolutional Neural Network ». In : *2015 IEEE International Confe-*

rence on Computer and Information Technology ; Ubiquitous Computing and Communications ; Dependable, Autonomic and Secure Computing ; Pervasive Intelligence and Computing. 2015, p. 2359-2364. DOI : 10.1109/CIT/IUCC/DASC/PICOM.2015.349 (cf. page 15).

[Par+19] Cheonbok Park, Inyoup Na, Yongjang Jo, Sungbok Shin, Jaehyo Yoo, Bum Chul Kwon, Jian Zhao, Hyungjong Noh, Yeonsoo Lee et Jaegul Choo. « Sanvis : Visual analytics for understanding self-attention networks ». In : *2019 IEEE Visualization Conference (VIS)*. IEEE. 2019, p. 146-150 (cf. pages 36, 45, 46).

[PC+14] Céline Paganelli, Viviane Clavier et al. « S'informer via des médias sociaux de santé : quelle place pour les experts ? » In : *Le Temps des médias* 2 (2014), p. 141-143 (cf. page 98).

[PE11] Vanaja Paul et Perumal Ekambaram. « Involvement of nitric oxide in learning & memory processes ». In : *The Indian journal of medical research* 133.5 (2011), p. 471 (cf. page 95).

[Pea00] Karl Pearson. « X. On the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling ». In : *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science* 50.302 (1900), p. 157-175. DOI : 10.1080/14786440009463897 (cf. pages 74, 77).

[Pea01] Karl Pearson F.R.S. « LIII. On lines and planes of closest fit to systems of points in space ». In : *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science* 2.11 (1901), p. 559-572. DOI : 10.1080/14786440109462720 (cf. pages 36, 37, 74).

[Pet+18a] Matthew E Peters, Mark Neumann, Luke Zettlemoyer et Wen-tau Yih. « Dissecting contextual word embeddings : Architecture and representation ». In : *arXiv preprint arXiv :1808.08949* (2018) (cf. pages 36, 40).

[Pet+18b] Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee et Luke Zettlemoyer. « Deep contextualized word representations ». In : *arXiv e-prints*, arXiv :1802.05365 (fév. 2018), arXiv :1802.05365. arXiv : 1802.05365 [cs.CL] (cf. pages 14, 20).

[PSM14] Jeffrey Pennington, Richard Socher et Christopher Manning. « GloVe : Global Vectors for Word Representation ». In : *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Doha, Qatar : Association for Computational Linguistics, oct. 2014, p. 1532-1543. DOI : 10.3115/v1/D14-1162. URL : https://aclanthology.org/D14-1162 (cf. page 12).

[QQH16] Peng Qian, Xipeng Qiu et Xuan-Jing Huang. « Analyzing linguistic knowledge in sequential model of sentence ». In : *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*. 2016, p. 826-835 (cf. pages 36, 43, 44).

[Rad+19] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever et al. « Language models are unsupervised multitask learners ». In : *OpenAI blog* 1.8 (2019), p. 9 (cf. pages 23, 80).

[Rav+16] Daniele Ravì, Charence Wong, Fani Deligianni, Melissa Berthelot, Javier Andreu-Perez, Benny Lo et Guang-Zhong Yang. « Deep learning for health informatics ». In : *IEEE journal of biomedical and health informatics* 21.1 (2016), p. 4-21 (cf. page 96).

[Rei+19] Emily Reif, Ann Yuan, Martin Wattenberg, Fernanda B Viegas, Andy Coenen, Adam Pearce et Been Kim. « Visualizing and measuring the geometry of BERT ». In : *Advances in Neural Information Processing Systems* 32 (2019) (cf. pages 36, 39).

[RHT08] Eduardo Rodriguez-Tello, Jin-Kao Hao et Jose Torres-Jimenez. « An effective two-stage simulated annealing algorithm for the minimum linear arrangement problem ». In : *Computers & Operations Research* 35.10 (2008), p. 3331-3346 (cf. pages 68, 82).

[RHW85] David E Rumelhart, Geoffrey E Hinton et Ronald J Williams. *Learning internal representations by error propagation*. Rapp. tech. California Univ San Diego La Jolla Inst for Cognitive Science, 1985 (cf. pages 14, 22, 25, 30).

[RHW86] David E. Rumelhart, Geoffrey E. Hinton et Ronald J. Williams. « Learning representations by back-propagating errors ». In : *Nature* 323.6088 (oct. 1986), p. 533-536. issn : 1476-4687. doi : 10.1038/323533a0. url : https://doi.org/10.1038/323533a0 (cf. page 16).

[Rin+21] Orlando Rincones, Sayeda Naher, Rebecca Mercieca-Bebber, Martin Stockler et al. « An updated systematic review of quantitative studies assessing anxiety, depression, fear of cancer recurrence or psychological distress in testicular cancer survivors ». In : *Cancer Management and Research* 13 (2021), p. 3803 (cf. page 118).

[RN18] Alec Radford et Karthik Narasimhan. « Improving Language Understanding by Generative Pre-Training ». In : 2018 (cf. pages 20, 22).

[Rod+19] Francisco C. M. Rodrigues, Mateus Espadoto, Roberto Hirata et Alexandru C. Telea. « Constructing and Visualizing High-Quality Classifier Decision Boundary Maps ». In : *Information* 10.9 (2019), p. 280. issn : 2078-2489. doi : 10.3390/info10090280 (cf. pages 36, 49, 50, 84, 85).

[Rom12] Hélène Romeyer. « La santé en ligne. Des enjeux au-delà de l'information ». In : *Communication. Information médias théories pratiques* 30.1 (2012) (cf. page 94).

[RR15] Kumar Ravi et Vadlamani Ravi. « A survey on opinion mining and sentiment analysis : Tasks, approaches and applications ». In : *Knowledge-Based Systems* 89 (2015), p. 14-46. issn : 0950-7051. doi : https://doi.org/10.1016/j.knosys.2015.06.015. url : https://www.sciencedirect.com/science/article/pii/S0950705115002336 (cf. page 11).

[RSG16] Marco Tulio Ribeiro, Sameer Singh et Carlos Guestrin. « "Why should I trust you ?" Explaining the predictions of any classifier ». In : *22nd ACM International Conference on Knowledge Discovery & Data Mining (SIGKDD)*. 2016, p. 1135-1144. doi : 10.1145/2939672.2939778 (cf. pages 27, 31, 34, 36, 47, 48, 56).

[RSG18] Marco Tulio Ribeiro, Sameer Singh et Carlos Guestrin. « Anchors : High-precision model-agnostic explanations ». In : *Proceedings of the AAAI conference on artificial intelligence*. T. 32. 1. 2018 (cf. pages 28, 29, 31, 36).

[Rud19] Cynthia Rudin. « Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead ». In : *Nature Machine Intelligence* 1.5 (2019), p. 206-215 (cf. page 26).

[RVM19] Karthikeyan Natesan Ramamurthy, Kush Varshney et Krishnan Mody. « Topological Data Analysis of Decision Boundaries with Application to Model Selection ». In : *36th International Conference on Machine Learning (ICML)*. T. 97. PMLR, 2019, p. 5351-5360 (cf. pages 36, 49, 84, 85).

[Sal+09] John M. Salsman, Suzanne C. Segerstrom, Emily H. Brechting, Charles R. Carlson et Michael A. Andrykowski. « Posttraumatic growth and PTSD symptomatology among colorectal cancer survivors : a 3-month longitudinal examination of cognitive processing ». In : *Psycho-Oncology* 18.1 (2009), p. 30-41. doi : https://doi.org/10.1002/pon.1367. eprint : https://onlinelibrary.wiley.com/doi/pdf/10.1002/pon.1367. url : https://onlinelibrary.wiley.com/doi/abs/10.1002/pon.1367 (cf. page 94).

[Sal+22] Abdul Salam, Alexander Woodman, Ashely Chu, Lamiaa H Al-Jamea, Mohammed Islam, Malek Sagher, Mohammed Sager et Mahmood Akhtar. « Effect of post-diagnosis exercise on depression symptoms, physical functioning and mortality in breast cancer survivors : A systematic review and meta-analysis of randomized control trials ». In : *Cancer Epidemiology* 77 (2022), p. 102111 (cf. page 118).

[San+19] Victor Sanh, Lysandre Debut, Julien Chaumond et Thomas Wolf. « DistilBERT, a distilled version of BERT : smaller, faster, cheaper and lighter ». In : *arXiv preprint arXiv :1910.01108* (2019) (cf. page 23).

[SBH21] Sandipan Sikdar, Parantapa Bhattacharya et Kieran Heese. « Integrated Directional Gradients : Feature Interaction Attribution for Neural NLP Models ». In : *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1 : Long Papers)*. 2021, p. 865-878 (cf. pages 31, 33, 34, 36).

[Set09] Burr Settles. *Active learning literature survey*. Rapp. tech. University of Wisconsin-Madison Department of Computer Sciences, 2009 (cf. pages 89, 123).

[Sev+21] Rita Sevastjanova, Aikaterini-Lida Kalouli, Christin Beck, Hanna Schäfer et Mennatallah El-Assady. « Explaining Contextualization in Language Models using Visual Analytics ». In : *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1 : Long Papers)*. 2021, p. 464-476 (cf. pages 36, 40, 45, 46, 128).

[SG10] Christin Seifert et Michael Granitzer. « User-based active learning ». In : *IEEE International Conference on Data Mining Workshops (ICDM)*. 2010, p. 418-425 (cf. pages 84, 86).

[SGK17] Avanti SHRIKUMAR, Peyton GREENSIDE et Anshul KUNDAJE. « Learning important features through propagating activation differences ». In : *International conference on machine learning*. PMLR. 2017, p. 3145-3153 (cf. pages 30, 31, 36).

[Sha53] Lloyd S SHAPLEY. *A value for n-person games, Contributions to the Theory of Games, 2, 307–317*. 1953 (cf. pages 30, 32).

[Shn03] Ben SHNEIDERMAN. « The eyes have it : A task by data type taxonomy for information visualizations ». In : *The craft of information visualization*. Elsevier, 2003, p. 364-371 (cf. page 128).

[ŠK10] Erik ŠTRUMBELJ et Igor KONONENKO. « An efficient explanation of individual classifications using game theory ». In : *The Journal of Machine Learning Research* 11 (2010), p. 1-18 (cf. pages 30-32, 36).

[ŠK14] Erik ŠTRUMBELJ et Igor KONONENKO. « Explaining prediction models and individual predictions with feature contributions ». In : *Knowledge and information systems* 41.3 (2014), p. 647-665 (cf. pages 30-32, 36).

[Ska+22] Ingjerd SKAFLE, Anders NORDAHL-HANSEN, Daniel S QUINTANA, Rolf WYNN et Elia GABARRON. « Misinformation About COVID-19 Vaccines on Social Media : Rapid Review ». In : *J Med Internet Res* 24.8 (août 2022), e37367. ISSN : 1438-8871. DOI : 10.2196/37367. URL : http://www.ncbi.nlm.nih.gov/pubmed/35816685 (cf. page 133).

[SKS12] Adam SADILEK, Henry KAUTZ et Vincent SILENZIO. « Modeling spread of disease from social interactions ». In : *Sixth International AAAI Conference on Weblogs and Social Media*. 2012 (cf. page 95).

[SM86] Gerard SALTON et Michael J MCGILL. « Introduction to modern information retrieval ». In : (1986) (cf. page 12).

[Smi+16] Daniel SMILKOV, Nikhil THORAT, Charles NICHOLSON, Emily REIF, Fernanda B VIÉGAS et Martin WATTENBERG. « Embedding projector : Interactive visualization and interpretation of embeddings ». In : *arXiv preprint arXiv :1611.05469* (2016) (cf. pages 37, 84).

[Smi+17] Daniel SMILKOV, Shan CARTER, D. SCULLEY, Fernanda B. VIÉGAS et Martin WATTENBERG. « Direct-Manipulation Visualization of Deep Networks ». In : *ArXiv* abs/1708.03788 (2017) (cf. pages 26, 36, 41, 42).

[SMY19] Chandan SINGH, W. James MURDOCH et Bin YU. « Hierarchical interpretations for neural network predictions ». In : *International Conference on Learning Representations*. 2019. URL : https://openreview.net/forum?id=SkEqro0ctQ (cf. pages 31, 33, 34, 36, 47).

[SO21] Waddah SAEED et Christian OMLIN. « Explainable AI (XAI) : A Systematic Meta-Survey of Current Challenges and Future Opportunities ». In : *arXiv preprint arXiv :2111.06420* (2021) (cf. page 34).

[SP97] Mike SCHUSTER et Kuldip PALIWAL. « Bidirectional recurrent neural networks ». In : *Signal Processing, IEEE Transactions on* 45 (déc. 1997), p. 2673-2681. DOI : 10.1109/78.650093 (cf. pages 16, 19).

[SS14] Carson SIEVERT et Kenneth SHIRLEY. « LDAvis : A method for visualizing and interpreting topics ». In : *ACL Workshop on Interactive Language Learning, Visualization, and Interfaces (ILLVI)*. 2014, p. 63-70. DOI : 10.3115/v1/W14-3110 (cf. page 72).

[SS18] Ozan Sener et Silvio Savarese. « Active Learning for Convolutional Neural Networks : A Core-Set Approach ». In : *International Conference on Learning Representations (ICLR)*. 2018 (cf. page 86).

[SSB17] Stanislau Semeniuta, Aliaksei Severyn et Erhardt Barth. « A Hybrid Convolutional Variational Autoencoder for Text Generation ». In : *ACL Conference on Empirical Methods in Natural Language Processing (EMNLP)*. 2017, p. 627-637. DOI : 10.18653/v1/D17-1066 (cf. page 89).

[Str+17] Hendrik Strobelt, Sebastian Gehrmann, Hanspeter Pfister et Alexander M Rush. « Lstmvis : A tool for visual analysis of hidden state dynamics in recurrent neural networks ». In : *IEEE transactions on visualization and computer graphics* 24.1 (2017), p. 667-676 (cf. pages 36, 43, 44, 84, 86, 128).

[Str+18] Hendrik Strobelt, Sebastian Gehrmann, Michael Behrisch, Adam Perer, Hanspeter Pfister et Alexander M Rush. « Seq2seq-vis : A visual debugging tool for sequence-to-sequence models ». In : *IEEE transactions on visualization and computer graphics* 25.1 (2018), p. 353-363 (cf. pages 36, 47).

[Str+19] Hendrik Strobelt, Sebastian Gehrmann, Michael Behrisch, Adam Perer, Hanspeter Pfister et Alexander M. Rush. « Seq2seq-Vis : A Visual Debugging Tool for Sequence-to-Sequence Models ». In : *IEEE Transactions on Visualization and Computer Graphics* 25.1 (2019), p. 353-363. DOI : 10.1109/TVCG.2018.2865044 (cf. pages 84, 86).

[STY17] Mukund Sundararajan, Ankur Taly et Qiqi Yan. « Axiomatic attribution for deep networks ». In : *International conference on machine learning*. PMLR. 2017, p. 3319-3328 (cf. pages 31, 32, 34, 36).

[SW65] S. S. Shapiro et M. B. Wilk. « An Analysis of Variance Test for Normality (Complete Samples) ». In : *Biometrika* 52.3/4 (1965), p. 591-611. DOI : 10.2307/2333709 (cf. page 66).

[Tan+19] Hui Fen Tan, Kuangyan Song, Madeilene Udell, Yiming Sun et Yujia Zhang. « Why should you trust my interpretation ? Understanding uncertainty in LIME predictions ». In : *ArXiv* abs/1904.12991 (2019) (cf. page 28).

[Tap20] Mike Tapi Nzali. « DEFT 2020 : détection de similarité entre phrases et extraction d'information ». In : *6e conférence conjointe Journées d'études sur la Parole (JEP, 33e édition), Traitement Automatique des Langues Naturelles (TALN, 27e édition), Rencontre des étudiants Chercheurs en Informatique pour le Traitement Automatique des Langues (RéCITAL, 22e édition). Atelier Défi Fouille de Textes*. ATALA ; AFCP. 2020, p. 91-96 (cf. pages 94, 95).

[Tor+20] Amirsina Torfi, Rouzbeh A Shirvani, Yaser Keneshloo, Nader Tavaf et Edward A Fox. « Natural language processing advancements by deep learning : A survey ». In : *arXiv preprint arXiv :2003.01200* (2020) (cf. pages 2, 6).

[Tur50] A. M. Turing. « I.—COMPUTING MACHINERY AND INTELLIGENCE ». In : *Mind* LIX.236 (oct. 1950), p. 433-460. ISSN : 0026-4423. DOI : 10.1093/mind/LIX.236.433. eprint : https://academic.oup.com/mind/article-pdf/LIX/236/433/30123314/lix-236-433.pdf. URL : https://doi.org/10.1093/mind/LIX.236.433 (cf. page 3).

[Ulr01] Brandes Ulrik. « A faster algorithm for betweenness centrality ». In : *The Journal of Mathematical Sociology* 25.2 (2001), p. 163-177. DOI : 10.1080/0022250X.2001.9990249 (cf. page 64).

[Vas+17] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser et Illia Polosukhin. « Attention is all you need ». In : *Advances in neural information processing systems*. 2017, p. 5998-6008 (cf. pages 20, 22).

[VB19] Jesse Vig et Yonatan Belinkov. « Analyzing the Structure of Attention in a Transformer Language Model ». In : *ACL Workshop BlackboxNLP : Analyzing and Interpreting Neural Networks for NLP*. Association for Computational Linguistics, 2019, p. 63-76. DOI : 10.18653/v1/W19-4808 (cf. pages 84, 86).

[Vig19] Jesse Vig. « A Multiscale Visualization of Attention in the Transformer Model ». In : *57th Annual Meeting of the Association for Computational Linguistics : System Demonstrations (ACL)*. Association for Computational Linguistics, 2019, p. 37-42. DOI : 10.18653/v1/P19-3007 (cf. pages 36, 45, 46, 84, 86).

[Voi+19] Elena Voita, David Talbot, Fedor Moiseev, Rico Sennrich et Ivan Titov. « Analyzing multi-head self-attention : Specialized heads do the heavy lifting, the rest can be pruned ». In : *arXiv preprint arXiv :1905.09418* (2019) (cf. pages 36, 45, 46).

[Wal+19] Emily Wall, Meeshu Agnihotri, Laura Matzen, Kristin Divis, Michael Haass, Alex Endert et John Stasko. « A Heuristic Approach to Value-Driven Evaluation of Visualizations ». In : *IEEE Transactions on Visualization and Computer Graphics* 25.1 (2019), p. 491-500. DOI : 10.1109/TVCG.2018.2865146 (cf. pages 75, 76).

[WCB15] Jason Weston, Sumit Chopra et Antoine Bordes. « Memory Networks ». In : *CoRR* abs/1410.3916 (2015) (cf. page 19).

[WTC21] Zijie J. Wang, Robert Turko et Duen Horng Chau. « Dodrio : Exploring Transformer Models with Interactive Visualization ». In : *59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing : System Demonstrations (ACL-IJCNLP)*. Association for Computational Linguistics, 2021, p. 132-141 (cf. pages 36, 45, 46, 84, 86).

[Wu+16] Yonghui Wu, Mike Schuster, Z. Chen, Quoc V. Le, Mohammad Norouzi, Wolfgang Macherey, Maxim Krikun, Yuan Cao, Qin Gao, Klaus Macherey, Jeff Klingner, Apurva Shah, Melvin Johnson, Xiaobing Liu, Lukasz Kaiser, Stephan Gouws, Yoshikiyo Kato, Taku Kudo, Hideto Kazawa, Keith Stevens, George Kurian, Nishant Patil, Wei Wang, Cliff Young, Jason R. Smith, Jason Riesa, Alex Rudnick, Oriol Vinyals, Gregory S. Corrado, Macduff Hughes et Jeffrey Dean. « Google's Neural Machine Translation System : Bridging the Gap between Human and Machine Translation ». In : *ArXiv* abs/1609.08144 (2016) (cf. pages 3, 11, 13).

[WV18] Bernhard Waltl et Roland Vogl. « Explainable artificial intelligence the new frontier in legal informatics ». In : *Jusletter IT* 4 (2018), p. 1-10 (cf. page 23).

[Xu+15] Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhudinov, Rich Zemel et Yoshua Bengio. « Show, attend and tell : Neural image caption generation with visual attention ». In : *International conference on machine learning*. PMLR. 2015, p. 2048-2057 (cf. pages 27, 36).

[Yan+19] Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Russ R Salakhutdinov et Quoc V Le. « Xlnet : Generalized autoregressive pretraining for language understanding ». In : *Advances in neural information processing systems* 32 (2019) (cf. page 23).

[YB18] Vikas Yadav et Steven Bethard. « A Survey on Recent Advances in Named Entity Recognition from Deep Learning models ». In : *COLING*. 2018 (cf. page 11).

[Yin+17] Wenpeng Yin, Katharina Kann, Mo Yu et Hinrich Schütze. « Comparative study of CNN and RNN for natural language processing ». In : *arXiv preprint arXiv :1702.01923* (2017) (cf. page 19).

[Yu+22] Wenhao Yu, Chenguang Zhu, Zaitang Li, Zhiting Hu, Qingyun Wang, Heng Ji et Meng Jiang. « A Survey of Knowledge-Enhanced Text Generation ». In : *ACM Comput. Surv.* (jan. 2022). issn : 0360-0300. doi : 10.1145/3512467. url : https://doi.org/10.1145/3512467 (cf. pages 3, 11).

[ZC08] Y. Zhiyong et Xu. Congfu. « Using decision boundary to analyze classifiers ». In : *3rd International Conference on Intelligent System and Knowledge Engineering (ISKE)*. T. 1. 2008, p. 302-307. doi : 10.1109/ISKE.2008.4730945 (cf. pages 36, 49, 84, 85).

[Zha+17] Zizhao Zhang, Yuanpu Xie, Fuyong Xing, Mason McGough et Lin Yang. « Mdnet : A semantically and visually interpretable medical image diagnosis network ». In : *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2017, p. 6428-6436 (cf. pages 27, 36).

[Zha+19] Jiawei Zhang, Yang Wang, Piero Molino, Lezhi Li et David S. Ebert. « Manifold : A Model-Agnostic Framework for Interpretation and Diagnosis of Machine Learning Models ». In : *IEEE Transactions on Visualization and Computer Graphics* 25.1 (2019), p. 364-373. doi : 10.1109/TVCG.2018.2864499 (cf. pages 36, 49, 84, 85).

[Zha+21] Yu Zhang, Peter Tiňo, Aleš Leonardis et Ke Tang. « A survey on neural network interpretability ». In : *IEEE Transactions on Emerging Topics in Computational Intelligence* (2021) (cf. pages 2, 96).

[Zho+15] Chunting Zhou, Chonglin Sun, Zhiyuan Liu et F. Lau. « A C-LSTM Neural Network for Text Classification ». In : *ArXiv* abs/1511.08630 (2015) (cf. page 23).

[ZZL15] Xiang Zhang, Junbo Zhao et Yann LeCun. « Character-level Convolutional Networks for Text Classification ». In : *Advances in Neural Information Processing Systems*. Sous la dir. de C. Cortes, N. Lawrence, D. Lee, M. Sugiyama et R. Garnett. T. 28. Curran Associates, Inc., 2015 (cf. page 16).

MATÉRIEL SUPPLÉMENTAIRE

Questionnaire EBBE-Text

Vous allez être amenés à utiliser l'outil d'aide à l'interprétabilité des prédictions produites dans le cas d'une classification dichotomique de données textuelles à l'aide d'EBBE-Text.

Quelques concepts

Tâche de classification relative aux sentiments dans les phrases.

Espace de représentation de dimension n : représentation d'une donnée à travers n valeurs. Réduction de dimension : processus qui permet de passer d'un espace de dimension n à un espace de dimension m avec n>m.

Classe / label : étiquette d'une donnée. Prédiction : prédiction de l'étiquette d'une donnée faite par un classifieur. Encoder : intégralité des couches du réseau de neurones précédant la dernière couche (de classification). Classifieur : algorithme ou fonction prédisant une ou des étiquettes (ici, fully-connected).

MCC : Coefficient de corrélation de Matthews (-1 corrélation négative parfaite, 0 aucune corrélation, 1 corrélation positive parfaite). Exemple si les résultats de classification sont à l'opposé des labels, la valeur sera égale à -1.

Exemple final : "Malgré l'accueil chaleureux de ce restaurant et le cadre idyllique j'en garde un souvenir très amer du fait de la qualité des plats."

Certaines tâches en classification nécessitent de pouvoir présenter les raisons de leurs prédictions : domaine médical etc... EBBE-Text aide à l'interprétabilité avec la volonté de visualiser la frontière de décision. Les techniques classiques de réduction de dimension ne prenant pas en compte la tâche associée, ces techniques ne nous permettent pas de comprendre et visualiser les performances du réseau. Notre méthode a pour objectif de visualiser cette frontière de décision pour comprendre la certitude avec laquelle le réseau classe les données.

N'hésitez pas à me contacter sur Discord, si quoi que ce soit n'est pas clair durant le questionnaire.

[Vidéo d'introduction]

Visualisation des différentes localités

Pour commencer cette question, vous devez ouvrir la page d'EBBE-Text sans cliquer nulle part. La question porte sur la visualisation présentée des différentes localités de l'espace de représentation des données. Une localité d'un espace de représentation est une zone de l'espace de représentation qui regroupe des données proches les unes des autres. Dans toutes les tâches de deep-learning, les données qui se ressemblent se regroupent, nous appelons, ici, ces groupes des localités.

Rappel : Le coefficient de corrélation de Matthews est abrégé par MCC.

[Vidéo de présentation des localités]

Le coefficient de corrélation de Matthews de la localité 392 est ... (Be.1, Be.3)

- indisponible dans cette visualisation.
- meilleur que le coefficient de corrélation de Matthews de l'intégralité des données.
- identique au coefficient de corrélation de Matthews de l'intégralité des données.
- moins bon que le coefficient de corrélation de Matthews de l'intégralité des données.

La localité 392 comporte... (Be.1, Be.2)

- plus de données positives.
- plus de données négatives.
- autant de données positives que négatives.
- des données dont on ne peut pas connaître le caractère positif ou négatif.

La localité 11 comporte... (Be.1, Be.2)

- plus de données classées positivement.
- plus de données classées négativement.
- autant de données classées positivement que négativement.
- des données dont on ne peut pas connaître la classification.

Visualisation de la frontière de décision

Pour commencer ces questions, vous devez cliquer sur le graphique représentant la localité N°75. Les questions portent sur la visualisation de la frontière de décision (panneau de gauche). La frontière est représentée par une ligne grise, les données les plus à gauche sont les données classées le plus négativement, les données le plus droite sont les données classées le plus positivement. La frontière représente le seuil auquel on passe d'une prédiction à une autre (de positif à négatif). Une donnée orange est négative dans le jeu de données d'entrée. Une donnée verte est une donnée positive dans le jeu de données d'entrée. Cliquer sur une donnée dans cette visualisation permet

de connaître la phrase qui lui est associée, celle-ci se trouvera en haut du panneau central.

Intéressez-vous seulement aux concepts présentés ci-dessus.

[Vidéo de présentation de la frontière de décision]

Une donnée verte est... (Be.2)

- une donnée classée positivement.
- une donnée classée négativement.
- une donnée positive.
- une donnée négative.

Une donnée orange à la droite de la frontière est... (Be.2)

- une donnée négative classée négativement.
- une donnée positive classée négativement.
- une donnée négative classée positivement.
- une donnée positive classée positivement.

Quelle donnée est classée la plus négativement ? (Be.2)

- "The book came fast, right at the time I needed it, and it was in great condition."
- "CD is alright, innovative use of outside instruments, but the vocals need some severe resuscitation."
- "We found it to be a sad display of bad jokes. Hubby didn't read it."
- "This was defective did not work at all.. do not waste your money on this..dena"

Quelle donnée positive est classée la plus positivement ? (Be.2)

- "Sailed better than I thought it would, fast to trim and tack.Two are fun to race."
- "This book has not been released yet-not for several months! (G)."
- "The book came fast, right at the time I needed it, and it was in great condition."
- "Not sold in the Part of Japan I'm at. Thank you Amazon and all who supports it."

Quelle donnée positive est classée la plus négativement ? (Be.2)

- "This book has not been released yet-not for several months! (G)."
- "It's been simple to use and it's small enough to hide away on the desk."
- "I decree that the distribution of all music documentaries shall cease until this is realeased on DVD."
- "The book came fast, right at the time I needed it, and it was in great condition."

Visualisation du voisinage d'une donnée

Pour commencer ces questions, vous devez cliquer (sélectionner), dans la visualisation représentant la frontière de décision, sur la donnée la plus à droite. Cette visualisation est zoomable via un scroll. Un chemin vers la frontière de décision permet de voir quelles sont les modifications dans une phrase qui amènent la classification a être moins certaine. Les lignes noires représentent les liens entre les voisins directs d'une donnée et cette donnée. Les lignes bleues représentent les liens constituant les chemins d'une donnée à la frontière de décision. Ces mêmes lignes bleues sont représentées en rose lorsque ce lien traverse la frontière de décision. Ces liens existent entre les données entre elles et entre les données et les données (synthétiques) de la frontière.

Intéressez-vous seulement aux concepts présentés ci-dessus.

[Vidéo de présentation du voisinage d'une donnée]

Combien de voisins directs la donnée sélectionnée a-t-elle (données de frontière comprises) ? (Be.2, Be.3)

- Un.
- Deux.
- Trois.
- Dix-sept.

Combien de voisins directs sur la frontière la donnée sélectionnée a-t-elle ? (Be.2, Be.3)

- Aucun.
- Un.
- Deux.
- Dix-huit.

Visualisation des chemins jusqu'à la frontière de décision

Pour commencer ces questions, vous devez cliquer (sélectionner) la donnée la plus à gauche dans la visualisation de la frontière de décision. Cette visualisation est zoomable via un scroll. Une arborescence des phrases composant les chemins de la donnée à la frontière de décision est proposée sur le panneau central. Un survol du descripteur (carré à gauche d'une phrase) d'une phrase permet de localiser la donnée liée à cette phrase dans la visualisation de la frontière de décision. Le descripteur de la phrase (ici en gras) de la donnée sélectionnée représente sur sa partie haute le nombre de chemins jusqu'à la frontière de décision, et sur sa partie basse-droite le nombre de voisins directs de cette donnée plus éloignés de la frontière de décision. Le descripteur d'une phrase appartenant aux chemins d'une donnée jusqu'à la frontière de décision représente en haut la classe et la prédiction de la donnée. La barre noire centrale représente la frontière. La position et la couleur du point représente respectivement sa

prédiction et sa classe. La partie basse-droite représente le nombre de voisins directs de cette donnée plus proches de la frontière de décision.

Intéressez-vous seulement aux concepts présentés ci-dessus.

[Vidéo de présentation des chemins d'une donnée à la frontière de décision]

Figure 3 – *Visualisation des descripteurs*

Descriptions des descripteurs Le descripteur de gauche est un descripteur de données plus éloigné de la frontière de décision que celui qui a été sélectionné. Le descripteur suivant est un un descripteur de la donnée sélectionnée. Les cases bleues montrent la classe de la donnée (négative ou positive). La position du point informe l'utilisateur sur la classification (à gauche comme négative et à droite comme positive). La case rose (uniquement présente sur le descripteur de la donnée sélectionnée) informe l'utilisateur sur le nombre de chemins entre cette donnée et la frontière de décision. La case jaune indique dans quelle mesure le classifieur est incertain de sa prédiction (voir rubrique suivante). Les cases dorées (uniquement présentes sur le descripteur de données sélectionné) indiquent le nombre de voisins plus éloignés de la frontière de décision puis les données sélectionnées (× signifie aucun). Les cases vertes (uniquement présentes pour les voisins plus proches de la frontière de précision que la donnée sélectionnée) indiquent le nombre de voisins plus proches de la frontière de décision (× signifie aucun).

En observant tous les chemins jusqu'à la frontière de décision, quelle phrase se trouve sur l'un de ces chemins ? (Be.3)

- "We found it to be a sad display of bad jokes. Hubby didn't read it."
- "This album is really stupid. Don't buy it. It's a waste of money."
- "Was unable to play the game for several days after purchase due to defective DRM."
- "The stones were not very clear and the cut was bad and rough. Very disappointed."

Pour commencer la question suivante, vous devez sélectionner la phrase "My dog can not open his mouth and get a drink this was a waste of my money." en double-cliquant sur son descripteur.

Combien de chemins mènent de la donnée sélectionnée à la frontière de décision ? (Be.3)

- Un.
- Deux.

- Trois.
- Quatre.

Combien de voisins directs plus proches de la frontière de décision la donnée sélectionnée a-t-elle ? (Be.3)

- Aucun.
- Un.
- Deux.
- Trois.

Visualisation des métriques de classification

Sur chacun des descripteurs de phrases, la partie basse-gauche présente l'incertitude liée à la prédiction associée à la phrase. Pour les faibles incertitudes un signe moins précède la valeur. Cela signifie que l'incertitude est d'ordre 10 puissance moins cette valeur. Pour les fortes incertitudes (précédées d'un point), la valeur est a prendre telle quelle (avec le point).

Intéressez-vous seulement aux concepts présentés ci-dessus.

[Présentation des métriques de classification]

FIGURE 4 – *Visualisation des descripteurs*

Descriptions des descripteurs La case jaune indique dans quelle mesure le réseau est incertain de sa prédiction. Une valeur précédée d'un symbole moins signifie qu'elle est d'ordre dix puissance moins la valeur. Un point signifie la valeur elle-même (par exemple, "-4" signifie une incertitude d'ordre 10e-4, ".4" signifie une incertitude proche de .4).

Parmi les propositions suivantes, quelle est la valeur la plus proche de l'incertitude liée de la prédiction associée à la phrase "There is about one sentence in this book that is new information to anyone. What a waste."? (Be.4)

- -2
- 0.2
- -3
- 0.01

Parmi les propositions suivantes, quelle est la valeur la plus proche de l'incertitude liée de la prédiction associée à la phrase "There are so many movies you should watch. Save you money and try THE others."? (Be.4)

- -2
- 0.2
- -3
- 0.01

Visualisation des mots les plus pertinents

Pour commencer ces questions, vous devez cliquer sur l'onglet "Top-10 Word" sur le panneau de droite.

Un tableau des dix mots les plus pertinents est alors affiché (du plus pertinent au moins pertinent). Ce tableau est trié en fonction de la valeur λ (que vous ne devez pas modifier pour l'instant). Ce tableau contient entre parenthèses à côté de chacun des mots le nombre de fois ou celui-ci est apparu dans la localité. Trois différentes options d'affichage existent pour ce tableau. Celle sous vos yeux (mode Cl) propose pour chaque mot la répartition des classes des phrases contenant ce mot dans l'intégralité des données (barre du haut) ou dans la localité (barre du bas). Cliquez maintenant sur le bouton "Cl" pour changer de mode. L'option d'affichage sous vos yeux (mode Lc) présente la même donnée sur la barre du bas, mais la barre du haut présente maintenant la répartition des prédictions faite par le réseau pour la localité. Cliquez maintenant sur le bouton "Lc" pour changer de mode. L'option d'affichage sous vos yeux (mode Pr) présente sur la barre du bas la répartition des prédictions faites par le réseau pour la localité. Celle du haut présente la même chose mais pour l'intégralité des données. Cliquez maintenant sur le bouton "Pr" pour revenir au mode initial. Enfin si les barres ne font pas la même largeur, cela représente a quel point un mot est plus représenté dans une des modalités (localité ou intégralité des données) que dans l'autre (e.g. une barre deux fois moins large indique qu'un mot a une probabilité d'apparition deux fois moins grande).

Intéressez-vous seulement aux concepts présentés ci-dessus.

[Vidéo de présentation des mots pertinents]

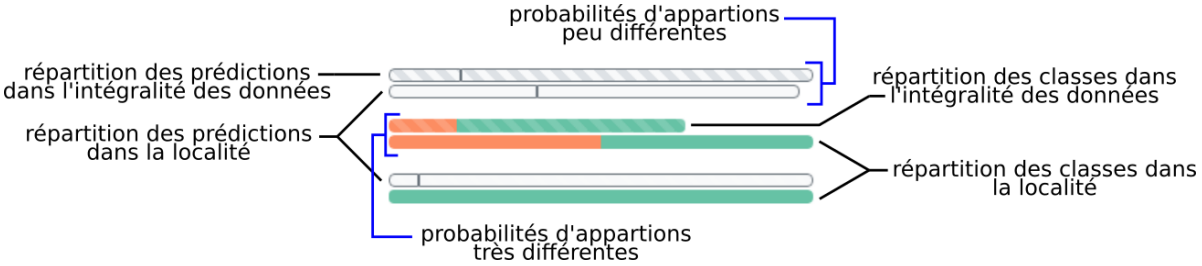


Figure 5 – *Mode Pr / Cl / Lo*

Quel est le mot le plus pertinent pour cette localité pour ce λ fixé? (Be.5)

- "just".
- "book".
- "exactly".
- "waste".

Définissez la valeur de λ à 1 et cliquez sur le bouton "Go".

Quel est le mot apparaissant le plus souvent dans les phrases de cette localité ? (Be.5)

- "book".
- "waste".
- "very".
- "good".

Le mot "good"... (Be.5)

- est apparu 48 fois dans les phrases contenues dans l'intégralité des données.
- est apparu 48 fois dans les phrases contenues dans la localité.
- a pour score moyen de prédiction des phrases qui le contiennent la valeur 48.
- a une probabilité d'apparition 48 fois plus élevée dans cette localité que dans l'intégralité des données.

Quel est le mot le plus représenté et exclusivement présent dans des phrases positives classées positivement dans cette localité ? (Be.5)

- "Condition".
- "Exactly".
- "Waste".
- "Come".

Le mot "condition" a... (Be.5)

- une tendance négative.
- aucune tendance particulière.
- une tendance positive.
- une tendance différente dans cette localité et dans l'intégralité des données.

Visualisation de l'espace de représentation en dimension réduite

Pour commencer ces questions, vous devez cliquer sur l'onglet "UMAP" sur le panneau de droite puis sélectionner la donnée la plus à droite dans la visualisation de la frontière de décision. UMAP et t-SNE sont des techniques de réduction de dimension

similaires à l'ACP (à savoir, elles produisent une réduction de dimension de l'espace de représentation).

Intéressez-vous seulement aux concepts présentés ci-dessus.

[Vidéo de présentation des espaces de dimension réduite]

A l'aide du survol du descripteur des phrases, comment caractérisez-vous la proximité des phrases contenues dans les chemins dans l'espace de représentation construit par UMAP ? (Be.6)

- Elles sont toutes proches.
- Elles sont proches sauf une ou deux qui sont plus éloignées.
- Elles sont toutes éloignées les unes des autres.
- Deux groupes de phrases proches, ces groupes sont loin l'un de l'autre.

Pour commencer cette question, vous devez cliquer sur l'onglet "t-SNE" sur le panneau de droite.

A l'aide du survol du descripteur des phrases, comment caractérisez-vous la proximité des phrases contenues dans les chemins dans l'espace de représentation construit par t-SNE ? (Be.6)

- Elles sont toutes proches.
- Elles sont proches sauf une ou deux qui sont plus éloignées.
- Elles sont toutes éloignées les unes des autres.
- Deux groupes de phrases proches, ces groupes sont loin l'un de l'autre.

Classification de phrase saisie

L'outil EBBE-Text permet une classification d'une phrase donnée en entrée. Pour commencer cette question, saisissez une phrase en anglais de moins de vingt mots dans la zone de saisie en bas du panneau de droite. Vous pouvez cliquer sur une phrase de la liste des phrases et la modifier légèrement si vous préférez. Cliquez ensuite sur "Classify". La ligne jaune qui apparaît symbolise la distance à la frontière de décision à laquelle votre phrase se trouverait.

Intéressez-vous seulement aux concepts présentés ci-dessus.

[Vidéo de présentation de la classification]

Comment est classée votre phrase par le classifieur ? (Be.7)

- Très négativement.
- Légèrement positivement.
- Moyennement positivement.
- Très positivement.