UNIVERSITÉ DU QUÉBEC À MONTRÉAL

DÉTECTION DE SIGNES ASSOCIÉS AUX TROUBLES DE LA SANTÉ MENTALE PAR ANALYSE AUTOMATIQUE DU LANGAGE NATUREL

MÉMOIRE

PRÉSENTÉ

COMME EXIGENCE PARTIELLE

DE LA MAÎTRISE EN INFORMATIQUE

PAR

YVES FERSTLER

MAI 2026

REMERCIEMENTS

Malgré les difficultés rencontrées au cours de ce mémoire, je suis avant tout très heureux d'avoir pu mener ce travail entouré de personnes que j'ai appris à connaître et à apprécier, dans un domaine qui n'a cessé de me passionner.

Je tiens d'abord à remercier profondément Catherine Lavoie, qui est l'une des premières personnes à qui je pense lorsque je me remémore les différentes étapes de ce parcours. Son aide, sa présence et son soutien ont eu une grande importance pour moi. Je remercie également Thomas Soulas, Valentyna Tsilinchuk, Yassine Chahdi, ainsi que Gaëlle Laperrière et les autres membres de l'équipe qui m'ont accompagné au fil du temps. Leur aide, que ce soit pour la recherche, la relecture ou leurs conseils, a été vraiment important pour moi. Au-delà de ça, ils ont surtout été des personnes incroyables avec qui j'ai partagé de très bons moments.

J'adresse aussi mes remerciements aux membres du CIRST, que j'ai eu la chance de côtoyer quotidiennement. Leur présence et nos échanges ont contribué à créer un environnement stimulant pour mon travail.

Finalement, je tiens avant tout à remercier ma directrice de recherche, Marie-Jean Meurs, pour la confiance qu'elle m'a accordée tout au long de ce mémoire. Son accompagnement, son soutien et ses encouragements ont été essentiels dans la réalisation de ce mémoire.

TABLE DES MATIÈRES

TABLE DES FIGURES

LISTE DES TABLEAUX

ACRONYMES

UQAM Université du Québec à Montréal.

RPA réseau de propagation avant.

PPC perceptron à plusieurs couches.

GPT Generative Pre-Trained.

BERT Bidirectional Encoder Representations from Transformers.

MNLI Multi-Genre Natural Language Inference.

TALN traitement automatique du langage naturel.

IA intelligence artificielle.

TA troubles alimentaires.

RÉSUMÉ

Les récentes avancées dans le domaine du traitement automatique du langage naturel ont permis d'améliorer les modèles en intelligence artificielle pour la compréhension de textes. Des tâches telles que la génération de textes, la détection d'entités nommées et le résumé automatique ont connu des progrès significatifs grâce à l'émergence de nouvelles méthodes. Cependant, certaines tâches, souvent associées à d'autres domaines, demandent une compréhension plus large que celle du simple texte. Par exemple, détecter des signes de troubles mentaux dans un texte nécessite une expertise dans le domaine de la santé mentale pour effectuer la tâche. L'objectif de ce manuscrit va alors être de d'examiner s'il est possible de détecter des comportements associés aux troubles de la santé mentale en utilisant les récentes approches issues du traitement automatique du langage.

Ce mémoire présente les notions préliminaires dans le domaine du traitement automatique du langage naturel qui ont permis la conception des nouvelles approches maintenant plus démocratisées dans de nombreux domaines. Le document met ensuite en évidence les avancées récentes appliquées au domaine de la santé mentale. Plus précisément, il examine les applications de ces méthodes dans la détection du risque en santé mentale et la reconnaissance des signes de troubles. La suite du mémoire présente deux articles qui traitent de cette question. Le premier se focalise sur la détection de comportements associés aux troubles alimentaires. L'article présente l'utilisation d'un modèle appliqué à des conversations en ligne pour entraîner un réseau de neurones à prédire des comportements. Il met en évidence les décisions prises en matière de représentation des données et les méthodes qui ont contribué à l'amélioration des résultats obtenus par le modèle. Le deuxième article porte sur la sélection de phrases associées aux symptômes de la dépression. Dans ce travail, l'objectif était de trouver des phrases qui facilitent la détection d'un symptôme associé à la dépression, mais aussi de comparer deux approches : l'utilisation de modèles de langage pré-entraînés et affinés et l'approche zéro-coup. L'article montre l'intérêt de cette méthode pour cette tâche.

Mots-clés : Modèle de langue, modèle de sujets, représentation d'historique conversationnel, classification zéro-coup, extraction de phrases, santé mentale, dépression, troubles alimentaires.

INTRODUCTION

Selon l'Organisation mondiale de la Santé (OMS, 2022), une personne sur huit dans le monde vit avec des troubles de santé mentale. McGrath *et al.* (2023) estiment même qu'à l'âge de 75 ans, une personne sur deux aura expérimenté une ou plusieurs formes de ces troubles. Au Canada, c'est approximativement 18% de la population âgée de 15 ans ou plus qui est concernée, et ce nombre semble augmenter (Stephenson, 2023; STATCAN, 2023). Les troubles de la santé mentale prennent différentes formes, comme la dépression ou les troubles alimentaires. Ces troubles provoquent des perturbations importantes de la pensée, de la régulation des émotions ou du comportement, entraînant un impact dans de nombreux aspects de la vie personnelle ou professionnelle. Les personnes souffrant de troubles sévères de santé mentale ont une espérance de vie réduite de 10 à 20 ans et sont beaucoup plus sujettes au suicide (OMS, 2019; NIH, 2022; STATCAN, 2019). Ces conséquences ont incontestablement un impact majeur sur le bien-être des personnes, leur participation au monde du travail et leur productivité, et donc sur l'économie mondiale (Knapp et Wong, 2020; Mental Health Commission of Canada, 2012).

Aujourd'hui, des solutions existent pour prévenir ces troubles (au moins en partie) et prendre en charge les personnes qui en souffrent, comme le suivi par des spécialistes ou des traitements personnalisés. Cependant, les personnes souffrant de troubles ont souvent de la difficulté à bénéficier des aides mises à leur disposition. Des facteurs sociaux peuvent empêcher certaines personnes d'accéder à de l'aide en santé mentale (CIHI, 2024; STATCAN, 2024; OMS, 2019). D'autres peuvent hésiter à en demander ou refuser en raison de la stigmatisation potentielle ou de fausses croyances, telles que le fait que leurs problèmes ne sont pas assez graves ou qu'ils disparaîtront d'eux-mêmes. (Doll *et al.*, 2021; Henderson *et al.*, 2013; Negash *et al.*, 2020). Ces barrières expliquent en partie pourquoi certaines personnes se tournent vers des solutions alternatives, plus accessibles, pour les aider à trouver du soutien ou améliorer leur bien-être (Odgers et Jensen, 2020; Wang *et al.*, 2023; Aye *et al.*, 2024). Internet et des forums, comme Reddit [1], sont notamment utilisés pour poser des questions de façon (pseudo-)anonyme, sur une grande variété de sujets, et pour obtenir des réponses d'autres personnes utilisatrices. Les échanges amènent les personnes à décrire leur situation, leur état émotionnel et leurs symptômes, parfois de façon très détaillée. Les messages publiés sur ces forums contiennent donc souvent des informations révélatrices d'un mal-être ou de signes précoces de troubles mentaux.

1. `https://www.reddit.com/`

Grâce aux avancées récentes en intelligence artificielle (IA) et en traitement automatique du langage naturel (TALN), il est désormais possible d'effectuer une analyse automatique de texte et les approches par analyse de sentiments peuvent s'avérer très utiles pour détecter des signes de troubles dans des messages. En effet, depuis l'introduction de l'architecture Transformeur (Vaswani *et al.*, 2017), de nombreuses approches ont été développées pour améliorer la compréhension et l'interprétation automatiques de textes. Parmi elles, le modèle BERT (Devlin *et al.*, 2019a) a permis des progrès significatifs dans des tâches telles que la compréhension de texte et la détection de sujet, montrant également des performances notables en analyse de sentiments. D'autres approches, comme l'utilisation de modèles pré-entraînés tels que BART (Lewis *et al.*, 2020) en configuration zéro-coup (Yin *et al.*, 2019), offrent la possibilité de réaliser des classifications textuelles sans nécessiter un entraînement spécifique sur un ensemble de données annoté. Ces techniques ont déjà démontré leur efficacité dans divers domaines, y compris la santé mentale, où elles peuvent exploiter le contenu textuel des forums pour identifier des signaux précoces de détresse.

Dans ce contexte, la **question de recherche** explorée dans les travaux présentés est :

> Comment l'analyse de sentiments appliquée à des messages issus de forums en ligne peut-elle, lorsqu'elle est combinée à la détection de signaux précoces, être utilisée par des modèles d'apprentissage automatique afin de permettre la détection précoce de risques de troubles en santé mentale ?

Le manuscrit est structuré de la manière suivante. Le chapitre 1 présente les concepts clés nécessaires pour comprendre les chapitres suivants. Ce chapitre présente les concepts fondamentaux du domaine de l'intelligence artificielle, jusqu'aux méthodes les plus avancées de l'état de l'art. Le chapitre 2 décrit comment les méthodes du domaine de l'intelligence artificielle peuvent être appliquées dans la détection de signes associés aux troubles de la santé mentale.

Le développement de la recherche, présenté dans ce mémoire, a été publié sous forme de deux articles scientifiques revus par les pairs. Ces deux articles ont été présentés lors de la conférence sur le Traitement Automatique des Langues Naturelles (TALN 2025) à Marseille. Ils portent tous les deux sur la détection précoce de risques liés à la santé mentale à partir de données issues des réseaux sociaux.

Le chapitre 3 présente le premier article, qui porte sur la détection automatique des comportements associés aux troubles alimentaires. Ce chapitre revient sur la conception d'un outil de modélisation de sujets,

dans le but d'affiner la représentation des échanges entre personnes utilisatrices et ainsi d'extraire les comportements associés à ces troubles. Le chapitre 4 porte sur le trouble de la dépression. L'objectif de cet article est de comparer les méthodes traditionnelles de prédiction, qui sont souvent coûteuses en ressources de calcul et en données, à une méthode de classification zéro-coup qui n'a pas ces inconvénients. La fin de ce mémoire conclut sur une rétrospective des applications d'outils en IA pour la détection de troubles en santé mentale et de l'avenir que portent ces outils dans ce domaine.

Les articles présentés en tant que chapitres du manuscrit sont :

Détecter des comportements associés aux troubles alimentaires par l'analyse automatique des publications textuelles en ligne
Yves Ferstler, Catherine Lavoie, Marie-Jean Meurs
Association pour le Traitement Automatique des Langues
Actes de CORIA-TALN-RJCRI-RECITAL 2025. Actes des 32ème Conférence sur le Traitement Automatique des Langues Naturelles (TALN), volume 1 : articles scientifiques originaux
`https://aclanthology.org/2025.jeptalnrecital-taln.12/`

Sélection ordonnée de phrases associées aux symptômes de la dépression par classification zéro-coup
Yves Ferstler, Catherine Lavoie, Marie-Jean Meurs
Association pour le Traitement Automatique des Langues
Actes de CORIA-TALN-RJCRI-RECITAL 2025. Actes de l'atelier Traitement du langage médical à l'époque des LLMs 2025 (MLP-LLM)
`https://aclanthology.org/2025.jeptalnrecital-mlpllm.4/`

Au cours de ma maîtrise, j'ai aussi contribué aux travaux de eRisk 2024 et à l'article suivant qui les décrit :

Automatically finding evidence, predicting answers in mental health self-report questionnaires
Diego Maupomé, Yves Ferstler,Sébastien Mosser, Marie-Jean Meurs
eRisk 2024 : Early Risk Prediction on the Internet
eRisk 2024 Workshop at the 15th International Conference of the CLEF Association, CLEF 2024, Grenoble, France, September 9–12, 2024
`https://ceur-ws.org/Vol-3740/`

CHAPITRE 1
NOTIONS PRÉLIMINAIRES

1.1 Réseaux de neurones

Les réseaux de neurones font partie d'une famille d'algorithmes d'apprentissage automatique. Un réseau de neurones est constitué d'unités appelées *neurones*, chacun connecté à un ou plusieurs autres neurones. Les neurones sont généralement regroupés par couche, connectés d'une couche à l'autre. On parle de réseau de propagation avant (RPA) lorsque les connexions entre neurones sont unidirectionnelles. La forme la plus simple de ce type de réseau de neurones est le perceptron classique (Rosenblatt, 1958), qui se compose de seulement deux couches : une d'entrée et une de sortie. D'autres formes, comme le perceptron à plusieurs couches (PPC), ajoutent des couches additionnelles entre celle d'entrée et celle de sortie. Ces couches supplémentaires sont appelées des couches cachées et on parle de réseau de neurones profond lorsque le nombre de couches cachées est au minimum de deux. La Figure 1.1 montre un exemple d'un perceptron à plusieurs couches.

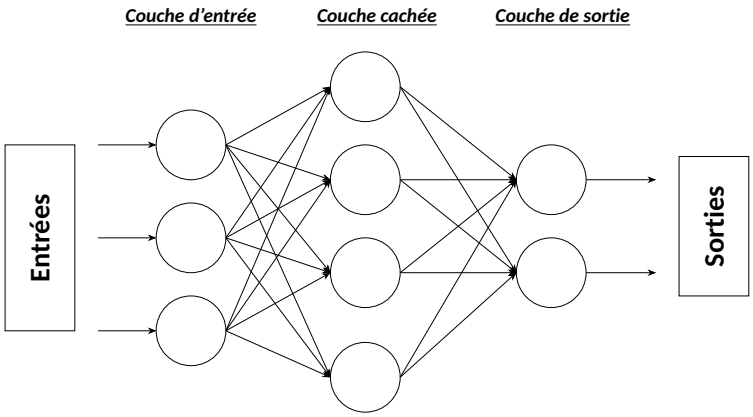


Figure 1.1 – Schéma d'un réseau de neurones perceptron à plusieurs couches et à propagation avant. Composé de trois couches : (1) une d'entrée, (2) une cachée, (3) une de sortie.

Les opérations entre neurones sont effectuées de la manière suivante : L'entrée d'un réseau de neurones correspond à une suite finie de valeurs $(u_i)_{i\in\mathbb{N}}$ appartenant à $\mathbb{R}$. Cette suite doit être de longueur $n_0 \in \mathbb{N}$, correspondant au nombre de neurones présents dans la couche d'entrée du réseau. La valeur d'un neurone est définie par y_{kj} avec $k \in \mathbb{N}$ l'index de sa couche dans le réseau de neurones, soit $k = 0$ pour la couche

d'entrée, et j l'index du neurone dans cette couche. On définit alors les valeurs des neurones d'entrée par :

$$y_{0,j} = u_i \quad \forall\, 1 \leq i, j \leq n_0 \quad \text{et} \quad i = j.$$

Pour les couches suivantes $k \geq 1$ de longueur $n_k \in \mathbb{N}$, la valeur d'un neurone y_{kj} est définie par la somme des valeurs des neurones de la couche précédente connectés à ce dernier, soit $(x_i)_{i\in\mathbb{N}} \in \mathbb{R}$ de longueur $m_{kj} \in \mathbb{N}$ correspondant au nombre de neurones connectés à y_{kj}. Chaque connexion est pondérée par un poids $w_i \in \mathbb{R}$. Un biais b_{kj} est ajouté à la suite de la somme pondérée. On définit alors les valeurs des neurones, se trouvant a posteriori de la couche d'entrée, par :

$$y_{k,j} = \sum_{i=1}^{m_{kj}} x_i w_i + b_{kj} \quad \forall\, 1 \leq i \leq m_{kj} \quad \text{et} \quad \forall\, 1 \leq j \leq n_k \quad \text{avec} \quad k \geq 1.$$

La Figure 1.2 illustre l'ensemble des opérations pour déterminer la valeur d'un neurone. Chaque neurone est défini par des opérations linéaires.Afin de permettre au réseau de neurones d'apprendre également des motifs non-linéaires, une fonction d'activation f est ajoutée aux sorties des neurones. Cette fonction permet au réseau de contrôler les sorties des neurones, notamment en permettant de désactiver certains d'entre eux. Désactiver un neurone consiste à réduire sa sortie à une valeur proche ou égale à zéro. Les fonctions d'activation les plus fréquentes sont les fonctions ReLU (Fred et Agarap, 2018) ou Sigmoïde (Han et Moraga, 1995), qui permettent notamment de désactiver les neurones lorsque sa valeur en sortie tend vers le négatif. La Figure 1.3 illustre les fonctions d'activation ReLU et Sigmoïde.

Selon l'approche la plus répandue, les valeurs des poids et des biais d'un réseau de neurones sont initialisées aléatoirement, et, lors de l'entraînement, ces valeurs sont mises à jour afin de permettre au réseau de prédire un résultat semblable à la vérité du terrain. Pour apprendre de ses erreurs, le réseau de neurones va calculer la perte, c'est-à-dire la différence entre le résultat obtenu et celui souhaité. En fonction de l'importance de la perte, le réseau de neurones ajuste les poids et les biais en utilisant la descente de gradient dans un processus de rétropropagation (Rumelhart *et al.*, 1986).

1.2 Algorithme de regroupement

Les algorithmes de regroupement ont pour objectif d'identifier des groupes de points dans un espace vectoriel en fonction de leurs similitudes, sans la nécessité de données étiquetées. Par exemple, si une personne reporte sur un graphique différentes caractéristiques de sifflements d'oiseaux, telles que le ton ou

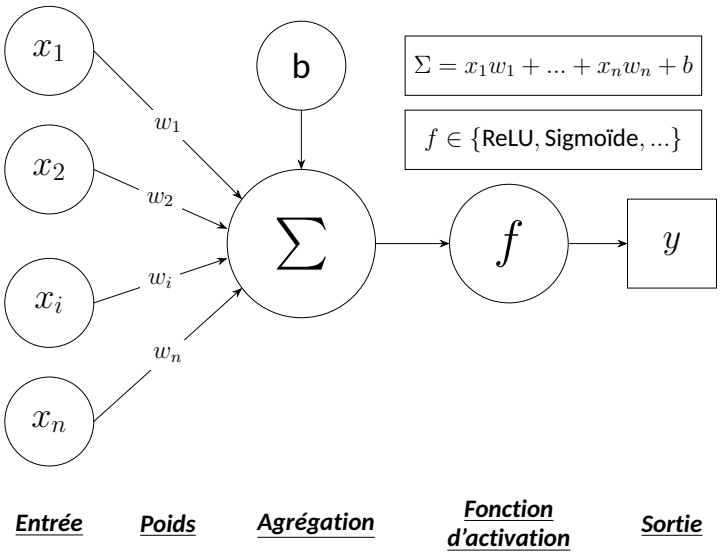


Figure 1.2 – Opérations entre valeurs d'entrée x, poids w, biais b et fonction d'activation $\int$ pour déterminer la valeur de neurone y

le nombre de répétitions, elle pourra naturellement observer des regroupements de points proches. Ces regroupements lui permettront de classer un sifflement particulier parmi l'un de ces groupes, sans pour autant connaître l'espèce de l'oiseau. Les algorithmes de regroupement fonctionnent selon un principe similaire : ils forment des groupes de points en se basant sur les distances ou similitudes entre ces derniers. Plusieurs méthodes existent pour réaliser ce type de regroupement, notamment K-moyennes (MacQueen, 1967) ou DBSCAN (Ester *et al.*, 1996), présentés ci-après.

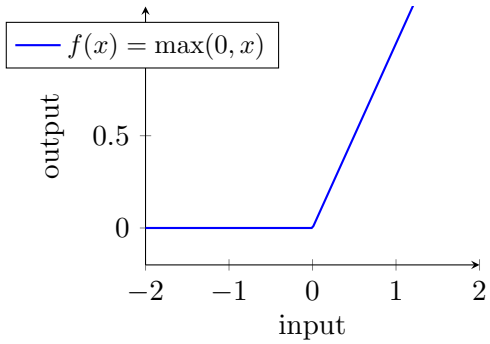
1.2.1 K-moyennes

L'algorithme K-moyennes (MacQueen, 1967) est utilisé pour identifier k groupes de points similaires dans un ensemble de données. Étant donné à l'instant t=1 un ensemble aléatoire de k centroïdes $m_1^{(1)}, ..., m_k^{(1)}$, l'approche standard (Forgy, 1965), alterne entre les deux étapes suivantes :

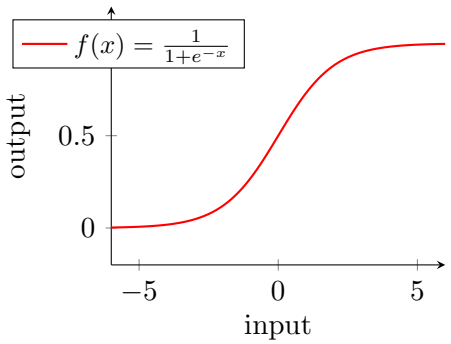
(1) Affecter chaque point au centroïde le plus proche en calculant la distance euclidienne au carré la plus courte.

$$S_i^{(t)} = \left\{ x_p : \left\| x_p - m_i^{(t)} \right\|^2 \leq \left\| x_p - m_j^{(t)} \right\|^2, \ \forall\, 1 \leq j \leq k \right\}, \ \forall\, 1 \leq i \leq k$$

où chaque point x_p est assigné à un seul groupe $S_i^{(t)}$.



(a) Fonction ReLU



(b) Fonction Sigmoïde

Figure 1.3 – Comparaison entre les fonctions d'activation ReLU (en bleu) et Sigmoïde (en rouge)

(2) Mettre à jour les coordonnées de chaque centroïde $m_i^{(t+1)}$ en calculant la moyenne des points affectés lors de la première étape.

$$m_i^{(t+1)} = \frac{1}{\left|S_i^{(t)}\right|} \sum_{x_p \in S_i^{(t)}} x_p$$

Les deux étapes sont répétées jusqu'à convergence, c'est-à-dire jusqu'à ce que les centroïdes ne changent plus ou que le nombre maximal d'itérations soit atteint.

L'algorithme K-moyennes est simple et rapide d'implémentation, mais il présente plusieurs limites. Premièrement, la mise à jour des centroïdes, telle que faite par K-moyennes, implique que l'algorithme tend à détecter des groupes de forme sphérique, ce qui limite sa capacité à identifier des groupes de formes plus complexes. Deuxièmement, l'algorithme nécessite de définir à l'avance la valeur de k, ce qui requiert un choix préalable du paramètre qui est parfois difficile à optimiser. Ainsi, selon le contexte, d'autres algorithmes comme DBSCAN sont préférables.

1.2.2 DBSCAN

DBSCAN (Ester *et al.*, 1996) est un algorithme de regroupement de densité qui identifie des groupes de points selon la densité de points dans l'espace. Les points proches les uns des autres seront regroupés, et ceux qui sont isolés seront considérés comme des valeurs aberrantes. DBSCAN a deux paramètres pour définir la densité de points dans un espace :

1. ε : rayon de voisinage. Les points se trouvant à une distance inférieure ou égale au rayon seront considérés comme voisins.

2. $\mathrm{MinPtsEps}$: nombre minimal de points devant se situer dans le rayon ε d'un point x_p pour former un groupe de densité.

DBSCAN peut être formulé à travers l'équation 1.1. Pour un ensemble de groupes possibles $\mathcal{C}$, l'objectif est de déterminer un sous-ensemble de groupes $C = \{C_1, ..., C_l\} \subset \mathcal{C}$ minimisant le nombre de groupes $|C|$ sous la contrainte que chaque paire de points $p, q \in C_i$ soient densément connectés sous un seuil ε :

$$\min_{d(p,q)\leq\varepsilon} |C| \quad \forall\, p, q \in C_i \quad \text{et} \quad \forall\, C_i \in C \tag{1.1}$$

$d(p, q)$ correspond à la plus petite valeur de ε pour que les points p et q soient densément connectés.

Le choix des paramètres est crucial lors de l'initialisation de l'algorithme, car il détermine la granularité des groupes à former. Cependant, avec des paramètres fixes pour définir la densité des groupes, cela peut poser un problème si toutes les zones de données ne présentent pas la même densité. Dans ce cas, DBSCAN risque de ne pas identifier correctement tous les groupes.

1.2.3 HDBSCAN

HDBSCAN (Campello *et al.*, 2013) est une extension de DBSCAN qui introduit une structure hiérarchique afin d'identifier des groupes de densité variable. L'algorithme commence par mesurer la densité locale autour de chaque point, puis redéfinit les distances entre les points en fonction de cette densité. Par exemple, deux points très proches dans une région peu dense sont considérés comme plus éloignés que deux points de même distance dans une région très dense. Après avoir ajusté ces distances, un arbre hiérarchique est construit à partir de l'arbre couvrant de poids minimal, permettant d'observer la formation et la disparition progressive des groupes. Enfin, il sélectionne les clusters les plus pertinents selon leur persistance dans la hiérarchie, tout en tenant compte du paramètre $\mathrm{MinPtsGrp}$, qui définit le nombre minimal de points qu'un groupe doit atteindre.

L'utilisation d'un algorithme basé sur la densité est généralement recommandée dans les espaces complexes, car il permet de segmenter la topologie des données, là où des algorithmes comme K-moyennes restent contraints à des formes de groupes plus sphériques. Un exemple de représentation des algorithmes de regroupement, pour quatre ensembles de données synthétiques (Lunes, Cercles, 3 Blobs et 3 Blobs à différentes densités) issus de Scikit-Learn [1], est présenté à la Figure 1.4.

1. `https://scikit-learn.org/stable/api/sklearn.datasets.html`

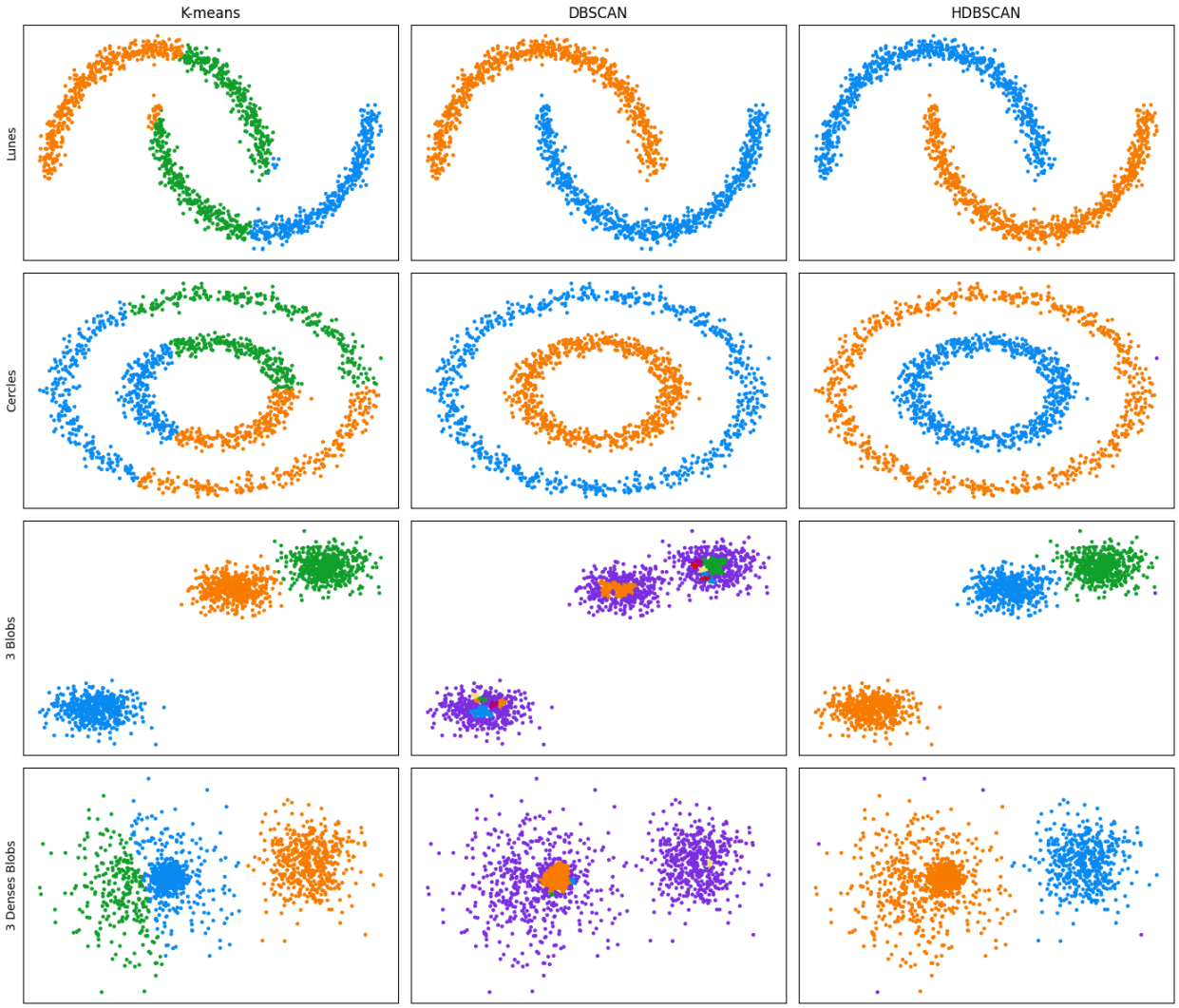


Figure 1.4 – Comparaison de trois algorithmes de regroupement : K-moyennes ($k = 3$), DBSCAN ($\varepsilon = 0.2$ et $\mathrm{MinPtsEps} = 10$) et HDBSCAN ($\mathrm{MinPtsGrp} = 30$). Quatre ensembles de données synthétiques issus de Scikit-Learn sont utilisés pour évaluer la capacité à identifier des groupes.

1.3 Manipulation de données textuelles

Nous venons de voir comment les réseaux de neurones et les algorithmes de regroupement fonctionnent, mais qu'en est-il de leur utilisation avec des données textuelles ?

Pour cela, le texte doit être représenté en valeurs numériques pour que les algorithmes puissent les utiliser comme entrée. Ces valeurs doivent être sélectionnées de manière à capturer le sens et les relations sémantiques entre les mots. Deux étapes permettent la représentation de texte en valeurs numériques : (1) La tokenisation, qui consiste à découper le texte en unités plus petites appelées tokens ; (2) La représentation vectorielle, également appelée plongement, capable de capturer le sens du mot ainsi que les relations syntaxiques et sémantiques avec les mots qui l'entourent.

1.3.1 Tokenisation

L'objectif des algorithmes de tokenisation est de découper chaque mot en token, et d'y associer un identifiant, appelé token ID. Chaque tokenisation permet d'obtenir un vocabulaire, qui est un dictionnaire indexant chaque token avec son identifiant. La taille du vocabulaire $\mathrm{taille}_{\text{vocabulaire}}$ sera alors égale au nombre total de tokens.

Une fois le vocabulaire créé, il peut être utilisé pour découper un nouveau corpus en tokens. Cependant, ce corpus peut présenter des tokens qui ne sont pas indexés dans le vocabulaire. Ainsi, un token spécial, inconnu (`[UNK]`), est ajouté au préalable au vocabulaire pour traiter tous les tokens n'ayant pas été indexés, mais ce token ne capturera aucune information sémantique. Plus il y aura de mots présents dans les données utilisées pour créer le vocabulaire, plus celui-ci sera riche. Pour garantir un vocabulaire suffisamment riche, les algorithmes de tokenisation proposent généralement des dictionnaires pré-établis.

Un vocabulaire est propre à son algorithme de tokenisation et à son ensemble de données. Les tokens peuvent être de différents types : des mots complets, des caractères ou des sous-groupes de mots.

1.3.1.1 Token de mots complets

La tokenisation par mots est l'approche la plus simple pour créer un vocabulaire. Elle consiste à découper un texte en une liste de mots, puis à associer un token à chaque mot distinct. Un exemple de tokenisation

par mots est présenté dans le tableau 1.1. Cependant, cette méthode présente deux inconvénients majeurs. Premièrement, pour couvrir l'ensemble des mots de la langue française, il faudrait indexer au moins 100 000 mots (Le Robert, 2025; Dictionnaire de l'Académie française, 2025; Larousse, 2025), sans compter ceux propres à un domaine particulier. Ensuite, pour chaque mot déjà indexé, il faudrait ajouter leurs formes fléchies, comme le pluriel, les conjugaisons ou les accords. Cet ajout augmente de manière importante le nombre déjà élevé de tokens dans le vocabulaire. Deuxièmement, même si ces formes sont incluses, chaque token sera traité indépendamment, malgré leurs similitudes. Cela rend plus difficile la suite du traitement, notamment lorsqu'on souhaite obtenir une représentation vectorielle.

Texte original	Ils	calculent	facilement	sans	calculatrice.	
Tokens	`Ils`	`calculent`	`facilement`	`sans`	`calculatrice`	`.`
Token ID	19	5183	2712	106	8025	99

Table 1.1 – Illustration de la représentation d'un texte original en tokens associée aux Token ID à partir d'un algorithme de tokenisation par mot.

1.3.1.2 Token de caractères

La tokenisation par caractères résout le problème précédent en créant un token pour chaque caractère d'un texte. Comme le nombre distinct de caractères est bien inférieur au nombre de mots dans un texte, le nombre de tokens inconnus sera par conséquent grandement réduit. La couverture du vocabulaire sera également plus facilement complète, permettant de mieux traiter de nouveaux mots. Un exemple de tokenisation par caractères est présenté au Tableau 1.2.

Cependant, séparer les mots en caractères augmente le nombre d'éléments à traiter dans un texte : un algorithme devra traiter une entrée pour chaque caractère, plutôt qu'un par mot. De plus, un caractère pour une langue latine contient peu d'information par rapport à un mot complet. Par exemple, le token `calculent` contient plus d'information sémantique que le token `c`. Une tokenisation par caractère rend donc la compréhension du texte plus difficile pour le modèle. En revanche, cette approche est adaptée à des langues comme le mandarin, où un caractère correspond souvent à un mot.

Texte original	ils calculent.												
Tokens	i	l	s	c	a	l	c	u	l	e	n	t	.
Token ID	3	4	1	7	0	4	7	8	4	5	9	2	6

Table 1.2 – Illustration de la représentation d'un texte original en tokens associée aux Token ID à partir d'un algorithme de tokenisation par caractères.

1.3.1.3 Token de sous-groupes de mots

La tokenisation de sous-groupes de mots consiste à décomposer les mots les moins fréquents en sous-groupes de mots plus fréquents. Chaque mot ou sous-groupe de mots est ajouté au vocabulaire en tant que token. Par exemple, si le mot "difficilement" est peu représenté dans le corpus, il sera décomposé en deux tokens plus courants, `difficile` et `#ment`. Par conséquent, si le mot "facilement" apparaît après la création du vocabulaire, mais que ce dernier a déjà rencontré "difficilement" et "facile", l'algorithme de tokenisation pourra tokeniser `facilement` en `facile` et `#ment`. Ce principe s'applique également à d'autres mots dérivés, composés ou au pluriel. Des techniques de tokenisation comme WordPiece (Wu *et al.*, 2016b) ou Byte Pair Encoding (BPE) (Sennrich *et al.*, 2016) reposent sur ce principe. Un exemple de tokenisation par sous-groupes de mots est présenté au Tableau 1.3.

Texte original	Ils	calculent	facilement	sans	calculatrice.			
Tokens	`Ils`	`calcul`	`#ent`	`facile`	`#ment`	`sans`	`calcul`	`#atrice`	`.`
Token ID	19	**1777**	2712	7712	106	8025	**1777**	1288	99

Table 1.3 – Illustration de la représentation d'un texte original en tokens associée aux Token ID à partir d'un algorithme de tokenisation par sous-groupes de mots.

1.3.2 Plongement des tokens

Après la création du vocabulaire, un corpus pourra être représenté par une suite d'identifiants numériques (Token ID), cependant, ces valeurs ont été initialisées sans tenir compte de leur sens dans la phrase. Deux tokens comme `voiture` et `véhicule` pourraient avoir deux identifiants éloignés alors que leur sens est similaire. Ce problème rend la tâche beaucoup plus compliquée pour l'apprentissage de réseaux de neurones ou d'algorithmes de regroupement. Afin d'éviter ce problème, des vecteurs, nommés plongements, capturent le sens d'un mot puis sont associés à chaque token présent dans le vocabulaire. Une illustration

des liens entre tokens, Token ID et plongements est présentée dans la Figure 1.5.

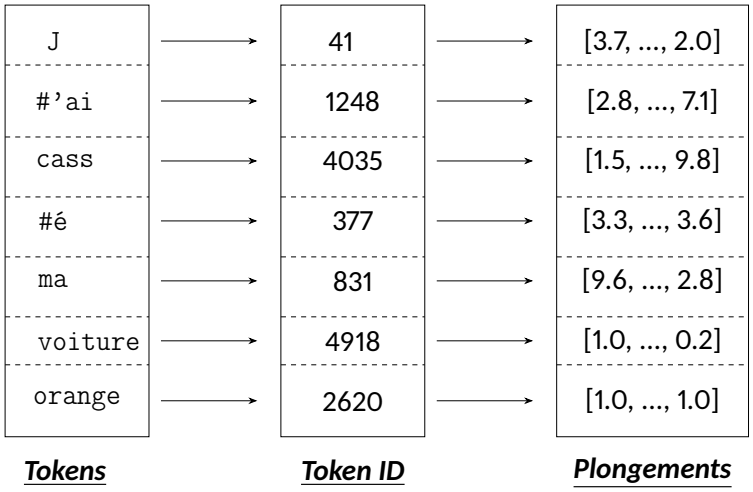


Figure 1.5 – Illustration entre une liste de tokens, les identifiants correspondant aux tokens (Token ID) et leurs plongements sous forme vectorielle.

On parle ainsi de **plongement de token** lorsqu'un token est représenté sous forme vectorielle. Ces vecteurs appartiennent à $\mathbb{R}^{d_{\text{modèle}}}$, avec $d_{\text{modèle}}$ correspondant à leur dimension, définie par l'architecture du modèle utilisé. Ils sont denses, numériques, et appartiennent à un espace de dimension fixe appelé **espace de plongement**. Pour chaque token présent dans le vocabulaire, un seul plongement sera associé. L'ensemble des plongements de token est contenu dans une **matrice de plongement** E de taille : $(\text{taille}_{\text{vocabulaire}}, d_{\text{modèle}})$, tel que :

$$e_i = E[i]$$

Le plongement e_i du token i est obtenu en accédant à la ligne i de la matrice E.

Si on prend les tokens `véhicule`, `voiture` et `orange` comme exemple, et une fonction hypothétique $\text{plongement}(token)$ qui prendrait en entrée un token et en sortirait un plongement de token, on aurait :

$$\begin{aligned} \text{plongement}(\texttt{véhicule}) &= [0,6\ ;\ 0,8\ ;\ 0,1] \\ \text{plongement}(\texttt{voiture}) &= [1\ ;\ 0,9\ ;\ 0,2] \\ \text{plongement}(\texttt{orange}) &= [1\ ;\ 0,2\ ;\ 1] \end{aligned} \tag{1.2}$$

Les vecteurs retournés dans l'équation 1.2 sont un exemple de plongement des tokens `véhicule`, `voiture` et `orange`. Une illustration de ces plongements de tokens dans un espace à trois dimensions est présenté à la Figure 1.6.

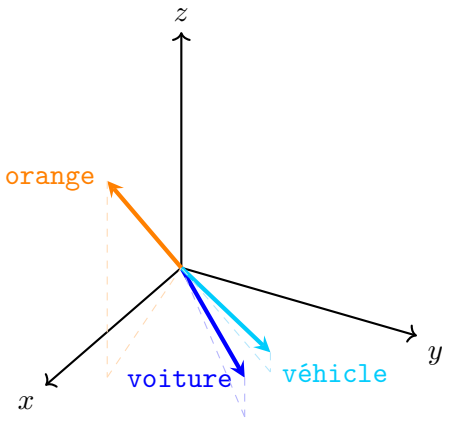


Figure 1.6 – Illustration de la représentation des tokens `orange`, `véhicule` et `voiture` dans un espace à 3-dimensions.

Cependant, si on suit l'exemple du token `orange`, le mot associé contient plusieurs sens (celui de la couleur ou du fruit, par exemple), mais uniquement un seul vecteur fixe (le plongement) est associé au token. Ainsi, comment permettre aux plongements de tokens de capturer différents sens selon leur contexte ?

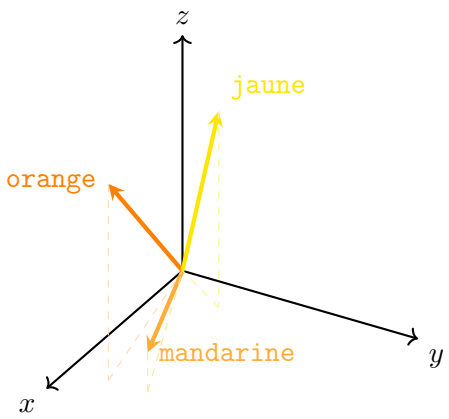
Pour cela, la représentation vectorielle des tokens doit pouvoir prendre en considération le contexte et les mots qui l'entourent. Parmi les différentes approches utilisées, l'architecture **transformeur** (Vaswani *et al.*, 2017) émerge comme la méthode la plus performante à ce jour.

1.4 Architecture transformeur

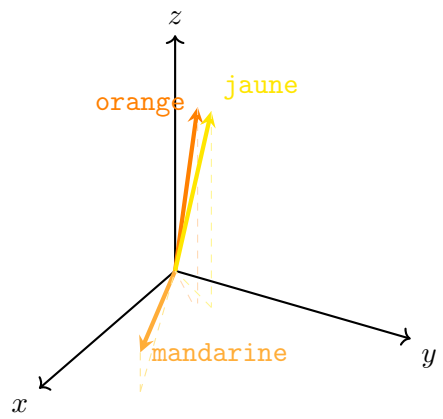
L'architecture transformeur est un système neuronale complexe du domaine de l'apprentissage profond. À l'aide d'un algorithme de tokenisation, elle convertit un texte en valeurs numériques (Token ID). Elle consulte ensuite la matrice de plongement, et au travers des différentes couches de l'architecture, affine le plongement en une représentation **contextuelle**. Ces plongements de tokens appartiennent à $\mathbb{R}^{d_{\text{modèle}}}$.

Le transformeur suit l'architecture encodeur-décodeur, proposée pour la première fois par Sutskever *et al.* (2014), puis régulièrement utilisée dans la tâche de traduction automatique. L'architecture est divisée en deux parties : l'une analyse et interprète les relations sémantiques entre tokens (encodeur), et l'autre génère des tokens pour former des phrases (décodeur). L'architecture encodeur-décodeur a mené à des résultats à l'état de l'art dans le domaine (Wu *et al.*, 2016a; Luong *et al.*, 2015) avant l'arrivée de l'architecture transformeur, basée principalement sur des *mécanismes d'attention* (Bahdanau *et al.*, 2014), notamment

l'auto-attention multi-têtes. Ces mécanismes permettent d'intégrer l'information contextuelle pertinente de chaque token à leur plongement. Si on reprend l'exemple de la Figure 1.5, cela permet de transformer le plongement initial du token `orange` en une représentation spécifique liée au contexte, ici la couleur plutôt que le fruit. Un exemple de cette transformation est présenté sur la Figure 1.7.



(a) Avant la propagation d'information du mécanisme d'attention : **plongement initial** du token `orange`.



(b) Après la propagation d'information du mécanisme d'attention : **représentation spécifique** du token `orange`.

Figure 1.7 – Illustration de la propagation d'information pour un token `orange` grâce à un mécanisme d'attention.

Par la suite, les architectures transformeur ont été utilisées dans une plus grande variété de tâches, que ce soit pour du texte, des images, de l'audio, etc. (Achiam *et al.*, 2023; Liu *et al.*, 2024; Radford *et al.*, 2023). D'autres types de modèles ont été développés à partir de l'architecture transformeur, tels que Bidirectional Encoder Representations from Transformers (BERT) (Devlin *et al.*, 2019a), qui reprend la partie encodeur de l'architecture pour créer des représentations textuelles denses, et Generative Pre-Trained (GPT) (Achiam *et al.*, 2023), qui reprend celle du décodeur pour la génération de texte. T5 (Raffel *et al.*, 2020) conserve l'architecture encodeur-décodeur et est entraîné pour des tâches plus diversifiées. La suite de ce manuscrit apporte des précisions sur la partie encodeur de l'architecture transformeur, illustrée par la Figure 1.8.

1.4.1 Entrée de l'architecture transformeur

La première étape du modèle consiste à appliquer un algorithme de tokenisation, associé à un vocabulaire, afin de convertir le contenu textuel en une séquence d'identifiants numériques. Le premier paramètre appris par le modèle correspond à la matrice de plongement, définie dans la section 1.3.2, qui permet de

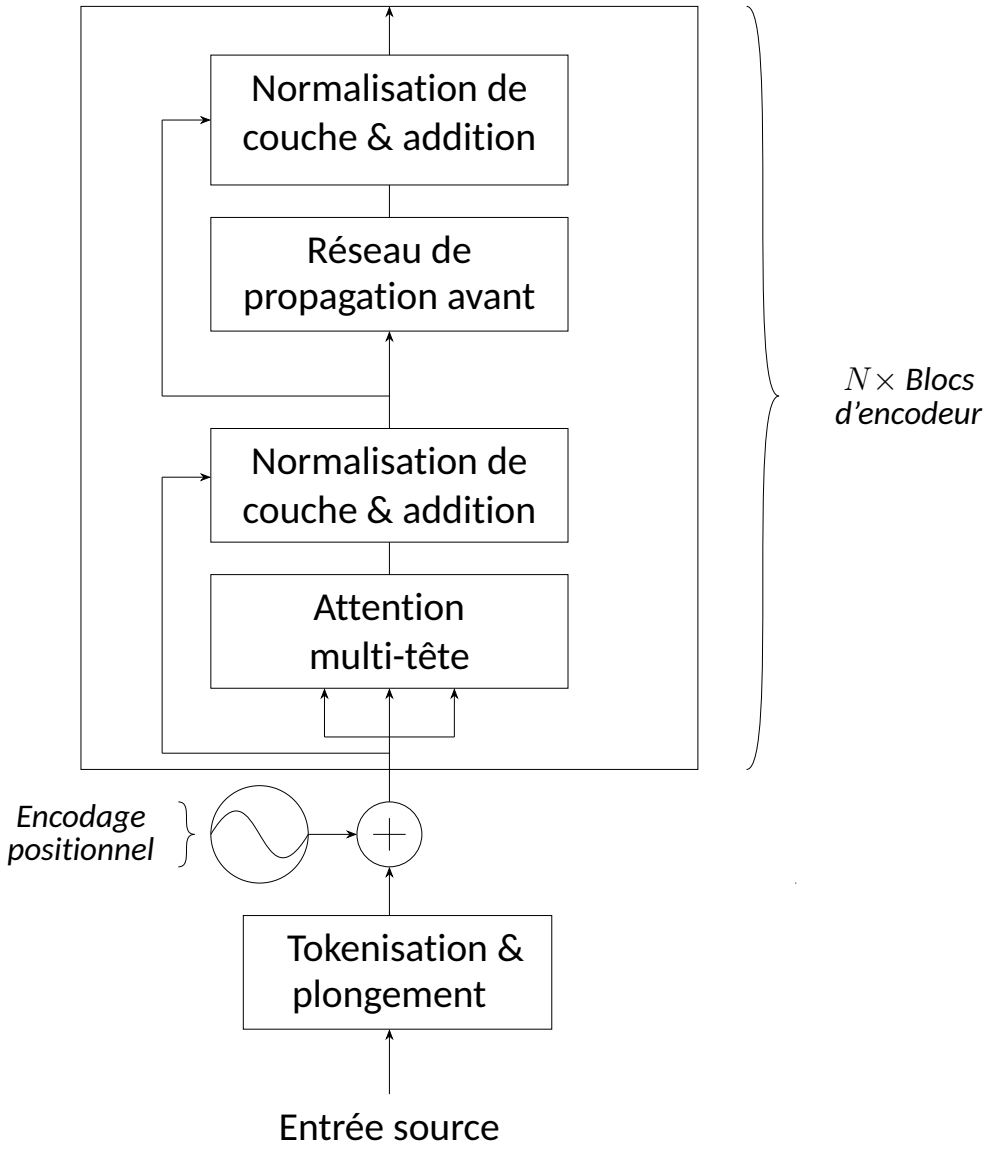


Figure 1.8 – Architecture de la partie encodeur d'une architecture transformeur

transformer les identifiants numériques en représentations vectorielles.

La première opération appliquée à ces représentations correspond à l'encodage positionnel, qui ajoute une information de position à chaque token, tel que :

$$x_i = e_i + p_i$$

Avec $p_i \in \mathbb{R}^{d_{\text{modèle}}}$ le vecteur d'encodage positionnel à la position i et x_i la première représentation du token i. Par exemple, la représentation du token `voiture` est différente entre les phrases “Ma chère voiture” et “Ma voiture chère”. Ainsi, cette étape permet d'apporter plus de précisions à la représentation initiale des

tokens en fonction de leur position dans la phrase.

Les représentations sont ensuite transmises à l'encodeur, chaque bloc d'encodage correspond à une séquence d'étapes permettant de propager l'information entre tokens. Une fois les étapes d'un bloc terminées, la sortie peut soit être réintroduite dans un autre bloc d'encodage, soit servir de résultat de l'encodeur. Selon le nombre de blocs choisis (N), l'encodeur affinera plus ou moins la représentation des tokens d'un point de vue contextuel et sémantique.

1.4.2 Blocs d'encodage

1.4.2.1 Attention Multi-Tête

La première étape d'encodage dans une architecture transformeur est l'*attention multi-tête*, celle qui consiste à appliquer le mécanisme d'attention entre les tokens. Le calcul de l'attention pour une séquence de tokens ne se fait pas une seule fois, mais plusieurs fois en parallèle, avec différents paramètres pour obtenir h résultats différents. L'intérêt d'avoir h différents calculs, appelés *têtes d'attention*, est de permettre au modèle d'identifier des relations entre les tokens.

En théorie, on pourrait imaginer qu'une tête d'attention cherche à trouver les liens entre les adjectifs, qu'une autre s'occupe du lien entre les noms propres, ou encore se charge de distinguer l'ironie (Manning *et al.*, 2020). Cependant, ces calculs sont beaucoup plus abstraits et les liens faits entre les tokens sont appris lors de l'entraînement du modèle.

Pour transformer une représentation grâce à une tête d'attention i, avec $i \in \mathbb{N}$ et $1 \leq i \leq h$, on la projette sur trois matrices de paramètres de dimension $d_h < d_{\text{modèle}}$: la matrice de requête ($W_i^Q \in \mathbb{R}^{d_{\text{modèle}} \times d_h}$), la matrice de clef ($W_i^K \in \mathbb{R}^{d_{\text{modèle}} \times d_h}$) et la matrice de valeur ($W_i^V \in \mathbb{R}^{d_{\text{modèle}} \times d_h}$).

La taille de la dimension réduite d_h dépend du nombre de têtes d'attention. Pour $h = 8$, la dimension sera calculée par : $d_h = d_{\text{modèle}}/h = 64$.

Pour comprendre comment le mécanisme d'attention fonctionne, nous allons prendre comme exemple la séquence de tokens suivante :

$$X = [\texttt{J}, \texttt{\#'ai}, \texttt{cass}, \texttt{\#é}, \texttt{ma}, \texttt{voiture}, \texttt{orange}]$$

Chaque token de X sera transformé en plongement de token, sans considération du contexte. Pour contextualiser ces plongements, les représentations de $X \in \mathbb{R}^{7 \times d_{\text{modèle}}}$ vont être multipliées par les matrices de paramètres pour donner : $Q = XW_0^Q$, $K = XW_0^K$ et $V = XW_0^V$, avec $Q, K, V \in \mathbb{R}^{7 \times d_h}$

Le produit matriciel QK^T permet d'obtenir une matrice de scores d'attention, appartenant ici à $\mathbb{R}^{7 \times 7}$, entre chaque représentation des tokens de X. Chaque ligne i de cette matrice correspond au token i de X, tandis que chaque colonne j représente le score d'attention entre le token i et le token j. Plus un score est élevé, plus le lien entre deux tokens est fort. Les scores sont ensuite normalisés par la racine carrée de la dimension réduite d_h. Puis, l'opération $\mathrm{softmax}$ est appliquée à l'ensemble des scores pour les convertir en une distribution de probabilité. La définition standard de la fonction $\mathrm{softmax}$ correspond à $\sigma : \mathbb{R}^L \to (0,1)^L$ où $L > 1$, prenant en entrée $z = (z_1, ..., z_L) \in \mathbb{R}^L$ (ici les scores obtenus par ligne) et calcule chaque composant vectoriel $\sigma(z) \in (0,1)^L$ avec :

$$\sigma(z)_i = \frac{e^{z_i}}{\sum_{j=1}^{L} e^{z_j}}$$

À ce stade, on a une matrice $A \in \mathbb{R}^{7 \times 7}$ où chaque ligne est une distribution de probabilité, et chaque élément $A_{i,j}$ indique comment un token i est pertinent face à un token j.

$$A = \sigma(\frac{QK^T}{\sqrt{d_h}})$$

Chaque ligne de V_i correspond à une représentation du token i, mais dans un espace de dimension d_h. La multiplication entre A et V permet d'obtenir la somme pondérée entre l'attention que portent les tokens (matrices A) et la valeur des tokens (matrice V). Le résultat de l'attention se calcule alors par :

$$\mathrm{Attention}(Q, K, V) = AV$$

Si on prend comme exemple la ligne A_6, chaque colonne correspondra à l'attention entre le token `voiture` et tous les autres tokens de la phrase. Supposons qu'on a $A_6 = [0\,;0\,;0,2\,;0,2\,;0,0\,;0,0\,;0,6]$. Dans ce cas, le token `orange` et ceux de `cassé` auront une influence sur la représentation du token `voiture`. La transformation de la représentation du token `voiture` sera alors calculé par la formule :

$$\mathrm{Attention}_{\texttt{voiture}} = [0,1V_3 + 0,1V_4 + 0,2V_7]$$

Une illustration de la transformation du plongement du token `voiture` par un mécanisme d'attention est présentée dans la Figure 1.9.

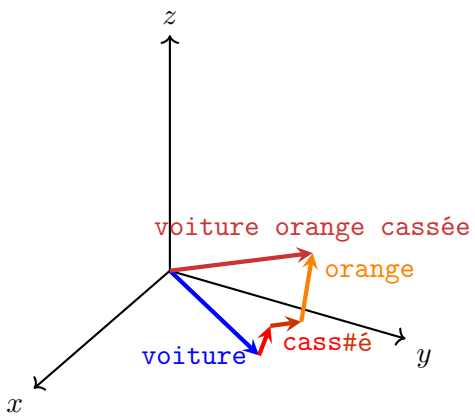


Figure 1.9 – Illustration de la transformation du plongement du token `voiture` par un mécanisme d'attention.

Cependant, la représentation calculée correspond à une seule tête d'attention, de dimension réduite d_h. Le calcul des nouvelles représentations des tokens en fonction du contexte est répété h fois pour chaque tête, en parallèle. Chacune des têtes capture une distribution différente du score d'attention. Les sorties des h têtes sont ensuite concaténées, puis projetées à nouveau dans la dimension du modèle à l'aide de la matrice $W^O \in \mathbb{R}^{hd_h \times d_{\text{modèle}}}$, par la formule suivante :

$$\text{MultiTete}(X) = \text{Concat}(\text{tete}_1, ..., \text{tete}_h)W^O$$

où

$$\text{tete}_i = \text{Attention}(XW_i^Q, XW_i^K, XW_i^V) \quad \forall i \in 1 \leq i \leq h$$

Avec $X \in \mathbb{R}^{n_{\text{séquence}} \times d_{\text{modèle}}}$ une séquence de tokens de taille fixe $n_{\text{séquence}}$ allant d'une dizaine à des millions de tokens de dimension $d_{\text{modèle}}$ selon le modèle.

Les matrices de paramètres W^Q et W^K sont apprises par descente de gradient afin que leurs projections de X évoluent au cours de l'entraînement pour renforcer les scores d'attention entre des tokens partageant une information. De même, la matrice W^V est entraînée similairement pour projeter les tokens pour construire des représentations contextuelles. Enfin, la matrice W^O est également apprise par descente de gradients afin de re-projeter et affiner la concaténation des h têtes d'attention dans la dimension du modèle.

1.4.2.2 Normalisation de couche et addition

La suite d'un bloc d'encodage est l'étape, ou la couche, de normalisation et d'addition, que l'on peut représenter de la manière suivante :

$$\text{Normalisation}(x + \text{SousCouche}(x)).$$

L'entrée $x + \text{SousCouche}(x)$ correspond à la somme entre la représentation initiale x et la sortie de l'étape précédente, ici celle de l'attention. Cette opération est appelée *connexion résiduelle*.

La connexion résiduelle permet d'éviter le problème de disparition du gradient, situation dans laquelle les gradients deviennent trop faibles pour que les paramètres du modèle continuent de s'ajuster, empêchant ainsi l'apprentissage. Si un tel problème se produit, la sous-couche va être ignorée ($\text{SousCouche}(x) = 0$) et la représentation initiale x va rester inchangée. Enfin, l'algorithme de normalisation de couche (Ba *et al.*, 2016) est appliqué pour stabiliser l'apprentissage à travers les couches du modèle.

1.4.2.3 Réseau à propagation avant

La dernière couche de l'architecture transformeur est un réseau de neurones à propagation avant (RPA), utilisé pour enrichir la représentation précédente, constituée de deux couches dont une d'activation ReLU (Fred et Agarap, 2018), tel que :

$$\text{RPA}(x) = \text{maximum}(0, xW_1 + b_1)W_2 + b_2.$$

Cette étape est importante pour le modèle, car, jusqu'à présent, les transformations appliquées aux représentations reposaient principalement sur des opérations linéaires. Comme vu dans la section 1.1, il est essentiel pour le modèle d'être capable d'apprendre des motifs non linéaires également.

Vaswani *et al.* (2017) ont choisi de construire le réseau de neurones en deux couches. La première couche (W_1, b_1) projette la représentation de dimension $d_{\text{modèle}=512}$ dans une dimension supérieure $d_{ff} = 2048$, également pour augmenter la capacité de représentation de chaque token. La seconde couche (W_2, b_2) ramène ensuite la représentation à la dimension initiale $d_{\text{modèle}}$.

Une couche d'addition et de normalisation suit celle du réseau RPA, ce qui marque la fin d'un bloc d'encodage. À partir de cette étape, les opérations d'un même bloc sont répétées N fois afin d'affiner progressivement les représentations des séquences. Après qu'un plongement soit passé par l'ensemble des blocs

d'encodages, les représentations des séquences pourront être traitées de deux manières différentes. Elles pourront être utilisées comme entrées pour d'autres tâches de traitement automatique du langage naturel (TALN), notamment pour l'extraction d'information textuelle à l'aide de BERT (Devlin *et al.*, 2019a). Elles pourront également être transmises à la partie décodeur du modèle afin de générer une nouvelle séquence de tokens, comme dans l'architecture BART (Lewis *et al.*, 2020).

Dans le chapitre suivant, nous allons voir leurs applications dans d'autres domaines, et notamment celui du TALN en santé mentale. L'objectif va être de montrer comment les modèles d'intelligence artificielle (IA) arrivent à tirer profit des informations présentes dans des dialogues entre personnes utilisatrices de réseaux sociaux pour détecter des signes associés aux troubles de la santé mentale.

CHAPITRE 2
DÉTECTION DU RISQUE EN SANTÉ MENTALE

Ce chapitre présente comment les architectures transformeur peuvent être utilisées dans différentes applications, telles que la modélisation de sujets et la classification zéro-coup. Ces applications sont utilisées dans de nombreux domaines, notamment celui de la santé mentale.

Dans les sections 2.1 et 2.2, nous voyons comment les architectures transformeur sont utilisées respectivement pour la modélisation de sujets, avec l'exemple de BERTopic (Grootendorst, 2022), et pour la classification zéro-coup, avec l'exemple du modèle BART entraîné sur MultiNLI [1]. Enfin, dans la section 2.3, nous voyons comment ces applications sont appliquées dans le domaine de la santé mentale.

2.1 Modélisation de sujets

BERTopic (Grootendorst, 2022), fait partie de la famille des *modèles de sujets*, une approche utilisée afin d'extraire les sujets présents dans un corpus textuel. À proprement parler, les modèles de sujets n'extraient pas directement les sujets présents dans un corpus, mais plutôt la distribution probabiliste de mots similaires, regroupés en sujets. Ainsi, une distribution de sujets va dépendre du corpus analysé, du prétraitement effectué et également des algorithmes utilisés pour calculer cette similarité. BERTopic s'utilise en deux parties, premièrement la création de la représentation de sujets, puis, à partir de cette représentation, l'extraction des sujets d'un corpus. BERTopic est composé de cinq étapes :

1. le plongement des données dans un espace latent,
2. la réduction de dimension,
3. le regroupement de données,
4. la représentation de sujets,
5. l'indexation des regroupements de sujets.

Chacune de ces étapes est modulable et BERTopic permet de les changer en fonction des besoins. Les prochaines sections présentent ces étapes avec la configuration par défaut de BERTopic.

1. `https://huggingface.co/facebook/bart-large-mnli`

2.1.1 Plongement des données

L'algorithme utilisé pour représenter les données textuelles dans un espace latent est crucial pour un modèle de sujets, car c'est cet algorithme qui détermine la similarité entre les mots. Comme évoqué dans les sections précédentes, les architectures transformeurs sont principalement utilisées lorsqu'on traite des données textuelles nécessitant du contexte. L'architecture BERT (Devlin *et al.*, 2019a) est souvent mentionnée lorsqu'il s'agit d'obtenir une représentation dans un espace latent textuel. BERT reprend la partie encodeur de l'architecture transformeur pour créer des représentations vectorielles du texte, puis une tête de classification est entraînée pour extraire, à partir d'un token, une sortie spécifique. Le modèle a été entraîné à détecter les relations entre deux phrases : un token `[SEP]` a été ajouté au vocabulaire pour indiquer une séparation entre une phrase A et une phrase B. BERT a également été entraîné à associer au token `[CLS]` un rôle de classification de la phrase. La méthode initialement proposée pour calculer la similarité entre deux phrases A et B consiste à concaténer les phrases A et B avec le token `[SEP]`, puis à entraîner un réseau de neurones à prédire leur similarité à partir du token `[CLS]`. Cependant, deux problèmes majeurs existent avec cette approche : premièrement, l'utilisation du token `[CLS]` pour représenter une phrase s'est révélé peu efficace (Reimers et Gurevych, 2019). Deuxièmement, calculer la similarité entre toutes les paires de phrases entraîne une complexité en $O((n(n-1))/2) = O(n^2)$. Pour éviter ces problèmes, BERTopic utilise SBERT (Reimers et Gurevych, 2019) (Sentence BERT) afin d'obtenir une meilleure représentation des phrases dans un espace latent. SBERT est entraîné à produire des plongements de phrases cohérents en : (1) Faisant passer deux phrases dans une architecture BERT siamoise (deux encodeurs partageant les mêmes poids). (2) Appliquant une opération d'agrégation sur les plongements des tokens de chaque phrase afin d'obtenir une représentation unique par phrase. (3) Calculant la similarité cosinus [2] entre les deux représentations, puis en mettant à jour les poids du modèle par rétropropagation en fonction du score de similarité attendu. En apprenant à rapprocher les phrases similaires et à éloigner les phrases différentes, SBERT construit un espace vectoriel dans lequel la similarité cosinus correspond à la similarité sémantique. Cette approche permet à BERTopic de projeter les phrases d'un corpus dans un espace vectoriel de grande dimension, où la proximité correspond à la similarité sémantique.

2. Sachant que A et B sont deux vecteurs d'attributs à n dimensions, la similarité cosinus $\cos\theta$ est définie tel que :

$$\textit{similarité cosinus} = S_C(A,B) := \cos(\theta) = \frac{A \cdot B}{\|A\| \, \|B\|} = \frac{\sum_{i=1}^{n} A_i B_i}{\sqrt{\sum_{i=1}^{n} A_i^2} \cdot \sqrt{\sum_{i=1}^{n} B_i^2}}$$

Avec A_i et B_i, respectivement, les $i$èmes composantes des vecteurs A et B.

2.1.2 Réduction de dimension

BERTopic se sert de SBERT pour représenter les phrases d'un corpus dans un espace latent et utiliser un algorithme de regroupement pour distinguer les sujets. Cependant l'espace produit par SBERT et d'autres approches similaires sont de grande dimension. Dans de tels cas, il est difficile pour les algorithmes de regroupement de former des groupes, particulièrement à cause de la malédiction de la dimensionnalité (Bellman, 1966), qui rend les distances de moins en moins significatives lorsque le nombre de dimensions augmente. Ainsi, BERTopic utilise une étape de réduction de dimension pour faire une transformation dans un espace de dimension exploitable (par exemple, 5) par les algorithmes de regroupement. UMAP (McInnes *et al.*, 2018) est l'algorithme utilisé par BERTopic pour réduire la dimension de l'espace. D'autres algorithmes, comme T-SNE (van der Maaten et Hinton, 2008), PCA (Maćkiewicz et Ratajczak, 1993) ou LDA (Blei *et al.*, 2003) peuvent être utilisés à la place, cependant UMAP a pour avantage d'apprendre la topologie de l'espace latent, de construire un graphe dans l'espace de grande dimension puis de le projeter et de l'optimiser dans un espace de plus petite dimension. UMAP fait ainsi une réduction non-linéaire, tout en préservant à la fois la structure globale et locale des points.

2.1.3 Algorithmes de regroupement

Après la réduction de dimension, BERTopic peut utiliser un algorithme de regroupement afin de distinguer les sujets. Par défaut BERTopic utilise HDBSCAN pour la création de groupes de sujets. Comme vu dans la section 1.2.3, HDBSCAN est un algorithme de regroupement basé sur la densité, ce qui améliore sa capacité à former des groupes de sujets complexes. Cependant BERTopic possède un argument pour obtenir le plongement d'un sujet en calculant le centroïde des points d'un groupe de sujets. Utiliser un algorithme de densité pourrait créer des plongements de sujets se situant en dehors du groupe de sujets. Ce problème est mentionné plus en détail dans la section 3, où une approche proposée est d'utiliser l'algorithme K-moyennes.

2.1.4 Indexation des groupes de sujets

Les deux dernières étapes de BERTopic sont la vectorisation et la construction de représentations de sujets. Ces étapes attribuent une étiquette textuelle à chaque sujet généré par BERTopic, facilitant ainsi l'interprétation des sujets. La vectorisation crée, pour chaque groupe de sujets, un sac de mots contenant les termes issus des plongements. Les représentations textuelles des sujets sont créées par BERTopic en utili-

sant une version modifiée de TF-IDF (Salton et Buckley, 1988). TF-IDF est utilisé pour mesurer l'importance des termes dans un document par rapport à l'ensemble d'un corpus. L'importance est calculée en fonction de la fréquence du terme dans un document (TF) et de l'inverse de son nombre d'occurrences dans l'ensemble du corpus (IDF).

L'équation 2.1 montre comment calculer le score TF-IDF, en multipliant la fréquence $\mathrm{TF}_{t,d}$ d'un terme t dans un document d par le logarithme du nombre total de documents N, divisé par le nombre d'occurrences DF du terme t dans l'ensemble des documents.

$$\text{TF-IDF}(t, d) = \mathrm{TF}_{t,d}. \log(\frac{1+N}{\mathrm{DF}_t}) \tag{2.1}$$

BERTopic utilise une version modifiée, c-TF-IDF pour mesurer l'importance des termes présents dans un groupe de sujets parmi l'ensemble des groupes générés par BERTopic.

L'équation 2.2 montre la formule de c-TF-IDF, avec $\mathrm{tf}_{t,c}$ qui représente la fréquence d'un terme t dans un groupe c, A, qui correspond au nombre moyen de termes par groupe, et f_t qui est la fréquence du terme t parmi l'ensemble des groupes.

$$\text{c-TF-IDF}(t, c) = \mathrm{TF}_{t,c}. \log(\frac{1+A}{f_t}) \tag{2.2}$$

2.2 Classification zéro-coup

Dans la section précédente, BERTopic a été présenté comme une approche permettant de projeter des données textuelles dans un espace latent afin de regrouper des documents par sujets. Cette méthode, comme d'autres techniques de regroupement ou de classification, nécessite des données pour créer un ensemble de sujets, ou effectuer de la classification. Récemment, de nouvelles approches basées sur les transformeurs ont permis l'émergence de modèles capables de réaliser certaines tâches spécifiques liées à un domaine sans nécessiter d'affinage. Traditionnellement, les modèles basés sur une architecture transformeur sont une première fois entraînés sur de larges corpus généraux (pré-entraînement), puis une seconde fois sur des données spécifiques à un domaine cible (affinage). On parle alors d'approches zéro-coup, et notamment de classification zéro-coup, lorsque le modèle est capable d'effectuer la classification d'un texte en s'appuyant uniquement sur son pré-entraînement, sans affinage spécifique dans le domaine cible.

2.2.1 Inférence en langage naturel

L'inférence en langage naturel (NLI) est une tâche utilisée pour de nombreux modèles de TALN afin d'apprendre une relation logique entre deux phrases : la prémisse et l'hypothèse. L'objectif de la tâche est de déterminer si l'hypothèse est vraie (implication) ou fausse (contradiction) face à la prémisse. De nombreux domaines du TALN, comme les systèmes de question-réponse, le résumé automatique ou la recherche d'information, sont entraînés ou évalués sur cette tâche pour aider le modèle à capturer le sens des phrases. Des ensembles de données comme Multi-Genre Natural Language Inference (MNLI) (Williams *et al.*, 2018) augmentent la complexité de la tâche en ajoutant la condition neutre entre deux phrases, c'est-à-dire qu'on ne peut pas conclure l'hypothèse, et en incluant des sources de données diverses.

2.2.2 Classification zéro-coup

Yin *et al.* (2019) proposent une nouvelle approche de classification zéro-coup en utilisant l'ensemble de données MNLI avec l'architecture transformeur BERT pour la classification zéro-coup. BERT est affiné pour prendre en entrée une prémisse et une hypothèse et apprendre une tête de classification à prédire une des sorties : implication, neutre ou contradiction. L'approche proposée consiste alors à prendre la phrase qu'on souhaite classifier comme prémisse et choisir les étiquettes comme hypothèses. Si le modèle prédit comme sortie : implication, alors c'est que l'étiquette est vraie pour la phrase donnée. L'architecture du modèle a été améliorée par la suite en utilisant BART (Lewis *et al.*, 2020), un modèle de type transformeur encodeur-décodeur entraîné sur un plus grand nombre de données que BERT.

L'utilisation d'un modèle basé sur l'architecture transformeur, en plus d'un modèle de tokenisation, permet à des données textuelles d'être plongées sous forme de représentation vectorielle, capturant la complexité syntaxique et sémantique des mots. Ces représentations peuvent ensuite être utilisées pour construire des modèles de sujets, faire de la classification zéro-coup, ou pour encore d'autres applications avec des outils issus du TALN.

La prochaine section a pour objectif de présenter les différentes applications de ces techniques dans le domaine de la santé mentale. Plus précisément, elle présentera les applications à l'état de l'art dans la détection des signes associés aux troubles de la santé mentale.

2.3 Applications en santé mentale

Avant la forte utilisation de l'architecture transformeur, différentes méthodes étaient choisies pour la détection de signes associés aux troubles en santé mentale. Chiong *et al.* (2021) utilisent des techniques d'apprentissage automatique, comme l'étiquetage morphosyntaxique (Ratnaparkhi, 1996) et la lemmatisation (Koskenniemi, 1984), afin d'entraîner un modèle à la détection des signes de dépression sur les réseaux sociaux. D'autres approches, comme celle de Ta *et al.* (2022), se concentrent sur la détection de signes d'anxiété au sein de communautés en ligne, en analysant les relations, à l'aide de graphes, entre les personnes utilisatrices et les sujets abordés.

L'architecture transformeur (Vaswani *et al.*, 2017) a par la suite introduit de nouvelles approches qui se sont largement répandues. Des comparaisons entre différentes techniques pour la détection de risques en santé mentale montrent que les approches liées à l'utilisation de modèles basés sur l'architecture transformeur donnent de meilleurs résultats (Sihab-Us-Sakib *et al.*, 2024; Bokolo et Liu, 2024). Benitez-Andrades *et al.* (2021) présentent une comparaison entre six modèles encodeur : BERT (Devlin *et al.*, 2019b), RoBERTa (Liu *et al.*, 2019), DistilBERT (Sanh *et al.*, 2020), CamemBERT (Martin *et al.*, 2020), ALBERT (Lan *et al.*, 2020) et FlauBERT (Le *et al.*, 2020) pour détecter les comportements liés aux troubles alimentaires (TA) sur les réseaux sociaux. MentalBERT (Ji *et al.*, 2021) est un modèle basé sur l'architecture BERT et entraîné sur des données spécifiques pour détecter des comportements associés aux symptômes de la dépression, de l'anxiété, du trouble de stress post-traumatique et du trouble de la personnalité limite. Karamat *et al.* (2024) créent une architecture hybride en utilisant MentalBERT pour la détection de signes associés aux troubles, et MelBERT (Choi *et al.*, 2021) pour la compréhension de métaphores. Cela a permis d'améliorer la détection pour des données issues de réseaux sociaux. Enfin, Lim *et al.* (2025); Chen *et al.* (2025) utilisent également des données audio pour enrichir la représentation textuelle en utilisant le modèle XLM-RoBERT (Conneau *et al.*, 2019).

L'utilisation de bases de connaissances est également présente dans de nombreuses approches afin de faciliter la détection de certains signes. Une base de connaissances correspond à une source de données externe au modèle, qui n'a pas été utilisée lors de l'apprentissage. Ces connaissances peuvent être utilisées par un modèle afin d'enrichir les représentations ou de guider l'interprétation des résultats en s'appuyant sur des informations validées par des personnes spécialistes du domaine. Yan *et al.* (2025) et Su *et al.* (2025) présentent un modèle de détection de signes associés aux troubles de la dépression à partir des symptômes décrits dans le livre médical de référence *Diagnostic and Statistical Manual of Mental Di-*

sorders (DSM-5) (Crocq *et al.*, 2015). D'autres approches utilisent comme bases de connaissances des questionnaires d'auto-évaluation pour entraîner un modèle à répondre à ses questions (Maupomé *et al.*, 2024). Cela permet d'éviter de faire une prédiction directe sur le trouble, mais plutôt d'évaluer les comportements ou les symptômes associés.

La modélisation de sujets est également utilisée dans ce domaine, notamment afin de représenter des sujets de discussion. Durant les campagnes d'évaluation eRisk [3] 2023 et 2024, la tâche 3 correspondait à la détection de signes du TA (Parapar *et al.*, 2024). En 2023, Rujas *et al.* (2023) ont utilisé BERTopic (Grootendorst, 2022) pour la prédiction des signes associés aux TA à partir de messages issus de réseaux sociaux. En 2024 également, une approche (Maupomé *et al.*, 2024) utilisant BERTopic a obtenu sur plusieurs métriques des résultats proches des meilleurs (Prasanna *et al.*, 2024).

Cependant, la plupart des méthodes présentées nécessitent à la fois des données d'entraînement et des ressources de calcul. Plus on a de ressources, plus les modèles montrent de bons résultats. Mais que faire lorsqu'on ne peut pas entraîner de modèle ?

Dans cette situation, une solution potentielle serait d'utiliser les méthodes de classification zéro-coup, une approche qui s'est révélée prometteuse dans d'autres domaines (Zhang *et al.*, 2020; Li *et al.*, 2017). Leow *et al.* (2025) montrent également que cette approche rivalise avec des modèles entraînés dans le domaine de la santé mentale.

Les deux prochains chapitres présentent des travaux réalisés et publiés dans deux articles, détaillant des applications d'outils du TALN en santé mentale. Le premier travail concerne l'application de modèle de sujets, sur des publications textuelles en ligne, pour la détection des comportements associés aux troubles de la santé mentale. Le second travail concerne l'application de modèle de classification zéro-coup pour la sélection ordonnée de phrases associées aux symptômes de la dépression.

3. eRisk `https://erisk.irlab.org`

CHAPITRE 3
DÉTECTER DES COMPORTEMENTS ASSOCIÉS AUX TROUBLES ALIMENTAIRES PAR L'ANALYSE AUTOMATIQUE DES PUBLICATIONS TEXTUELLES EN LIGNE

3.1 Contexte et références

Les travaux présentés dans cette section s'inscrivent dans la continuité de ceux réalisés lors de la tâche 3 de la campagne d'évaluation eRisk 2024 (Parapar *et al.*, 2024). L'objectif de cette tâche était de concevoir une approche permettant de détecter automatiquement des comportements associés aux troubles alimentaires chez des personnes utilisatrices de réseaux sociaux. Les recherches initiales ont porté sur le développement de modèles de sujets pour l'analyse textuelle des publications, suivi de l'entraînement d'un réseau de neurones afin d'identifier des comportements associés aux troubles alimentaires. Les travaux décrits dans la section 3.2 prolongent cette démarche en explorant plus en détails la manière dont les données textuelles sont représentées dans un espace latent afin de former des regroupements de sujets. Ils examinent également l'intérêt d'un filtrage par sujet pour affiner la détection des comportements ciblés. Cette analyse vise à approfondir la compréhension des méthodes de modélisation de sujets et à identifier les paramètres influençant la qualité de l'extraction sémantique à partir des ensembles de données textuelles.

L'article correspondant, présenté en section 3.2, a été rédigé et présenté par l'auteur de ce mémoire à la conférence *Traitement Automatique des Langues Naturelles (TALN)*, CORIA-TALN 2025. L'article est publié en libre accès (Ferstler *et al.*, 2025a), et le code source des travaux est également disponible publiquement sous licence libre GPL-v3, dans le dépôt GitLab (GitLab, 2025b). L'auteur du mémoire est à l'origine de la formulation de l'hypothèse, de la rédaction principale du manuscrit, ainsi que de l'intégralité de la conception et de l'implémentation du code source, avec la collaboration des co-autrices pour la rédaction, la relecture et la validation scientifique.

3.2 Article

Détecter des comportements associés aux troubles alimentaires par l'analyse automatique des publications textuelles en ligne

Yves Ferstler, Catherine Lavoie, Marie-Jean Meurs
Université du Québec à Montréal

Résumé

Cet article présente une méthode pour détecter des aspects du comportement liés aux troubles alimentaires à partir de publications textuelles échangées sur les réseaux sociaux. Nos travaux comparent différentes représentations d'historiques de publications permettant d'entraîner un modèle neuronal pour la prédiction. Les approches étudiées sont : (1) la représentation de sujet par fréquence, en calculant le nombre de sujets apparus dans un historique, (2) une représentation par plongement, en calculant la moyenne des représentations de sujets présents dans l'historique de publications, (3) une représentation par documents représentatifs, qui cherche à représenter un sujet par un document sémantiquement proche. Un filtrage de sujets est également étudié, pour sélectionner les sujets reliés aux troubles alimentaires. Les résultats montrent que l'utilisation de filtrage permet d'améliorer les performances des systèmes de détection. La méthode basée sur un document représentatif obtient les meilleurs résultats, parmi les autres représentations évaluées mais également parmi d'autres méthodes appliquées à la même tâche lors de la campagne d'évaluation eRisk 2024.

Abstract Detecting signs of eating disorders through automatic analysis of online text conversations

This paper leverages a topic modeling algorithm to predict aspects of eating disorders from a dataset of social network publications. Specifically, it investigates three methods for representing users' histories : (1) a frequency-based approach that calculates topic frequencies across a user's history, providing a quick but potentially insightful representation. (2) an embedding-based approach that uses a cluster topic representation where two clustering algorithms dense and centroid-based are tested to enrich semantic information. (3) a document-based approach that refines cluster embeddings by identifying the single most represen-

tative document, thereby mitigating the limitations observed in the embedding-based approach. A filter approach is also evaluated, focused on documents that share similarities with a standard questionnaire for self-evaluation of eating disorders. The results indicate that all representative approaches with filtering improved performance. Among the three representation methods, the representative document achieves the highest scores. Compared to other approaches applied on the same task, this method shows better results in 4 over 8 metrics compared to all other approaches proposed by the eRisk 2024 evaluation campagn participants.

Mots-clés : Modèle de sujet, Troubles alimentaires, Représentation d'historique conversationnel
Keywords : Topic model, Eating disorders, Representation of user history of publications

3.2.1 Introduction

Les troubles alimentaires (TA) sont complexes et ont de nombreux impacts sur la vie d'une personne et son bien-être. Caractérisés par une alimentation anormale culturellement, des craintes nutritionnelles, ou des préoccupations persistantes liées à l'image corporelle (Crocq *et al.*, 2015), les TA sont variés. De l'anorexie mentale à la boulimie ou encore les accès hyperphagiques, chacun de ces troubles montre des comportements associés variables et des symptômes différents. Plusieurs facteurs génétiques, psychologiques, sociaux ou culturels, peuvent favoriser le développement de troubles alimentaires chez une personne (Crocq *et al.*, 2015). Les symptômes pouvant être difficiles à détecter et confondus avec ceux d'autres troubles, l'identification de TA chez une personne atteinte peut se révéler complexe. C'est pour cette raison que des questionnaires d'auto-évaluation (Aardoom *et al.*, 2012a) sont utilisés dans le processus diagnostique.
Il est aujourd'hui courant de partager ses sentiments ou ses expériences en ligne : communiquer avec d'autres personnes, partager ses préoccupations et chercher de l'aide sur internet est devenu une habitude pour de nombreuses personnes (Sowles *et al.*, 2018). Des corpora de messages et de discussions entre internautes sont disponibles pour la communauté de recherche et peuvent être utilisés pour des projets liant le domaine du traitement automatique du langage naturel (TALN) et celui de la santé mentale. L'analyse automatique des conversations textuelles peut en effet aider à identifier des comportements associés aux troubles mentaux, et notamment de TA. Pour entraîner un réseau neuronal sur des tâches textuelles, une méthode de représentation des données doit être choisie au préalable.
Dans ce travail, l'approche utilisée pour représenter le texte brut en message interprétable est l'utilisation

d'un modèle de sujet, qui permet d'identifier les différents sujets abordés par une personne utilisatrice dans un historique conversationnel.
Dans un premier temps nous nous intéressons à la représentation de l'historique des publications par le modèle de sujet, en utilisant les méthodes de fréquences, de plongement et de document représentatif sémantiquement proche. Dans un second temps, nous examinons comment utiliser ces représentations et si une sélection de sujets permet d'améliorer les performances des systèmes de détection proposés.

3.2.2 État de l'art

Selon l'organisation mondiale de la santé, en 2022, une personne sur 8 dans le monde était affectée par un ou plusieurs troubles mentaux et 14 millions de personnes dont 3 millions d'enfants et adolescents avaient déjà souffert de TA [1]. L'aide que peut apporter le TALN aux recherches dans le domaine de la santé mentale et plus spécifiquement des TA est donc essentielle pour tenter d'aider les personnes qui souffrent. Parmi les travaux récents en TALN, Chiong *et al.* (2021) proposent une méthode de détection des signes de dépression sur les réseaux sociaux en utilisant des techniques de correction de mots, d'étiquetage Parts-Of-Speech (POS) et de lemmatisation, afin d'entraîner un modèle de classification. D'autres techniques utilisent des transformeurs puis des modèles d'apprentissage pour détecter des signes de cyberharcèlement et de comportements associés aux TA (Sihab-Us-Sakib *et al.*, 2024; Karamat *et al.*, 2024).

Rujas *et al.* (2023) présentent une méthode qui combine le prétraitement de données avec la modélisation de sujet en utilisant BERTopic (Grootendorst, 2022), et qui obtient des résultats prometteurs pour la détection précoce des TA à partir de messages textuels. Benitez-Andrades *et al.* (2021) proposent une comparaison de six modèles de langue préentrainés – BERT (Devlin *et al.*, 2019b), RoBERTa (Liu *et al.*, 2019), DistilBERT (Sanh *et al.*, 2020), CamemBERT (Martin *et al.*, 2020), ALBERT (Lan *et al.*, 2020), FlauBERT (Le *et al.*, 2020) – pour détecter les comportements liés aux TA sur les réseaux sociaux. Ces modèles obtiennent des résultats intéressants, particulièrement avec l'utilisation de RoBERTa. Tout en limitant la surconsommation de ressources computationnelles, des modèles plus frugaux comme ceux de Wang *et al.* (2020) obtiennent de bons résultats autant en termes d'efficacité que de performance.
Les travaux présentés précédemment indiquent que l'emploi de modèles de sujet pour représenter des historiques conversationnels est un choix pertinent. Parmi ces modèles, BERTopic permet d'obtenir des résultats à l'état de l'art dans le domaine de la détection du risque en santé mentale. Dans le cadre de la cam-

1. OMS, 2022 `https://www.who.int/news-room/fact-sheets/detail/mental-disorders`

pagne d'évaluation eRisk 2024 [2], les équipes participant à la tâche 3 ont travaillé sur l'évaluation de signes de troubles alimentaires, et parmi les résultats (Parapar *et al.*, 2024), l'approche utilisant BERTopic (Maupomé *et al.*, 2024) a obtenu sur plusieurs métriques des résultats proches des meilleurs (Prasanna *et al.*, 2024). Cette équipe a obtenu les meilleurs résultats à partir de 3 méthodes différentes, elles combinent des techniques de Word2Vec, de rétrotraduction, de réduction de dimensions (PCA) et de classifieurs SVM.

3.2.3 Méthodologie

Notre objectif est de représenter les historiques de publications en données interprétables pour l'entraînement d'un réseau neuronal, tout en préservant leur richesse informationnelle. Nos travaux ont donc porté sur le choix de l'ensemble de données et la représentation des publications et des historiques pour entraîner le modèle prédictif.

3.2.3.1 Ensemble de données

Le choix d'un ensemble de données dans un contexte de TA nécessite de porter une attention particulière à l'équilibre des données, à la présence de faux positifs et de faux négatifs. L'ensemble de données choisi dans les travaux présentés est celui de la tâche 3 de la conférence eRisk 2024, qui propose un ensemble de données issues de réseaux sociaux. La tâche de 2024 est la même que celles de 2022 et 2023. L'ensemble est constitué des données de ces deux années pour l'entraînement du modèle, et des données de test de 2024 pour l'évaluation. Ces ensembles contiennent l'historique conversationnel d'une personne utilisatrice, associée à ses réponses au questionnaire EDE-Q (*The Eating Disorder Examination Questionnaire*) (Fairburn et Beglin, 2008), un questionnaire standard d'auto-évaluation sur les comportements généralement associés aux TA.

Le questionnaire EDE-Q comporte 28 questions visant à évaluer les comportements associés à certains troubles alimentaires durant une période de temps précise. Les réponses varient sur une échelle de 0 à 6, où **0** indique que le ou les comportements associés aux TA évalués dans la question ne se sont jamais manifestés, et **6** indique qu'ils se sont produits quotidiennement. Ainsi, plus le score est élevé, plus forte est la prévalence du ou des comportements.

Quatre sous-échelles portent sur différents aspects des TA : la restriction alimentaire, les préoccupations

2. eRisk 2024 `https://erisk.irlab.org/2024/index.html`

alimentaires, la silhouette, le poids. Le score global est obtenu en combinant les résultats des quatre sous-échelles et en faisant la moyenne des réponses attribuées à chaque question. Le tableau 3.1 fournit des exemples de questions présentes dans le EDE-Q et leur traduction libre en français.

#	Question	Sous-échelle
1.	*Have you been deliberately **trying** to limit the amount of food you eat to influence your shape or weight (whether or not you have succeeded) ?* – Avez-vous délibérément essayé de limiter la quantité de nourriture que vous mangez pour influencer votre forme ou votre poids (que vous ayez réussi ou non) ?	Restriction alimentaire
7.	*Has thinking about **food, eating or calories** made it very difcult to concentrate on things you are interested in (for example, working, following a conversation, or reading) ?* – Penser à la nourriture, à l'alimentation ou aux calories vous a-t-il empêché de vous concentrer sur des choses qui vous intéressent (par exemple, travailler, suivre une conversation ou lire) ?	Préoccupations alimentaires
6.	*Have you had a definite desire to have a **totally flat** stomach ?* – Avez-vous déjà eu le désir absolu d'avoir un ventre totalement plat ?	Silhouette
12.	*Have you had a strong desire to lose weight ?* – Avez-vous eu un fort désir de perdre du poids ?	Poids

Table 3.1 – Exemples de questions du EDE-Q avec leurs sous-échelles

Le score global et les scores des sous-échelles sont utilisés pour interpréter les réponses du questionnaire (Mond *et al.*, 2004b; Institute for Eating Disorders, 2 31; Aardoom *et al.*, 2012b; Mond *et al.*, 2004a). Une interprétation courante du score global est qu'un score plus élevé reflète des comportements associés aux TA plus marqués (Aardoom *et al.*, 2012b), un score proche de 3 est également utilisé comme seuil pour différencier les cas positifs des cas négatifs (Institute for Eating Disorders, 2 31). Parmi les 28 questions, seuls

Statistiques		**Mots les plus fréquents**	
Métrique	**Valeur**	**Mot**	**Nb. occurences**
Nombre de personnes utilisatrices	74	***like***	7 214
Nombre de publications	32 806	***people***	4 319
Nombre de mots total	1 185 532	*get*	3 909
Taille du vocabulaire	104 757	***think***	3 470
Médiane de mots par publication	17,0	*would*	3 445
1er Quartile de mots par publication	6,0	*one*	3 142
3ème Quartile de mots par publication	43,0	*know*	3 098
Moyenne de publications par personne	443,32	*really*	2 938
Médiane de publications par personne	196,0	*time*	2 844
1er Quartile de publications par personne	86,5	***feel***	2 584
3ème Quartile de publications par personne	912,5	*much*	2 246
Moyenne de publications mensuelles par personne	144,15	*also*	2 219
Médiane de publications mensuelles par personnes	59,5	***want***	2 152

Table 3.2 – Description de l'ensemble de données d'entrainement de la tâche 3 de eRisk 2024

22 sont utilisées pour la cotation des sous-échelles et du score global (les questions 13 à 18 sont exclues car les réponses attendues sont en texte libre).

Le corpus d'entraînement contient 74 historiques de conversations textuelles, chacun étant associé à une personne utilisatrice. Parmi les 74 personnes utilisatrices, la moitié montre des comportements associés aux TA, le corpus est donc équilibré. Une description détaillée est fournie dans le tableau 3.2.
Les Figures 3.1 et 3.2 montrent en bleu le groupe présentant des comportements associés aux TA et en orange le groupe sans ces comportements, en utilisant le seuil de 3 du score global. Les tendances des scores entre les deux groupes sont similaires, que ce soit pour les sous-échelles ou pour les questions individuelles. Les personnes utilisatrices sans comportements associés aux TA partagent certaines préoccupations des personnes présentant ces comportements, mais avec une fréquence moindre. Une différence majeure entre les groupes réside dans les scores EDE-Q, qui sont proportionnellement plus élevés pour les personnes utilisatrices présentant des comportements associés aux TA.

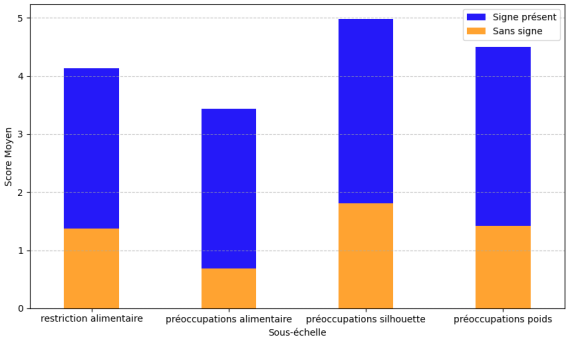


Figure 3.1 – Score moyen des sous-échelles entre les personnes utilisatrices ayant des comportements associés aux TA et celles sans.

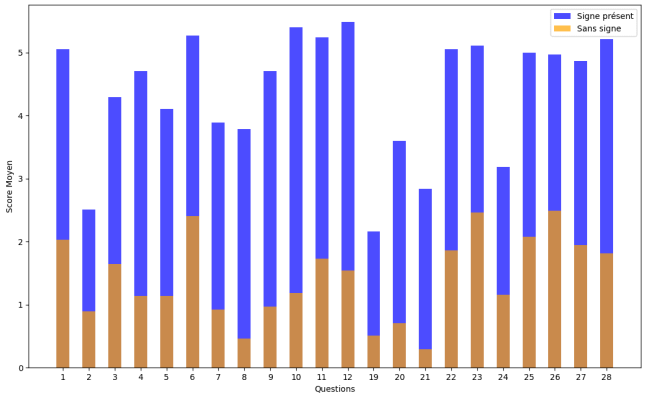


Figure 3.2 – Score moyen des questions entre les personnes ayant des comportements associés aux TA et celles sans.

Les statistiques du tableau 3.2 montrent un nombre de publications relativement élevé (32 806) pour un nombre de personnes utilisatrices faible (74). Avec une médiane de 196 publications et un troisième quartile de 912 publications par personne utilisatrice, une majorité de l'information de l'ensemble de données sera contenu dans une minorité d'historiques conversationnels. Pour permettre au modèle neuronal d'observer une plus grande diversité d'historiques, nous avons segmenté les historiques originaux en portions d'historiques de 350 publications (au maximum). Les réponses originales aux EDE-Q sont associées à la portion d'historique contenant les publications les plus récentes. Pour éviter le sur-apprentissage, et sur recommandation de personnes cliniciennes, les autres portions d'historiques sont associées aux réponses EDE-Q originales, en ajoutant du bruit. Ainsi, pour 10% des réponses, la valeur est modifiée de $\pm$ 1 entre 1 et 5, de + 1 pour la valeur 0 et de -1 pour la valeur 6. Le corpus d'entraînement ainsi segmenté comporte 138 historiques conversationnels, majorés à 350 publication par historiques, avec une médiane de publications par personne passant de 196,0 à 335,0 et d'un 3ème quartile de publications par personne passant de 912,5 à 349,0.

3.2.3.2 Représentation

Représenter les caractéristiques des historiques sous une forme riche est essentiel pour obtenir du modèle neuronal la meilleure prédiction possible. Notre choix de représentation d'un historique comporte deux étapes : (1) la transformation des publications en représentation individuelles, puis (2) l'agrégation des représentations individuelles pour représenter complètement l'historique.

3.2.3.2.1 Représentation des publications

Nous avons choisi d'explorer une représentation par modélisation de sujet en utilisant BERTopic (Grootendorst, 2022), qui génère, lors de son l'entraînement, un sujet pour chaque publication. Les publications sémantiquement proches seront donc associées au même sujet et les sujets générés pendant l'entraînement seront utilisés pour étiqueter les nouvelles publications pendant la phase de test. BERTopic fournit également trois publications représentatives pour chaque sujet.

Nous avons choisi de représenter chaque sujet de trois manières différentes : soit par une valeur numérique, soit par son plongement dans BERTopic, soit par sa publication représentative la plus proche de la nouvelle publication à étiqueter. Dans ce dernier cas, le choix d'une publication représentative parmi les trois fournies par BERTopic se fait dynamiquement et la distance utilisée est la distance euclidienne.

Pour générer les sujets, BERTopic utilise HDBSCAN (Malzer et Baum, 2020), un algorithme de clustering hiérarchique basé sur des zones de forte densité des points représentant les documents, puis calcule le centroïde des clusters obtenus. Tel qu'illustré à la Figure 3.3, cette approche peut être problématique. En effet, bien que deux clusters distincts puissent être correctement identifiés, il est possible que leurs sujets soient confondus.

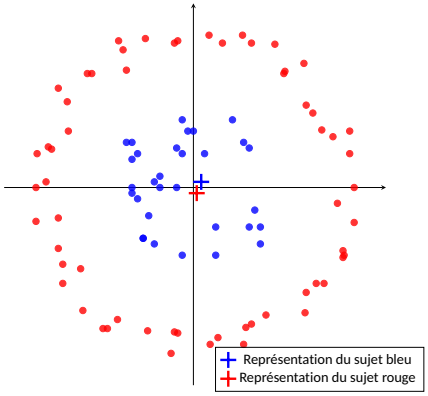


Figure 3.3 – Exemple de scénario problématique : deux clusters (un rouge, un bleu) et leurs centroïdes sous forme de plus (+) situées à des coordonnées voisines.

Pour éviter ce problème, nous avons aussi utilisé l'agorithme K-means car il génère des clusters et donc des centroïdes topologiquement éloignés. Le nombre de clusters imposé à K-means est le même que le celui déterminé par HDBSCAN, soit 413.

3.2.3.2.2 Représentation des historiques

Trois représentations d'historique sont évaluées : une basée sur la fréquence des sujets présents et deux basées sur les plongements – par publications et par documents représentatifs.
L'approche basée sur la fréquence représente l'historique d'une personne utilisatrice en calculant la fréquence d'apparition de chaque sujet. Cette représentation est efficace en termes de calcul et facilement interprétable puisqu'elle associe une liste de sujets à un historique. Cependant, cette simplicité limite la richesse de l'information encodée, ce qui peut affecter la performance prédictive du modèle neuronal. De plus, lorsque l'ensemble des sujets est large ou que le nombre de publications dans l'historique est faible, la représentation peut contenir une proportion significative de valeurs nulles, réduisant la performance du modèle neuronal.

L'approche basée sur la moyenne permet de représenter l'historique en calculant la moyenne des plongements associés aux publications ou la moyenne des plongements associés aux documents représentatifs. La taille de cette représentation est identique à celle des plongements utilisés. Cette méthode favorise une plus grande richesse de représentation.

3.2.3.2.3 Filtrage

Quelle que soit la technique de représentation, toutes les informations extraites ne sont pas forcément pertinentes dans le contexte de la détection des comportements associés aux TA. Ainsi, les sujets générés reflètent la globalité des conversations et l'utilisation de tous les sujets peut être donc inadaptée à la détection des comportements associés aux TA. Par conséquent, un filtrage des publications est appliqué afin d'identifier les plus pertinentes pour la représentation de l'historique. Pour cela, une sélection de sujets est faite regroupant les sujets attribués à chaque question du questionnaire EDE-Q par BERTopic. Afin d'élargir cette sélection, des sujets sémantiquement proches de ceux associés aux questions sont également ajoutés à la liste. Un ratio est utilisé pour définir la proportion de sujets à filtrer. Comme ce ratio est appliqué indépendamment à chaque question, il ne reflète pas directement le nombre total de sujets filtrés.

3.2.3.2.4 Modèle de prédiction de l'EDE-Q

Le modèle de prédiction pour le questionnaire EDE-Q est inspirée de (Maupomé *et al.*, 2024). Ce modèle utilise trois réseaux neuronaux à action directe (*feed-forward networks*). Le premier réseau, entraîné à par-

tir des représentations d'historiques conversationnels, prédit le score global obtenu au questionnaire. Le second réseau, entraîné à partir du modèle du score global et des représentations d'historiques conversationnels, prédit les scores des sous-échelles. Enfin, le troisième réseau, entraîné à partir des modèles du score global, des sous-échelles et des représentations d'historiques conversationnels, prédit les réponses aux questions.

3.2.4 Expériences et résultats

3.2.4.1 Métriques

Les métriques utilisées pour mesurer la performance des approches proposées sont celles de la campagne d'évaluation eRisk 2024, ce qui permet de comparer nos résultats à ceux des équipes participantes. Ces métriques sont les suivantes :

- **Mean Zero-One Error (MZOE)** mesure la proportion de réponses des personnes utilisatrices pour lesquelles la réponse prédite est incorrecte, peu importe le degré de l'erreur.
- **Mean Absolute Error (MAE)** est la moyenne sur tous les historiques des écarts moyens entre la réponse prédite et la réponse réelle des personnes utilisatrices.
- **Macro-averaged MAE (MAE_macro)** est la moyenne des MAE sur tous les historiques des écarts moyens pour chaque catégorie de réponse (définie par un score allant de 0 à 6).
- **Restriction (RS, questions 1 à 5), Préoccupation alimentaire (ECS, questions 7, 9, 19, 21, 20), Préoccupation de la silhouette (SCS questions 6, 8, 10, 11, 23, 26, 27, 28) et Préoccupation du poids (WCS, questions 8, 12, 22, 24, 25)** se concentrent sur des sous-échelles spécifiques des troubles alimentaires en calculant la racine de l'écart quadratique moyen entre la moyenne des scores prédits par le modèle et ceux des vraies réponses sur les questions associées.
- **Score global (GED)** est la même métrique appliquée sur les quatre sous-échelles précédentes afin de fournir une évaluation globale des performances prédictives sur l'ensemble du questionnaire.

3.2.4.2 Contrainte temporelle imposée par le questionnaire EDE-Q

Le EDE-Q 6.0 déclare que les questions doivent être répondues selon les évènements des 28 derniers jours (Fairburn et Beglin, 2008). Cependant, la proportion de données datant des 28 derniers jours est minimale comparée au corpus complet. L'utilisation de filtrage de sujets réduisant la taille de l'ensemble de données, une restriction supplémentaire du corpus aux publications des 28 jours les plus récents limite

	Sans filtrage	Filtrage moyen	Filtrage élevé	Filtrage fort
K-means	441	231	144	83
HDBSCAN	412	224	148	85

Table 3.3 – Proportion de sujets générés par BERTopic en fonction du filtrage est présentée sur le tableau.

la capacité d'entraîner le modèle neuronal. Nous avons donc examiné si les publications plus anciennes pourraient aider à identifier des comportements associés aux TA en comparant les résultats obtenus sur le sous-corpus contenant uniquement les publications des 28 derniers jours et sur un corpus sans restriction temporelle. Afin d'assurer une répartition équitable du nombre de publications entre les deux corpus, le nombre de publications du second est réduit à celui du premier en supprimant de manière aléatoire certaines publications de l'historique complet. Les résultats obtenus montrent que l'utilisation du corpus sans restriction de date et contenant la même quantité de données permet d'obtenir de meilleures prédictions que le corpus restreint à 28 jours. Bien que le questionnaire EDE-Q exige que les personnes participantes répondent aux questions en fonction de leurs comportements des derniers 28 jours, il est donc pertinent de considérer les historiques conversationnels des personnes sur une période non contrainte.

3.2.4.3 Expériences

Les expériences comparent les trois types de représentation : par fréquence, par plongement et par document représentatif le plus proche. Les configurations étudiées font varier les paramètres de filtrage de sujets. HDBSCAN et K-means sont comparés pour la représentation par plongement. Trois filtrages sont appliqués sur les sujets pour conserver : les 10% les plus pertinents (filtrage fort), les 25% les plus pertinents (filtrage élevé) et les 50% les plus pertinents (filtrage moyen). Une configuration sans filtrage est également testée. La proportion de sujets générés par BERTopic en fonction du filtrage est présentée dans le tableau 3.3.

BERTopic utilise le plongement all-MiniLM-L6-v2 [3] pour chacune des représentations, et HDBSCAN est utilisé comme algorithme de clustering pour les représentations de fréquence et de document représentatifs. Les modèles ont été entraînés et validés à partir de l'ensemble de données d'entraînement (124 historiques) et de validation (14 historiques) de la campagne eRisk 2022-2023. L'ensemble de test (18 historiques) est celui

3. https://huggingface.co/sentence-transformers/all-MiniLM-L6-v2

run eRisk	MAE	MZOE	MAE_macro	GED	RS	ECS	SCS	WCS
baseline tous 0	3,790	0,813	4,254	4,472	3,869	4,479	4,363	3,361
baseline tous 6	1,937	**0,551**	3,018	3,076	3,352	2,868	3,029	2,472
baseline moyenne	1,965	0,594	3,137	2,875	3,361	2,102	2,229	2,306
RELAI_0 (Maupomé *et al.*, 2024)	2,331	0,914	2,243	2,394	2,222	2,324	2,340	**1,812**
SCaLAR-NITK_0 (Prasanna *et al.*, 2024)	1,912	0,591	**1,643**	2,495	2,713	1,568	**1,536**	2,098
SCaLAR-NITK_1	1,980	0,664	1,972	2,570	2,562	**1,553**	1,960	2,066
SCaLAR-NITK_2	1,879	**0,568**	1,942	2,158	2,477	2,222	2,245	2,364
SCaLAR-NITK_3	1,932	0,586	1,868	**2,117**	2,430	2,046	2,242	2,407
SCaLAR-NITK_4	**1,874**	0,672	1,820	2,292	**2,140**	1,557	1,880	2,061
Représentations évaluées								
Fréquence	**1,775**	0,747	1,692	**1,967**	**2,024**	2,003	1,896	1,942
Plongement	1,840	0,767	1,727	2,126	2,103	2,217	2,067	2,112
Document représentatif	**1,694**	0,739	**1,620**	**2,061**	**2,013**	2,184	1,989	2,052

Table 3.4 – Baseline, meilleurs résultats obtenus à la tâche 3 de eRisk 2024 et résultats obtenus par nos trois représentations. Un score **bas** montre de meilleurs résultats pour l'ensemble des métriques.

de eRisk 2024.

3.2.4.4 Résultats

Les configurations ayant obtenu les meilleurs résultats sont :

- HDBSCAN et filtrage fort ($\sim$10%) pour la représentation par fréquence.
- K-means et filtrage moyen ($\sim$ 50%) pour la représentation par plongement.
- HDBSCAN et filtrage élevé ($\sim$ 25%) pour la représentation par document représentatif.

Les résultats de l'évaluation des trois représentations sont présentés dans le bas du tableau 3.4. La représentation par document représentatif obtient les meilleurs résultats pour 4 des 8 métriques (MZOE, MAE, MAE_macro et RS). La représentation par fréquence montre les meilleurs résultats pour chacune des sous-échelles et pour le score global.

Comparée avec les meilleurs scores obtenus à la tâche 3 d'eRisk 2024 présentés en haut du tableau 3.4, la représentation par document représentatif obtient des meilleurs scores sur 4 métriques (MAE, MAE_macro, GED et RS). Les résultats des représentations par fréquence et par plongement restent intéressants en comparaison des résultats obtenus à eRisk 2024.

Les résultats confortent dans le choix d'utiliser un modèle de sujet pour représenter les publications des historiques. L'approche par document représentatif confirme qu'une représentation plus riche des données conduit à de meilleurs résultats.
La représentation basée sur la fréquence produit également des résultats intéressants. Associée à un filtrage fort, elle semble avoir empêché l'extraction d'attributs non pertinents. Enfin, bien que la représentation par plongement présente des performances légèrement inférieures aux deux autres méthodes, ses résultats restent proches de l'état de l'art.

3.2.5 Conclusion

Ces travaux comparent trois méthodes de représentation d'historiques conversationnels pour prédire des comportement associés aux TA : représentation par fréquence, par plongement et par document représentatif. Un filtrage de sujets, effectué à partir des questions du EDE-Q, est également évalué dans la représentation d'historique. Cette représentation, une fois obtenue, permet d'entraîner un modèle neuronal pour la prédiction. Les résultats montrent que le filtrage permet d'améliorer la prédiction pour chacune des représentations. L'approche par document représentatif présente les meilleurs résultats, supérieurs à l'état de l'art de la tâche sur quatre métriques .

Nos travaux en cours se penchent sur le choix de l'algorithme de cluster permettant la création de sujets, les méthodes de filtrage possibles des listes de sujets et l'évaluation de techniques de réduction de dimension bien adaptées à la modélisation de sujet. Les travaux sont menés en collaboration avec des personnes expertes en santé mentale, ce qui permet d'explorer également l'ajout potentiel de ressources cliniques externes pour enrichir les outils développés.

Pour assurer la reproductibilité des expériences, le code source est disponible sous licence libre ici : `https://gitlab.labikb.ca/ikb-lab/articles/taln-2025`.

3.2.6 Remerciements

Ces travaux ont été réalisés grâce aux ressources de calcul mises à notre disposition par Calcul Québec et l'Alliance de recherche numérique du Canada, et grâce au soutien financier du Conseil de recherches en sciences naturelles et en génie du Canada (CRSNG) [MJ Meurs, CRSNG à la découverte #06487-2017] et de la Chaire de recherche du Québec sur la découvrabilité des contenus scientifiques en français [MJ Meurs, DOI 10.69777/358425]. Nous souhaitons également remercier Yassine Chahdi pour sa participation enthousiaste aux réflexions et aux analyses post-évaluation.

CHAPITRE 4
SÉLECTION ORDONNÉE DE PHRASES ASSOCIÉES AUX SYMPTÔMES DE LA DÉPRESSION PAR CLASSIFICATION ZÉRO-COUP

4.1 Contexte et références

Les travaux présentés dans cette section visent à évaluer une approche de classification zéro-coup pour la détection automatique de phrases associées à des symptômes de la dépression. Comme discuté dans la section 2.2, l'utilisation d'une approche zéro-coup a pour avantage de générer un résultat sans nécessiter un entraînement supervisé au préalable. Dans ce chapitre, nous examinons si cette approche permet d'identifier efficacement des comportements associés aux troubles de la dépression et dans quelle mesure ses performances se rapprochent de celles obtenues par des modèles affinés. Les données utilisées proviennent de la tâche 1 de la campagne d'évaluation eRisk 2024. Ce corpus contient environ 3,8 millions de phrases, dont 16100 d'entre elles ont été étiquetées manuellement. Ces phrases ont été utilisées par les personnes participantes pour l'entraînement et l'évaluation de leurs modèles. Dans le cadre de ces travaux, ces données ont servi à comparer les performances d'une approche zéro-coup à celles de modèles entraînés spécifiquement sur la même tâche. Bien que l'auteur de ce mémoire n'ait pas participé officiellement à la campagne, l'accès aux données a permis d'effectuer une analyse comparative indirecte avec les résultats publiés par les équipes participantes. Les résultats obtenus montrent que l'approche zéro-coup peut constituer une alternative compétitive, tout en éliminant le besoin d'un processus d'affinage coûteux en données annotées.

L'article correspondant, présenté en section 4.2, a été rédigé et présenté par l'auteur de ce mémoire lors de l'atelier *Traitement du Langage Médical à l'Époque des LLMs (MLP-LLM 2025)*, tenu dans le cadre de la conférence CORIA-TALN 2025. L'article est publié en libre accès (Ferstler *et al.*, 2025b), et le code source des travaux est également disponible publiquement sous licence libre GPL-v3, dans le dépôt GitLab (GitLab, 2025a). L'auteur du mémoire est à l'origine de la formulation de l'hypothèse, de la rédaction principale du manuscrit, ainsi que de l'intégralité de la conception et de l'implémentation du code source, avec la collaboration des co-autrices pour la rédaction, la relecture et la validation scientifique.

4.2 Article

Sélection ordonnée de phrases associées aux symptômes de la dépression par classification zéro-coup

Yves Ferstler, Catherine Lavoie, Marie-Jean Meurs

Université du Québec à Montréal

Résumé

Cet article présente une méthode pour extraire d'un corpus les phrases les plus pertinentes pour répondre à un questionnaire d'auto-évaluation. Un modèle de classification zéro-coup évalue la similarité entre les phrases et les réponses du questionnaire. Les résultats obtenus par ce modèle frugal sont prometteurs par comparaison avec ceux d'autres grands modèles de langue.

Abstract

Ordered selection of phrases associated with depression symptoms by zero-shot classification

This paper leverages a method for selecting in a corpus the most relevant sentences to answer a self-report questionnaire. A zero-shot classification model is used to evaluate the similarities between the sentences and the possible answers in the questionnaire. The results obtained by this frugal model are promissing when compared to those of large language models.

Mots-clés : Classification zéro-coup, Extraction de phrases, Dépression
Keywords : Zero-shot classification, Sentences extraction, Depression

4.2.1 Introduction

Selon l'Organisation mondiale de la Santé (World Health Organization, 2025), environ 5% des adultes dans le monde souffraient de dépression en 2023. Les symptômes de cette maladie sont liés par exemple à des sentiments de tristesse, d'irritation, à de la faible concentration et des changements du niveau d'appétit.

Plusieurs questionnaires de référence existent pour dépister et mesurer la sévérité de ces symptômes. Les questionnaires d'auto-évaluations sont généralement composées de questions à choix multiples, à échelles graduées ou de questions ouvertes axées sur les comportements associés aux différents symptômes. Ces outils offrent un support au corps médical dans le processus diagnostique et permettent l'auto-évaluation de l'évolution des symptômes. La détection automatique de ces comportements constitue un domaine de recherche important en traitement automatique du langage naturelle (TALN), notamment grâce à la disponibilité de corpora de messages et de discussions entre internautes issus de réseaux sociaux. Actuellement, l'état de l'art de la détection de symptômes de la dépression repose principalement sur l'affinage de grands modèles de langue (LLM) pré-entrainés, qui remplace souvent d'autres méthodes performantes. Parmi ces méthodes, la classification zéro-coup (ZSC) présente un intérêt particulier, car elle permet de classer des données sans nécessiter l'affinage d'un LLM ou l'entraînement préalable d'un modèle dédié, ce qui constitue un avantage quand les données sont limitées. Les travaux présentés dans cet article ont pour objectif de déterminer, dans un corpus de phrases, quelles sont celles qui sont pertinentes par rapport à des états émotionnels associés aux symptômes de dépression et de les classer par ordre de pertinence décroissante. L'approche utilise un modèle ZSC, dont les résultats sont comparés à d'autres approches basées sur l'affinage de LLM.

4.2.2 État-de-l'art

Parmi les travaux récents en TALN centrés sur la santé mentale, Karamat *et al.* (2024) proposent une approche hybride permettant la détection de plusieurs comportements associés à des symptômes de troubles de santé mentale sur les réseaux sociaux. En utilisant deux modèles préentraînés, MentalBERT et MelBERT (Ji *et al.*, 2021; Choi *et al.*, 2021), et en extrayant les caractéristiques avec un réseau de neurones convolutifs, leurs approches permettent une détection de comportements associés aux symptômes de la dépression, de l'anxiété, du trouble post-traumatique et du trouble de la personnalité limite. D'autres aspects de la santé mentale sont étudiés par Ta *et al.* (2022), comme la détection des symptômes associés à l'anxiété au sein de communautés en ligne en analysant les relations entre les personnes et les sujets abordés. Les méthodes pour détecter des comportements associés aux symptômes de la dépression sont variées. Lim *et al.* (2025); Chen *et al.* (2025) présentent des approches de conception de modèles capturant à la fois les caractéristiques audio d'un dialogue et sa transcription textuelle utilisant XLM-RoBERTa (Conneau *et al.*, 2019). De nouvelles approches utilisent des modèles accompagnés de base de connaissances pour améliorer la détection de comportements associés aux symptômes de la dépression. Yan *et al.* (2025) pro-

posent une méthode détectant des émotions négatives, basée sur les symptômes de dépression définis dans le *Diagnostic and Statistical Manual of Mental Disorders* (DSM-5) (Crocq *et al.*, 2015), et l'utilisation du transformeur BERT (Devlin *et al.*, 2018). Ces travaux présentent de meilleurs résultats que d'autres approches n'utilisant pas de base de connaissance. Su *et al.* (2025) utilisent également le DSM-5 comme base de connaissance. Dans le cadre de la campagne d'évaluation eRisk 2024 (Crestani *et al.*, 2022), les équipes participant à la tâche 1 ont travaillé sur l'évaluation de pertinence de phrases associées aux symptômes de la dépression. Parmi les résultats (Parapar *et al.*, 2024), les approches ayant obtenu les meilleurs scores sont basées sur l'affinage de LLM (Ang *et al.*, 2024; Bascuñana et Bedmar, 2024; Barachanou *et al.*, 2024). Cependant, d'autres approches utilisant des modèles de langue telle que la classification zéro-coup (ZSC) montrent des résultats prometteurs (Zhang *et al.*, 2020; Li *et al.*, 2017). Leow *et al.* (2025) présentent une comparaison entre l'utilisation de LLM ou de ZSC dans la détection de signes de la dépression. Les résultats indiquent que l'utilisation des LLM permet d'obtenir de meilleurs scores mais la méthode ZSC est plus frugale et aussi très performante.

4.2.3 Méthodologie

Pour déterminer quelles sont les phrases d'un corpus qui sont pertinentes par rapport à des symptômes de dépression, notre approche : (1) calcule, avec un modèle ZSC, un score de similitude entre phrases présentes dans le corpus et les énoncés pré-écrits d'un questionnaire d'auto-évaluation, (2) élimine les phrases ne contenant pas de pronoms personnels au singulier, (3) agrège les scores des phrases restantes en un score global qui permet de classer ces phrases par ordre de pertinence décroissante, (4) utilise la médiane des scores globaux pour définir un seuil au-delà duquel les phrases sont jugées pertinentes.

Ensemble de données et questionnaire de référence. Le corpus choisi dans les travaux présentés est le corpus d'évaluation de la tâche 1 de la campagne d'évaluation eRisk 2024 (Crestani *et al.*, 2022). Il contient 11,8k phrases, manuellement annotées, issues d'un historique conversationnel de personnes utilisatrices de réseaux sociaux. Chaque phrase est enregistrée avec ses phrases précédente et suivante pour conserver son contexte. Un exemple fictif est présenté en annexe, Figure 4.1. À partir des phrases manuellement annotées, deux ensembles de référence ont été créés : un contenant les phrases jugées pertinentes par consensus des personnes annotatrices et l'autre par la majorité de ces personnes. Le questionnaire d'auto-évaluation utilisé pour évaluer la pertinence entre les phrases et ses items est l'inventaire de dépression de Beck (Beck *et al.*, 1996) 2ème édition (BDI-II). Ce questionnaire est composé de 21 items associés à différents états

émotionnels liés à des symptômes de la dépression. Chacun de ces items est qualifié par plusieurs énoncés pré-écrits reflétant la gravité de l'état considéré. Des exemples et leur traduction libre en français sont fournis en annexe (table 4.3).

Modèle d'évaluation de pertinence. Les directives d'annotation manuelle des phrases indiquent qu'une phrase pertinente doit donner de l'information sur l'état de la personne en rapport avec l'item d'auto-évaluation. Ainsi, une phrase exprimant du bien-être, de la joie, de l'appétit, ou d'autres manifestations positives est aussi pertinente qu'une phrase présentant des émotions ou des comportements associés aux symptômes de la dépression. Afin d'identifier les phrases pertinentes, le modèle ZSC doit les comparer à chacun des énoncés pré-écrits. Avant de calculer les similarités, un filtrage inspiré de Ang *et al.* (2024) est appliqué pour supprimer les phrases sans pronoms personnels au singulier. Ne se rapportant pas à la personne autrice de la publication, elles sont jugées non pertinentes pour le questionnaire. Le modèle ZSC choisi est bart-large-mnli (Hugging Face, 2025), pour sa légèreté et ses performances à l'état de l'art. Les scores ZSC des énoncés pré-écrits sont combinés pour produire deux versions de score global, l'une étant la somme des scores et l'autre le maximum. Le score global permet de classer les phrases et d'identifier celles qui sont les plus pertinentes pour chaque item. Un exemple de scores est présenté dans la table 4.1. La phrase « je vais bien, mais je n'ai juste pas envie de faire quelque chose » a été évaluée par le modèle ZSC sur les trois items présents dans la table 4.3 en annexe. Les résultats montrent que, pour les items 1 (tristesse) et 12 (perte d'intérêt), les étiquettes avec le score le plus haut sont « *I do not feel sad.* » et « *I am less interested in other people or things than before.* », ce qui est cohérent pour la phrase évaluée. Pour l'item 7 (dégoût de soi), les étiquettes obtiennent des scores plus faibles, ce qui est également cohérent, dans la mesure où aucun sentiment de dégoût de soi n'est exprimé.

Phrase	Je vais bien, mais je n'ai juste pas envie de faire quelque chose					
Scores	1er	2ème	3ème	4ème	somme	maximum
1.	r1 : 0,951	r2 : 0,000	r3 : 0,000	r4 : 0,000	0,951	0,951
7.	r1 : 0,114	r2 : 0,013	r4 : 0,000	r3 : 0,000	0,127	0,114
12.	r2 : 0,876	r3 : 0,354	r1 : 0,300	r4 : 0,038	1,568	0,876

Table 4.1 – Exemple de résultat obtenu par un modèle ZSC entre une phrase et 3 items du BDI-II caractérisés par les 4 réponses pré-écrites (r1 à r4)

4.2.4 Expériences et résultats

Le corpus de référence, utilisé pour l'évaluation, a été construit à partir des soumissions des participants à la tâche 1 de eRisk 2024 (Parapar *et al.*, 2024). N'ayant pas accès à toutes les soumissions, il est impossible de reproduire les mêmes conditions d'évaluation. Nous avons donc évalué nos résultats en les comparant aux résultats fournis par les organisateurs pour les phrases du corpus de référence. Cette approche permet d'évaluer les capacités de classification de pertinence du modèle ZSC en utilisant les métriques classiques de précision, rappel et F-mesure. L'ordonnancement des phrases selon leur pertinence est mesuré par le gain cumulé normalisé actualisé (NDCG) qui, pour chaque item du BDI-II, additionne les scores de pertinence des phrases selon leur position dans le classement généré par le modèle, en accordant plus d'importance aux phrases placées en haut et normalise ensuite ce score en le divisant par le score idéal, obtenu avec un classement parfait.

Expériences. Les phrases du corpus de référence (manuellement étiquetées en deux versions : *majorité* et *consensus* entre annotations) sont filtrées puis évaluées par le modèle ZSC pour produire les deux versions de score global (somme et maximum). Un seuil est alors appliqué au score global des phrases pour distinguer celles qui sont pertinentes pour un item de celles qui ne le sont pas. Un seuil spécifique à chaque item est calculé en prenant la médiane des scores obtenus. Le code source est disponible sous licence libre ici.

Résultats. Les résultats de classification de phrases obtenus par le modèle ZSC, selon les calculs de scores, sont présentés dans la table 4.2. Les approches proposées donnent des résultats intéressants, en particulier sur le corpus de référence *majorité*. Les résultats montrent que le calcul du score global par maximum est plus performant que le calcul utilisant la somme. Ainsi les phrases sélectionnées par maximum pourraient potentiellement permettre de choisir une réponse parmi celles pré-écrites pour les items du BDI-II. Il sera donc intéressant d'évaluer la capacité du modèle à prédire directement les réponses au BDI-II à partir des phrases qu'il sélectionne dans les écrits d'une personne. Le modèle ZSC avec les deux options de calcul de score obtient un NDCG de 0,7 qui correspond à celui des systèmes présentés à eRisk 2024, confirmant la capacité d'ordonnancement de notre approche.

4.2.5 Conclusion

Ces travaux présentent une méthode basée sur un modèle ZSC pour identifier et classer la pertinence de phrases en fonction des items du BDI II (Beck *et al.*, 1996), questionnaire d'auto-évaluation d'états émotionnels spécifiques associées aux symptômes de la dépression qui est un des principaux outils utilisés pour le dépistage et la mesure de sévérité de ces symptômes chez les personnes adolescentes et adultes. Les résul-

tats obtenus grâce au score maximum attribué aux phrases par le modèle sont prometteurs. Ils confirment qu'affiner un LLM n'est pas toujours indispensable quand un modèle plus frugal comme ZSC peut être mis en oeuvre.

4.2.6 Remerciements

Ces travaux ont été réalisés grâce aux ressources de calcul mises à notre disposition par Calcul Québec et l'Alliance de recherche numérique du Canada, et grâce au soutien financier du Conseil de recherches en sciences naturelles et en génie du Canada (CRSNG) [MJ Meurs, CRSNG à la découverte #06487-2017] et de la Chaire de recherche du Québec sur la découvrabilité des contenus scientifiques en français [MJ Meurs, DOI 10.69777/358425]. Nous souhaitons également remercier Yassine Chahdi pour sa participation enthousiaste aux réflexions et aux analyses post-évaluation.

4.2.7 Annexe

Figure 4.1 – **Exemple fictif d'extrait du corpus de la tâche 1 d'eRisk 2024**

```
<DOC>
    <DOCNO>5001_0_1</DOCNO>
    <PRE></PRE>
    <TEXT>Je vais bien, mais je n'ai juste pas envie de faire quelque chose</TEXT>
    <POST>Je pense que je vais passer ma journée chez moi</POST>
</DOC>
<DOC>
    <DOCNO>5001_0_2</DOCNO>
    <PRE>Je vais bien, mais je n'ai juste pas envie de faire quelque chose</PRE>
    <TEXT>Je pense que je vais passer ma journée chez moi</TEXT>
    <POST>Arrête de me dire ça, tu vas me faire culpabiliser</POST>
</DOC>
```

Évaluation sur le corpus de référence ***majorité***

Score global	Précision	Rappel	F-mesure
ZSC-somme	0,7883	0,6644	0,7210
ZSC-maximum	**0,8151**	0,6858	**0,7449**
Meilleurs scores par question ordonnée par F-mesure			
ZSC-maximum-question-14 *(Dévalorisation)*	0,9756	0,7751	0,8639
ZSC-somme-question-14 *(Dévalorisation)*	0,9707	0,7713	0,8596
Pires scores par question ordonnée par F-mesure			
ZSC-somme-question-13 *(Indécision)*	0,5847	0,5487	0,5661
ZSC-maximum-question-13 *(Indécision)*	0,5862	0,5519	0,5685

Évaluation sur le corpus de référence ***consensus***

Score global	Précision	Rappel	F-mesure
ZSC-somme	0,5631	0,7212	0,6300
ZSC-maximum	0,5862	**0,7494**	0,6600
Meilleurs scores par question ordonnés par F-mesure			
ZSC-maximum-question-14 *(Dévalorisation)*	0,9317	0,8603	0,8946
ZSC-somme-question-14 *(Dévalorisation)*	0,9121	0,8423	0,8758
Pires scores par question ordonnée par F-mesure			
ZSC-somme-question-2 *(Pessimisme)*	0,2846	0,6557	0,3970
ZSC-somme-question-13 *(Indécision)*	0,3356	0,5914	0,4282

Table 4.2 – Résultats de l'approche ZSC (somme et maximum) appliquée au corpus de référence *majorité* et *consensus* de la tâche 1 de eRisk 2024

Table 4.3 – **Exemples d'items du BDI-II avec leurs énoncés associés**

Réponses Pré-écrites
1. *Sadness* – Tristesse
r1. *I do not feel sad.* – Je ne me sens pas triste.
r2. *I feel sad much of the time.* – Je me sens triste la plupart du temps.
r3. *I am sad all the time.* – Je suis toujours triste.
r4. *I am so sad all the time.* – Je suis tellement triste tout le temps.
7. *Self-Dislike* – Dégoût de soi
r1. *I feel the same about myself as ever.* – Je me sens toujours aussi bien.
r2. *I have lost confidence in myself.* – J'ai perdu confiance en moi.
r3. *I am disappointed in myself.* – Je suis déçu par moi-même.
r4. *I dislike myself.* – Je ne m'aime pas.
12. *Loss of Interest* – Perte d'intérêt
r1. *I have not lost interest in other people or activities.* – Je ne me suis pas désintéressé des autres personnes ou des activités.
r2. *I am less interested in other people or things than before.* – Je m'intéresse moins aux autres personnes ou aux choses qu'auparavant.
r3. *I have lost most of my interest in other people or things.* – J'ai perdu la plupart de mon intérêt pour les autres personnes ou les choses.
r4. *It's hard to get interested in anything.* – Il est difficile de s'intéresser à quoi que ce soit.

CONCLUSION

Les récents changements dans le TALN ont permis l'amélioration des performances de nombreuses tâches de compréhension de texte. Cela a conduit à des avancées majeures dans plusieurs domaines et parfois même à la découverte de nouvelles applications. Parmi les domaines profitant du TALN, celui de la santé mentale est l'un de ceux pour lesquels c'est le plus utile.

L'objectif, tout au long de ces travaux, était de montrer comment les modèles d'apprentissage automatique utilisés en TALN pouvaient contribuer à la détection précoce de risques de troubles en santé mentale.

Pour cela, deux approches basées sur l'architecture transformeur ont été évaluées dans la tâche d'analyse de texte et de sentiments. La première approche repose sur BERTopic et l'analyse de sujets conversationnels à partir d'un modèle de type encodeur. La seconde utilise BART et le calcul de proximité syntaxique et sémantique à partir d'un modèle de type encodeur-décodeur.

Le Chapitre 3 a présenté la première approche, qui se focalise sur la détection de signes de troubles alimentaires à partir de conversations issues de réseaux sociaux. BERTopic a été utilisé afin de réduire les conversations en représentations vectorielles et de permettre à un algorithme de regroupement, après réduction de dimension, d'identifier plusieurs groupes de sujets de discussion. Cette méthode a permis de modéliser des groupes de sujets de discussion propres à chaque personne utilisatrice. Ces groupes peuvent ensuite être utilisés pour entraîner un réseau de neurones à prédire des signes de troubles alimentaires. La modélisation de discussions en groupes de sujets permet aussi l'analyse de ces sujets ainsi que la possibilité d'en filtrer certains. Les performances obtenues pour la détection de signes de troubles alimentaires sont meilleures que celles des approches précédentes proposées lors de la campagne d'évaluation eRisk 2024. Les travaux ont également montré que l'utilisation de l'algorithme de regroupement K-moyennes dans ce contexte pouvait améliorer les résultats, tout comme l'usage de techniques de filtrage de sujets. Le Chapitre 3 a ainsi montré que l'utilisation de modèles encodeurs pour la modélisation de sujets, associée à l'analyse du corpus et des algorithmes utilisés, pouvait permettre la détection précoce de risques de troubles en santé mentale.

Le Chapitre 4 a présenté la seconde approche, visant à prédire des comportements associés aux troubles de la dépression. L'approche proposée repose sur BART et sur un calcul de similarité obtenu par une méthode

de classification en zéro-coup. L'objectif était de calculer un score pour chaque phrase issue de réseaux sociaux et de les associer, de manière ordonnée, à des comportements liés aux troubles de la dépression. Les comportements auxquels les phrases devaient être associées proviennent de l'inventaire de dépression de Beck. Pour chaque phrase, BART a été utilisé pour calculer un score de similarité entre une phrase issue du corpus et les réponses pré-écrites issues de l'inventaire. Un score final, obtenu par maximum ou par somme des scores, a ensuite été calculé pour chaque phrase et chaque catégorie de l'inventaire. Les résultats ont été comparés à ceux d'autres approches lors de la campagne d'évaluation eRisk 2024, dont la majorité des modèles ont été affinés lors de cette tâche. Malgré l'absence d'affinage dans l'approche en zéro-coup, les résultats ont montré que cette méthode permettait d'obtenir des performances encourageantes pour la détection de signes de troubles dépressifs tout en réduisant les besoins en calcul et en données liés à l'affinage.

Les approches proposées lors de ces travaux ont montré que l'utilisation de modèles d'apprentissage automatique issus du TALN pour l'analyse des sentiments dans le but de détecter de manière précoce des risques de troubles en santé mentale était pertinente. Toutefois, l'analyse du corpus et des algorithmes employés a été déterminante pour améliorer leurs performances. L'utilisation d'outils propres au domaine de la santé mentale, comme l'inventaire de Beck ou l'EDE Q, a également contribué à la détection de signes de troubles.

Enfin, les prédictions faites par les approches présentées, ainsi que d'autres techniques de détection automatique, ne permettent pas une réelle identification de signe de troubles sans l'expertise de personnes formées dans ce domaine.

BIBLIOGRAPHIE

Aardoom, J. J., Dingemans, A. E., Slof Op't Landt, M. C. et Van Furth, E. F. (2012a). Norms and discriminative validity of the Eating Disorder Examination Questionnaire (EDE-Q). *Eating Behaviors*, *13*(4), 305–309. https://doi.org/10.1016/j.eatbeh.2012.09.002

Aardoom, J. J., Dingemans, A. E., Slof Op't Landt, M. C. et Van Furth, E. F. (2012b). Norms and discriminative validity of the Eating Disorder Examination Questionnaire (EDE-Q). *Eating Behaviors*, *13*(4), 305–309. https://doi.org/https://doi.org/10.1016/j.eatbeh.2012.09.002

Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F. L., Almeida, D., Altenschmidt, J., Altman, S., Anadkat, S. *et al.* (2023). Gpt-4 technical report. *arXiv preprint arXiv :2303.08774*.

Ang, B. H., Gollapalli, S. D. et Ng, S.-K. (2024). NUS-IDS@ eRisk2024 : ranking sentences for depression symptoms using early maladaptive schemas and ensembles. Dans *Experimental IR Meets Multilinguality, Multimodality, and Interaction : 15th International Conference of the CLEF Association, CLEF 2024, Grenoble, France, September 9–12, 2024, Proceedings, Part II*, (p. 782–793).

Aye, L. M., Tan, M. M., Schaefer, A., Thurairajasingam, S., Geldsetzer, P., Soon, L. K., Reininghaus, U., Bärnighausen, T. et Su, T. T. (2024). Self-help digital mental health intervention in improving burnout and mental health outcomes among healthcare workers : A narrative review. *DIGITAL HEALTH*, *10*, 20552076241278313. https://doi.org/10.1177/20552076241278313

Ba, J. L., Kiros, J. R. et Hinton, G. E. (2016). Layer normalization. *arXiv preprint arXiv :1607.06450*.

Bahdanau, D., Cho, K. et Bengio, Y. (2014). Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv :1409.0473*.

Barachanou, A., Tsalakanidou, F. et Papadopoulos, S. (2024). REBECCA at eRisk 2024 : search for symptoms of depression using sentence embeddings and prompt-based filtering. Dans *Experimental IR Meets Multilinguality, Multimodality, and Interaction : 15th International Conference of the CLEF Association, CLEF 2024, Grenoble, France, September 9–12, 2024, Proceedings, Part II*, (p. 794–802).

Bascuñana, A. et Bedmar, I. S. (2024). APB-UC3M at eRisk 2024 : natural language processing and deep learning for the early detection of mental disorders. Dans *Experimental IR Meets Multilinguality, Multimodality, and Interaction : 15th International Conference of the CLEF Association, CLEF 2024, Grenoble, France, September 9–12, 2024, Proceedings, Part II*, (p. 871–880).

Beck, A. T., Steer, R. A. et Brown, G. K. (1996). Beck Depression Inventory (BDI-II). *Psychological assessment*, *10*.

Bellman, R. (1966). Dynamic programming. *science*, *153*(3731), 34–37.

Benitez-Andrades, J. A., Alija-Perez, J. M., Garcia-Rodriguez, I., Benavides, C., Alaiz-Moreton, H., Vargas, R. P. et Garcia-Ordas, M. T. (2021). BERT Model-Based Approach For Detecting Categories of Tweets in the Field of Eating Disorders (ED). Dans *2021 IEEE 34th International Symposium on Computer-Based Medical Systems (CBMS)*, (p. 586–590). IEEE. https://doi.org/10.1109/CBMS52027.2021.00105

Blei, D. M., Ng, A. Y. et Jordan, M. I. (2003). Latent dirichlet allocation. *Journal of machine Learning research*, *3*(Jan), 993–1022.

Bokolo, B. G. et Liu, Q. (2024). Advanced comparative analysis of machine learning and transformer models for depression and suicide detection in social media texts. *Electronics*, *13*(20), 3980.

Campello, R. J., Moulavi, D. et Sander, J. (2013). Density-based clustering based on hierarchical density estimates. Dans *Pacific-Asia conference on knowledge discovery and data mining*, (p. 160–172). Springer.

Chen, Z., Wang, D., Lou, L., Zhang, S., Zhao, X., Jiang, S., Yu, J. et Xiao, J. (2025). Text-guided multimodal depression detection via cross-modal feature reconstruction and decomposition. *Information Fusion*, *117*, 102861. https://doi.org/https://doi.org/10.1016/j.inffus.2024.102861

Chiong, R., Budhi, G. S., Dhakal, S. et Chiong, F. (2021). A textual-based featuring approach for depression detection using machine learning classifiers and social media texts. *Computers in Biology and Medicine*, *135*, 104499. https://doi.org/https://doi.org/10.1016/j.compbiomed.2021.104499

Choi, M., Lee, S., Choi, E., Park, H., Lee, J., Lee, D. et Lee, J. (2021). MelBERT : Metaphor Detection via Contextualized Late Interaction using Metaphorical Identification Theories. https://arxiv.org/abs/2104.13615

CIHI (2024). Institut canadien d'information sur la santé : Les Canadiens signalent que leurs besoins en matière de soins de santé mentale augmentent, tout comme les obstacles qu'ils rencontrent pour accéder à ces soins. https://www.cihi.ca/fr/les-canadiens-signalent-que-leurs-besoins-en-matiere-de-soins-de-sante-mentale-augmentent-tout [Accessed : (15/08/2025)].

Conneau, A., Khandelwal, K., Goyal, N., Chaudhary, V., Wenzek, G., Guzmán, F., Grave, E., Ott, M., Zettlemoyer, L. et Stoyanov, V. (2019). Unsupervised Cross-lingual Representation Learning at Scale. *CoRR*, *abs/1911.02116*. http://arxiv.org/abs/1911.02116

Crestani, F., Losada, D. et Parapar, J. (2022). *Early Detection of Mental Health Disorders by Social Media Monitoring : The First Five Years of the eRisk Project*. Studies in Computational Intelligence. Springer International Publishing. https://doi.org/10.1007/978-3-031-04431-1

Crocq, M.-A., Guelfi, J. D., Association, A. P. et Force, A. P. A. D.-. T. (2015). *DSM-5 : manuel diagnostique et statistique des troubles mentaux (5e édition)*. Elsevier Masson. https://ebookcentral.proquest.com/lib/umontreal-ebooks/detail.action?docID=4337396

Devlin, J., Chang, M., Lee, K. et Toutanova, K. (2018). BERT : Pre-training of Deep Bidirectional Transformers for Language Understanding. *CoRR*, *abs/1810.04805*. http://arxiv.org/abs/1810.04805

Devlin, J., Chang, M.-W., Lee, K. et Toutanova, K. (2019a). Bert : Pre-training of deep bidirectional transformers for language understanding. Dans *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics : human language technologies (long and short papers)*, Vol. 1, (p. 4171–4186).

Devlin, J., Chang, M.-W., Lee, K. et Toutanova, K. (2019b). BERT : Pre-training of Deep Bidirectional Transformers for Language Understanding. https://arxiv.org/abs/1810.04805

Dictionnaire de l'Académie française (2025). Dictionnaire de l'Académie française, 9e édition (actuelle). https://www.dictionnaire-academie.fr/article/QDL056 [Accessed : (25/11/2025)].

Doll, C. M., Michel, C., Rosen, M., Osman, N., Schimmelmann, B. G. et Schultze-Lutter, F. (2021). Predictors of help-seeking behaviour in people with mental health problems : a 3-year prospective community study. *BMC Psychiatry*, *21*(1), 432. https://doi.org/10.1186/s12888-021-03435-4

Ester, M., Kriegel, H.-P., Sander, J. et Xu, X. (1996). A density-based algorithm for discovering clusters in large spatial databases with noise. Dans *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining*, KDD'96, (p. 226–231). AAAI Press.

Fairburn, C. et Beglin, S. (2008). Eating Disorder Examination Questionnaire (EDE-Q 6.0). *Cognitive behavior therapy and eating disorders*, 309–313. https://doi.org/10.1037/t03974-000

Ferstler, Y., Lavoie, C. et Meurs, M.-J. (2025a). Détecter des comportements associés aux troubles alimentaires par l'analyse automatique des publications textuelles en ligne. Dans *Actes de CORIA-TALN-RJCRI-RECITAL 2025. Actes des 32ème Conférence sur le Traitement Automatique des Langues Naturelles (TALN), volume 1 : articles scientifiques originaux*, Vol. 1, (p. 206–217). Association pour le Traitement Automatique des Langues. https://talnarchives.atala.org/TALN/TALN-2025/136.pdf

Ferstler, Y., Lavoie, C. et Meurs, M.-J. (2025b). Sélection ordonnée de phrases associées aux symptômes de la dépression par classification zéro-coup. Dans *Actes de CORIA-TALN-RJCRI-RECITAL 2025. Actes de l'atelier Traitement du langage médical à l' époque des LLMs 2025 (MLP-LLM)*, (p. 42–48). Association pour le Traitement Automatique des Langues. https://talnarchives.atala.org/ateliers/2025/MLPLLM/203.pdf

Forgy, E. W. (1965). Cluster analysis of multivariate data : efficiency versus interpretability of classifications. *biometrics*, *21*, 768–769.

Fred, A. et Agarap, M. (2018). Deep learning using rectified linear units (relu). *arXiv preprint arXiv :1803.08375*, 1–6.

GitLab (2025a). Conférence MLP-LLM 2025. https://gitlab.labikb.ca/ikb-lab/articles/mlp-llm-2025 [Accessed : (22/12/2025)].

GitLab (2025b). Conférence TALN 2025. https://gitlab.labikb.ca/ikb-lab/articles/taln-2025 [Accessed : (22/12/2025)].

Grootendorst, M. (2022). BERTopic : Neural topic modeling with a class-based TF-IDF procedure. https://doi.org/https://doi.org/10.48550/arXiv.2203.05794

Han, J. et Moraga, C. (1995). The influence of the sigmoid function parameters on the speed of backpropagation learning. Dans J. Mira et F. Sandoval (dir.), *From Natural to Artificial Neural Computation*, (p. 195–201). Springer Berlin Heidelberg.

Henderson, C., Evans-Lacko, S. et Thornicroft, G. (2013). Mental Illness Stigma, Help Seeking, and Public Health Programs. *American Journal of Public Health*, *103*(5), 777–780. https://doi.org/10.2105/AJPH.2012.301056

Hugging Face (2025). facebook/bart-large-mnli. https://huggingface.co/facebook/bart-large-mnli [Accessed : (03/05/2025)].

Institute for Eating Disorders (2024-12-31). Eating Disorder Examination Questionnaire EDE-Q. https://insideoutinstitute.org.au/resource-library/eating-disorder-examination-questionnaire-ede-q. Accessed : 2024-12-31.

Ji, S., Zhang, T., Ansari, L., Fu, J., Tiwari, P. et Cambria, E. (2021). MentalBERT : Publicly Available Pretrained Language Models for Mental Healthcare. https://arxiv.org/abs/2110.15621

Karamat, A., Imran, M., Yaseen, M. U., Bukhsh, R., Aslam, S. et Ashraf, N. (2024). A Hybrid Transformer Architecture for Multiclass Mental Illness Prediction using Social Media Text. *IEEE Access*, 1–1. https://doi.org/10.1109/ACCESS.2024.3519308

Knapp, M. et Wong, G. (2020). Economics and mental health : the current scenario. *World Psychiatry*, *19*(1), 3–14. https://doi.org/10.1002/wps.20692

Koskenniemi, K. (1984). A general computational model for word-form recognition and production. Dans *10th International Conference on Computational Linguistics and 22nd Annual Meeting of the Association for Computational Linguistics*, (p. 178–181).

Lan, Z., Chen, M., Goodman, S., Gimpel, K., Sharma, P. et Soricut, R. (2020). ALBERT : A Lite BERT for Self-supervised Learning of Language Representations. https://arxiv.org/abs/1909.11942

Larousse (2025). LAROUSSE Dictionaire De Français. https://www.larousse.fr/dictionnaires/francais-monolingue [Accessed : (25/11/2025)].

Le, H., Vial, L., Frej, J., Segonne, V., Coavoux, M., Lecouteux, B., Allauzen, A., Crabbé, B., Besacier, L. et Schwab, D. (2020). FlauBERT : Unsupervised Language Model Pre-training for French. https://arxiv.org/abs/1912.05372

Le Robert (2025). Dictionnaire Le Petit Robert de la langue française. https://www.lerobert.com/dictionnaires/francais/langue/dictionnaire-le-petit-robert-de-la-langue-francaise-edition-abonnes-3133099010272.html [Accessed : (25/11/2025)].

Leow, J. J. D., Chua, H. N., Jasser, M. B., Issa, B. et Wong, R. T. (2025). Comparison of Depression Detection Between LLMs and Zero-Shot Learning Using DAD Dataset. Dans *2025 21st IEEE International Colloquium on Signal Processing & Its Applications (CSPA)*, Vol. 21, (p. 295–300). https://doi.org/10.1109/CSPA64953.2025.10933098

Lewis, M., Liu, Y., Goyal, N., Ghazvininejad, M., Mohamed, A., Levy, O., Stoyanov, V. et Zettlemoyer, L. (2020). BART : Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. Dans *Proceedings of the 58th annual meeting of the association for computational linguistics*, (p. 7871–7880).

Li, X., Fang, M. et Wu, J. (2017). Zero-shot classification by transferring knowledge and preserving data structure. *Neurocomputing*, *238*, 76–83. https://doi.org/https://doi.org/10.1016/j.neucom.2017.01.038

Lim, E., Jhon, M., Kim, J.-W., Kim, S.-H., Kim, S. et Yang, H.-J. (2025). A lightweight approach based on cross-modality for depression detection. *Computers in Biology and Medicine*, *186*, 109618. https://doi.org/https://doi.org/10.1016/j.compbiomed.2024.109618

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L. et Stoyanov, V. (2019). RoBERTa : A Robustly Optimized BERT Pretraining Approach. https://arxiv.org/abs/1907.11692

Liu, Y., Zhang, K., Li, Y., Yan, Z., Gao, C., Chen, R., Yuan, Z., Huang, Y., Sun, H., Gao, J. *et al.* (2024). Sora : A review on background, technology, limitations, and opportunities of large vision models. *arXiv preprint arXiv :2402.17177*.

Luong, M.-T., Pham, H. et Manning, C. D. (2015). Effective approaches to attention-based neural machine translation. *arXiv preprint arXiv :1508.04025*.

Maćkiewicz, A. et Ratajczak, W. (1993). Principal components analysis (PCA). *Computers & Geosciences*, *19*(3), 303–342.

MacQueen, J. (1967). Multivariate observations. Dans *Proceedings ofthe 5th Berkeley Symposium on Mathematical Statisticsand Probability*, Vol. 1, (p. 281–297).

Malzer, C. et Baum, M. (2020). A Hybrid Approach To Hierarchical Density-based Cluster Selection. Dans *2020 IEEE International Conference on Multisensor Fusion and Integration for Intelligent Systems (MFI)*, (p. 223–228). IEEE. https://doi.org/10.1109/mfi49285.2020.9235263

Manning, C. D., Clark, K., Hewitt, J., Khandelwal, U. et Levy, O. (2020). Emergent linguistic structure in artificial neural networks trained by self-supervision. *Proceedings of the National Academy of Sciences*, *117*(48), 30046–30054.

Martin, L., Muller, B., Ortiz Suárez, P. J., Dupont, Y., Romary, L., de la Clergerie, É., Seddah, D. et Sagot, B. (2020). CamemBERT : a Tasty French Language Model. Dans D. Jurafsky, J. Chai, N. Schluter, et J. Tetreault (dir.), *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, (p. 7203–7219). Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.acl-main.645

Maupomé, D., Ferstler, Y., Mosser, S. et Meurs, M.-J. (2024). Automatically finding evidence, predicting answers in mental health self-report questionnaires. Dans *eRisk 2024 Workshop at the 15th International Conference of the CLEF Association, CLEF 2024, Grenoble, France, September 9–12, 2024*, (p. 841–850).

McGrath, J. J., Al-Hamzawi, A., Alonso, J., Altwaijri, Y., Andrade, L. H., Bromet, E. J., Bruffaerts, R., De Almeida, J. M. C., Chardoul, S., Chiu, W. T., Degenhardt, L., Demler, O. V., Ferry, F., Gureje, O., Haro, J. M., Karam, E. G., Karam, G., Khaled, S. M., Kovess-Masfety, V., … Zaslavsky, A. M. (2023). Age of onset and cumulative risk of mental disorders : a cross-national analysis of population surveys from 29 countries. *The Lancet Psychiatry*, *10*(9), 668–681. https://doi.org/https://doi.org/10.1016/S2215-0366(23)00193-1

McInnes, L., Healy, J. et Melville, J. (2018). Umap : Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv :1802.03426*.

Mental Health Commission of Canada (2012). Making the case for investing in mental health in canada. [Accessed : (15/08/2025)], https://www.mentalhealthcommission.ca/wp-content/uploads/drupal/MHStrategy_CaseForInvestment_ENG_0_1.pdf

Mond, J., Hay, P., Rodgers, B., Owen, C. et Beumont, P. (2004a). Validity of the Eating Disorder Examination Questionnaire (EDE-Q) in screening for eating disorders in community samples.

Behaviour Research and Therapy, *42*(5), 551–567. https://doi.org/https://doi.org/10.1016/S0005-7967(03)00161-X

Mond, J. M., Hay, P. J., Rodgers, B., Owen, C. et Beumont, P. J. (2004b). Validity of the Eating Disorder Examination Questionnaire (EDE-Q) in screening for eating disorders in community samples. *Behaviour research and therapy*, *42*(5), 551–567.

Negash, A., Khan, M. A., Medhin, G., Wondimagegn, D. et Araya, M. (2020). Mental distress, perceived need, and barriers to receive professional mental health care among university students in Ethiopia. *BMC Psychiatry*, *20*(1), 187. https://doi.org/10.1186/s12888-020-02602-3

NIH (2022). Mental illness statistics. https://www.nimh.nih.gov/health/statistics/mental-illness [Accessed : (15/08/2025)].

Odgers, C. L. et Jensen, M. R. (2020). Annual Research Review : Adolescent mental health in the digital age : facts, fears, and future directions. *Journal of Child Psychology and Psychiatry*, *61*(3), 336–348. https://doi.org/10.1111/jcpp.13190

OMS (2019). Oms. https://www.who.int/health-topics/mental-health#tab=tab_2 [Accessed : (15/08/2025)].

OMS (2022). Oms. https://www.who.int/news-room/fact-sheets/detail/depression [Accessed : (15/08/2025)].

Parapar, J., Martín-Rodilla, P., Losada, D. E. et Crestani, F. (2024). Overview of eRisk 2024 : Early Risk Prediction on the Internet. Dans *Experimental IR Meets Multilinguality, Multimodality, and Interaction : 15th International Conference of the CLEF Association, CLEF 2024, Grenoble, France, September 9–12, 2024, Proceedings, Part II*, (p. 73–92). Springer-Verlag. https://doi.org/10.1007/978-3-031-71908-0_4

Prasanna, S., Gulati, A. S., Karmakar, S., Hiranmayi, M. Y. et Madasamy, A. K. (2024). Measuring the severity of the signs of eating disorders using machine learning techniques. Dans *Experimental IR Meets Multilinguality, Multimodality, and Interaction : 15th International Conference of the CLEF Association, CLEF 2024, Grenoble, France, September 9–12, 2024, Proceedings, Part II*, (p. 881–887).

Radford, A., Kim, J. W., Xu, T., Brockman, G., McLeavey, C. et Sutskever, I. (2023). Robust speech recognition via large-scale weak supervision. Dans *International conference on machine learning*, (p. 28492–28518). PMLR.

Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W. et Liu, P. J. (2020). Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of machine learning research*, *21*(140), 1–67.

Ratnaparkhi, A. (1996). A maximum entropy model for part-of-speech tagging. Dans *Conference on empirical methods in natural language processing*.

Reimers, N. et Gurevych, I. (2019). Sentence-BERT : Sentence Embeddings using Siamese BERT-Networks. Dans *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics. https://arxiv.org/abs/1908.10084

Rosenblatt, F. (1958). The perceptron : A probabilistic model for information storage and organization in the brain. *Psychological Review*, *65*(6), 386–408. https://doi.org/10.1037/h0042519

Rujas, M., Merino-Barbancho, B., Arroyo, P. et Fico, G. (2023). Development of a Natural Language Processing-Based System for Characterizing Eating Disorders. Dans *Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2023)*. CEUR-WS. https://ceur-ws.org/Vol-3496/mentalriskes-paper13.pdf

Rumelhart, D. E., Hinton, G. E. et Williams, R. J. (1986). Learning representations by back-propagating errors. *Nature*, *323*(6088), 533–536. https://doi.org/10.1038/323533a0

Salton, G. et Buckley, C. (1988). Term-weighting approaches in automatic text retrieval. *Information processing & management*, *24*(5), 513–523.

Sanh, V., Debut, L., Chaumond, J. et Wolf, T. (2020). DistilBERT, a distilled version of BERT : smaller, faster, cheaper and lighter. https://arxiv.org/abs/1910.01108

Sennrich, R., Haddow, B. et Birch, A. (2016). Neural Machine Translation of Rare Words with Subword Units. arXiv :1508.07909, https://doi.org/10.48550/arXiv.1508.07909

Sihab-Us-Sakib, S., Rahman, M. R., Forhad, M. S. A. et Aziz, M. A. (2024). Cyberbullying detection of resource constrained language from social media using transformer-based approach. *Natural Language Processing Journal*, *9*, 100104. https://doi.org/https://doi.org/10.1016/j.nlp.2024.100104

Sowles, S. J., McLeary, M., Optican, A., Cahn, E., Krauss, M. J., Fitzsimmons-Craft, E. E., Wilfley, D. E. et Cavazos-Rehg, P. A. (2018). A content analysis of an online pro-eating disorder community on Reddit. *Body Image*, *24*, 137–144. https://doi.org/10.1016/j.bodyim.2018.01.001

STATCAN (2019). Statistique canada : Facteurs de protection et de risque en santé mentale. https://www.canada.ca/fr/sante-publique/services/facteurs-protection-et-risque-sante-mentale.html [Accessed : (15/08/2025)].

STATCAN (2023). Statistique canada : Enquête canadienne sur la santé des enfants et des jeunes (ECSEJ). https://www23.statcan.gc.ca/imdb/p2SV_f.pl?Function=getSurvey&Id=1504253 [Accessed : (15/08/2025)].

STATCAN (2024). Statistique canada : Profil socioéconomique de la population 2ELGBTQ+ âgée de 15 ans et plus, 2019 à 2021. https://www150.statcan.gc.ca/n1/daily-quotidien/240125/dq240125b-fra.htm [Accessed : (15/08/2025)].

Stephenson, E. (2023). Statistique canada : Troubles mentaux et accès aux soins de santé mentale. https://www150.statcan.gc.ca/n1/pub/75-006-x/2023001/article/00011-fra.htm [Accessed : (15/08/2025)].

Su, Y., Zheng, X., Lu, J., Gong, Y., Gao, Q., Shen, S. et Liu, Q. (2025). Semantic distillation and enhanced diagnostic alignment : A novel approach for depression detection in social media. *Expert Systems with Applications*, *279*, 127346. https://doi.org/https://doi.org/10.1016/j.eswa.2025.127346

Sutskever, I., Vinyals, O. et Le, Q. V. (2014). Sequence to Sequence Learning with Neural Networks. Dans Z. Ghahramani, M. Welling, C. Cortes, N. Lawrence, et K. Weinberger (dir.), *Advances in Neural Information Processing Systems*, Vol. 27. Curran Associates, Inc. https://proceedings.neurips.cc/paper_files/paper/2014/file/5a18e133cbf9f257297f410bb7eca942-Paper.pdf

Ta, N., Li, K., Yang, Y., Jiao, F., Tang, Z. et Li, G. (2022). Evaluating Public Anxiety for Topic-Based Communities in Social Networks. *IEEE Transactions on Knowledge and Data Engineering*, *34*(3), 1191–1205. https://doi.org/10.1109/TKDE.2020.2989759

van der Maaten, L. et Hinton, G. (2008). Visualizing Data using t-SNE. *Journal of Machine Learning Research*, *9*(86), 2579–2605. http://jmlr.org/papers/v9/vandermaaten08a.html

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L. u. et Polosukhin, I. (2017). Attention is All you Need. Dans I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, et R. Garnett (dir.), *Advances in Neural Information Processing Systems*, Vol. 30. Curran Associates, Inc. https://proceedings.neurips.cc/paper_files/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf

Wang, Q., Zhang, W. et An, S. (2023). A systematic review and meta-analysis of Internet-based self-help interventions for mental health among adolescents and college students. *Internet Interventions*, *34*, 100690. https://doi.org/10.1016/j.invent.2023.100690

Wang, W., Wei, F., Dong, L., Bao, H., Yang, N. et Zhou, M. (2020). MiniLM : Deep Self-Attention Distillation for Task-Agnostic Compression of Pre-Trained Transformers. https://arxiv.org/abs/2002.10957

Williams, A., Nangia, N. et Bowman, S. (2018). A Broad-Coverage Challenge Corpus for Sentence Understanding through Inference. Dans *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics : Human Language Technologies, Volume 1 (Long Papers)*, (p. 1112–1122). Association for Computational Linguistics. http://aclweb.org/anthology/N18-1101

World Health Organization (2025). World Health Organization (depression). https://www.who.int/news-room/fact-sheets/detail/depression [Accessed : (03/05/2025)].

Wu, Y., Schuster, M., Chen, Z., Le, Q. V., Norouzi, M., Macherey, W., Krikun, M., Cao, Y., Gao, Q., Macherey, K. *et al.* (2016a). Google's neural machine translation system : Bridging the gap between human and machine translation. *arXiv preprint arXiv :1609.08144*.

Wu, Y., Schuster, M., Chen, Z., Le, Q. V., Norouzi, M., Macherey, W., Krikun, M., Cao, Y., Gao, Q., Macherey, K., Klingner, J., Shah, A., Johnson, M., Liu, X., Kaiser, L., Gouws, S., Kato, Y., Kudo, T., Kazawa, H., … Dean, J. (2016b). Google's Neural Machine Translation System : Bridging the Gap between Human and Machine Translation. arXiv :1609.08144, https://doi.org/10.48550/arXiv.1609.08144

Yan, Z., Peng, F. et Zhang, D. (2025). DECEN : A deep learning model enhanced by depressive emotions for depression detection from social media content. *Decision Support Systems*, *191*, 114421. https://doi.org/https://doi.org/10.1016/j.dss.2025.114421

Yin, W., Hay, J. et Roth, D. (2019). Benchmarking zero-shot text classification : Datasets, evaluation and entailment approach. *arXiv preprint arXiv :1909.00161*.

Zhang, J., Chen, Y. et Zhai, Y. (2020). Zero-Shot Classification Based on Word Vector Enhancement and Distance Metric Learning. *IEEE Access*, *8*, 102292–102302. https://doi.org/10.1109/ACCESS.2020.2998495