L’intelligence artificielle pour la classification automatisée des textes courts issus de la vigie psychosociale : des méthodes classiques aux grands modèles de langage

Mémoire présenté

dans le cadre du programme de maîtrise en informatique

en vue de l’obtention du grade de maître ès sciences (M.Sc.)

PAR

© Fatima Azzahrae Adnane

Octobre 2025

Composition du jury :

Chan Wang Park , président du jury, Université du Québec à Rimouski

Mehdi Adda, directeur de recherche, Université du Québec à Rimouski

Lily Lessard, codirectrice de recherche, Université du Québec à Rimouski

Bruno Bouchard, examinateur interne, Université du Québec à Chicoutimi

Dépôt initial le 20 Mai 2025 Dépôt final le 08 Octobre 2025

Je dédie ce travail à ma mère et à mon père, en témoignage de ma profonde gratitude pour leur soutien constant et inconditionnel. Merci d’avoir toujours été à mes côtés.

REMERCIEMENTS

Tout d’abord, je souhaite adresser mes remerciements les plus sincères à mon directeur de recherche, Dr Mehdi Adda, pour son soutien continu, sa disponibilité, et sa précieuse collaboration. Grâce à son expertise et à son accompagnement, j’ai pu explorer de nouveaux aspects méthodologiques et élargir considérablement mes compétences. Ses conseils avisés et son encadrement rigoureux ont joué un rôle crucial dans l’aboutissement de ce mémoire.

Un immense merci à Dr Lily Lessard pour son soutien indéfectible, tant sur le plan professionnel qu’émotionnel. Sa contribution et son encadrement bienveillant ont été d’une grande importance tout au long de ce parcours. Sa présence constante, notamment pour m’épauler sur les aspects humains de ce projet, a été une source précieuse de motivation et de réconfort.

Je souhaite également exprimer ma profonde reconnaissance aux membres du jury qui ont accepté d’évaluer ce mémoire, Dr Chan Wang Park et Dr Bruno Bouchard. Leur temps, leur expertise et leurs commentaires constructifs ont contribué à enrichir ce travail et à en améliorer la qualité scientifique.

Je tiens également à remercier l'équipe de la direction régionale de santé publique de Chaudière-Appalaches qui soutient la vigie psychosociale et du Centre de recherche du CISSS de Chaudière-Appalaches pour m’avoir accordé leur confiance en mettant à ma disposition des données confidentielles essentielles à la conduite de ce projet. Leur collaboration a été un élément clé du développement et des avancées réalisées dans ce travail.

Enfin, je tiens à exprimer ma gratitude infinie à mes parents, mon frère et mes sœurs pour leur amour inconditionnel, leur soutien moral et leur confiance en moi. Leur présence à

mes côtés, même à distance, m'a donné la force de relever les défis et d'accomplir cette étape importante de ma vie académique.

AVANT-PROPOS

Présenté dans le cadre de la collaboration entre la direction régionale de la santé publique de Chaudière-Appalaches et l'UQAR, ce mémoire reflète les efforts des deux organisations pour concevoir de nouvelles solutions innovantes et technologiques au bien-être psychosocial des collectivités. Ce travail s'inscrit dans la volonté de mobiliser les avancées récentes en intelligence artificielle pour répondre à des besoins concrets en santé publique, notamment en facilitant l'analyse qualitative de données recueillies auprès de populations confrontées à des perturbations psychosociales majeures. Il trouve son origine dans des contextes précis tels que la pandémie de COVID-19, qui a mis en évidence la nécessité d'outils adaptatifs capables d'accompagner les intervenants de première ligne.

Mon intérêt pour ce sujet a été suscité par le fait que les préoccupations psychosociales sont souvent sous-évaluées ou mal catégorisées en raison de leur complexité et de leur évolution constante. Ce projet m'a permis d'explorer des approches innovantes dans le domaine de l'intelligence artificielle, afin de traiter automatiquement des textes rédigés en langage naturel, tels que les descriptions libres de préoccupations psychosociales exprimées par des citoyens, en passant par la classification traditionnelle, la modélisation thématique, et l'intégration des grands modèles de langage pour proposer des solutions explicables et adaptatives aux équipes qui administrent une vigie psychosociale et qui utilisent ses résultats.

L'objectif de ce travail est donc double, d'abord concevoir des outils pratiques permettant de classifier et analyser automatiquement les préoccupations psychosociales, provenant de la collectivité, afin de faciliter leur interprétation par les intervenants en santé publique non-initiés à la technicité de l'IA. Ensuite évaluer les limites des méthodes actuelles

et proposer des solutions fondées sur des technologies émergentes, tout en respectant la confidentialité des données sensibles.

RÉSUMÉ

Cette étude explore différentes approches pour automatiser la classification et l'analyse des préoccupations psychosociales exprimées par la population d'une région du Québec. Ces préoccupations sont recueillies sous forme de textes courts dans le cadre d'une veille psychosociale mise en œuvre pendant la pandémie de COVID-19 et qui se poursuit. L'objectif est de concevoir une application web permettant d'assister les intervenants en santé publique dans l'automatisation de la gestion de données.

Trois études ont été menées. Tout d'abord, une analyse comparative des algorithmes traditionnels de machine learning (k-NN, SVM, XGBoost) et des modèles de deep learning basés sur les transformateurs, combinés au Finetuning et au SetFit, a montré que les approches utilisant des sentence-transformers sont les plus performantes, atteignant une précision de 70,74 %, contre 68,69 % pour le modèle SVM. Ensuite, une modélisation thématique appliquée à un corpus d'environ 2000 entrées collectées dans les enquêtes mensuelles de la Vigie a permis d'identifier des tendances psychosociales émergentes. La méthode LDA avec unigrams a obtenu les meilleurs résultats (cohérence de 0,59). Enfin, des grands modèles de langage(LLMs), comme LLama3.1, LLama2 et Mistral, ont été intégrés pour transformer la classification en une tâche de génération de texte explicatif. Cette méthode a permis une classification multi-étiquettes et la génération d'explications adaptées aux besoins des équipes en santé publique , avec une précision atteignant 97.2 %.

Mots clés : intelligence artificielle, classification automatique, modélisation thématique, grands modèles de langage, préoccupations psychosociales, génération adaptative.

ABSTRACT

This study explores different approaches to automating the classification and analysis of psychosocial concerns expressed by the population of a region in Québec. These concerns are collected in the form of short texts as part of a psychosocial surveillance effort launched during the COVID-19 pandemic and currently continuing. The aim is to design an web application to assist healthcare actors in automating data management.

Three studies were carried out. Firstly, a comparative analysis of traditional machine learning algorithms (k-NN, SVM, XGBoost) and deep learning models based on transformers, combined with Finetuning and SetFit, showed that approaches using sentence-transformers performed best, achieving an accuracy of 70.74%, compared with 68.69% for the SVM model. Next, thematic modeling applied to a corpus of around 2000 entries collected in the monthly Vigie surveys identified emerging psychosocial trends. The LDA method with unigrams obtained the best results (consistency of 0.59). Finally, massive language models (LLMs), such as LLama3.1, LLama2 and Mistral, were integrated to transform classification into an explanatory text generation task. This method enabled multi-label classification and the generation of explanations tailored to the needs of professionals, with an accuracy of up to 97%.

Keywords: artificial intelligence, automated classification, topic modeling, large language models, psychosocial concerns, adaptive generation.

TABLE DES MATIÈRES

INTRODUCTION GÉNÉRALE

1. CONTEXTE ET PERTINENCE DES PREOCCUPATIONS PSYCHOSOCIALES

La pandémie de COVID-19 a occasionné divers impacts sur les populations humaines, notamment sous l'angle du bien-être psychosocial. Des portraits de ce bien-être ont été produit rapidement au Québec dès les premiers mois de la pandémie par l'Institut national de santé publique du Québec. Ces portraits, publiés au niveau régional ou national, étaient toutefois trop agrégés pour guider la prise de décision locale et les interventions en santé et en résilience communautaire.

En 2021, soucieuse de prendre le pouls des besoins de sa population à une échelle plus fine pour mieux orienter les actions, la direction de santé publique de Chaudière-Appalaches, en collaboration avec une équipe de recherche de l'Université du Québec à Rimouski (UQAR), ont établi une méthode pour mesurer en continu l'état de bien-être de la population à l'échelle des municipalités régionales de comité (MRC) et détecter rapidement les signaux faibles révélant de nouvelles fragilités sociales afin d'agir rapidement pour les prévenir ou les limiter. Cette vigie psychosociale (Vigie) a été conçue comme un outil de veille apte à capter, mois après mois, les variations fines de la santé psychosociale au sein des communautés.

La Vigie diffère des questionnaires d'auto-évaluation centrée sur l'individu habituels. Elle mobilise plutôt les observations d'intervenants communautaires, de responsables associatifs et de citoyens impliqués capables de porter un regard sur l'état d'un réseau social élargi.

Un autre élément d'originalité de cette Vigie réside dans le mélange entre des formats déclaratifs fermés, pour évaluer les niveaux de bien-être, de dynamisme et de bienveillance communautaires, et des questions ouvertes. Pour ces dernière, chaque répondant est invité à décrire, avec ses propres mots, les principales préoccupations psychosociales qu'il observe dans son entourage et à en prioriser trois.

Pour apprécier la pertinence scientifique de cette catégorie d'informations, il est essentiel de la replacer dans le cadre théorique des impacts psychosociaux défini par l'impact psychologique ou social, vécus par une personne ou un groupe, causés par des facteurs personnels ou environnementaux (Oliveira, 2013). Cette définition rappelle que les effets d'une crise ne se réduisent pas à la morbidité ou à la mortalité : ils englobent également les perturbations cognitives, émotionnelles, relationnelles et structurelles qui, cumulées, façonnent la résilience ou la vulnérabilité d'une collectivité (Patel V., 2019). Dans le contexte pandémique, ces impacts se sont traduits par une augmentation documentée des troubles anxiodépressifs, une détérioration du tissu économique local et une recrudescence des tensions intrafamiliales, de la précarité résidentielle ou de la stigmatisation.

L'analyse qualitative conduite par des experts humains permet une lecture approfondie du contexte, une interprétation fine des nuances sémantiques et une détection pertinente des signaux faibles ou émergents. Toutefois, lorsque le volume de données devient très important et que l'analyse doit être effectuée rapidement, cette approche demande beaucoup de ressources et devient difficile à appliquer à grande échelle.

C'est la problématique que l'équipe de la santé publique a rapidement atteint. Le seuil de saturation est d'environ 500 formulaires pouvant être analysés manuellement chaque mois, au prix d'un investissement important en personnel qualifié. Sans automatisation du traitement des données qualitatives, le processus d'analyse, comprenant la lecture, la catégorisation et la synthèse des réponses textuelles, représentait une charge de travail importante pour les équipes. Cette lourdeur risquait de retarder la diffusion des résultats de plusieurs semaines, réduisant ainsi la capacité du dispositif à répondre aux dynamiques psychosociales qu'il visait à éclairer.

Dans ce contexte, de multiples travaux ont mis en évidence l'efficacité des modèles de machine Learning (Venkatesh, 2024) pour la détection des thématiques associées à la santé mentale. En parallèle, la recherche actuelle, notamment dans le domaine de l'intelligence artificielle appliquée à la santé, cherche de plus en plus à rendre les approches explicables, et donc capables de justifier les décisions d'une manière beaucoup plus compréhensible et transparente (Huang, 2024).

C'est dans cette dynamique que s'inscrit le présent mémoire, en explorant des méthodes avancées d'automatisation de la classification et de l'analyse des préoccupations psychosociales, tout en intégrant des mécanismes d'explicabilité adaptés à un public non initié aux technologies de l'intelligence artificielle.

2. PROBLEMATIQUE

Malgré les avancées en intelligence artificielle, la classification des préoccupations psychosociales exprimées en texte libre demeure un défi, notamment lorsqu'il s'agit de textes courts. Ces énoncés, souvent très brefs sont marqués par une forte variabilité lexicale, une ambiguïté contextuelle, et une absence de structure grammaticale formelle, ce qui limite la performance des algorithmes traditionnels, pourtant efficaces sur des données plus longues, structurées et annotées (Buda et al., 2018). De plus, la rapidité d'émergence de nouvelles thématiques exige des modèles adaptatifs, capables non seulement de reconnaître des catégories connues, mais aussi d'identifier des préoccupations inédites et évolutives.

Un second défi d'importance se pose autour de l'explicabilité. Les grands modèles de langage , comme les modèles BERT, GPT ou LLaMA (H.Touvron, 2023), ont effectivement marqué le traitement du langage naturel, mais leur fonctionnement demeure en grande partie opaque (Abdurahman, 2023). Or, dans un secteur aussi sensible que la santé psychosociale, il est indispensable de produire des justifications explicites et compréhensibles pour

permettre aux praticiens et acteurs décisionnels d'avoir confiance dans ces modèles d'apprentissage. À ces enjeux s'ajoute la nature évolutive des préoccupations psychosociales: de nouvelles thématiques apparaissent régulièrement, ce qui exige des méthodes capables de s'adapter à des catégories d'analyse en constante transformation, tout en maintenant une classification cohérente et pertinente dans le temps. Deux questionnements de recherche apparaissent donc prioritaires : Comment classifier efficacement des données textuelles issues de préoccupations psychosociales exprimées librement par des centaines de répondants, alors même que les catégories d'analyse évoluent au fil du temps ? Comment mobiliser les grands modèles de langage afin de fournir des explications transparentes, contextualisées et appropriées au terrain ?

3. OBJECTIF

Ce mémoire s'inscrit dans le projet de la Vigie psychosociale et consiste à concevoir un outil technologique pour l'automatisation de la classification des préoccupations psychosociales récoltées chaque mois auprès de centaines de répondant sous la forme de texte libre. Ces préoccupations étant susceptibles d'évoluer au fil du temps, l'approche proposée vise à concevoir un système flexible, capable d'intégrer de nouvelles catégories sans nécessiter un réentraînement complet.

Ce travail a pour ambition de proposer des solutions permettant non seulement de classifier les préoccupations exprimées mais également de fournir aux intervenants en santé publique, des explications interprétables et adaptatives pour guider le processus de prise de décision et améliorer les services communautaires ainsi que les services de santé et sociaux offerts aux populations qui soulèvent des préoccupations ou besoins particuliers. Les objectifs de ce mémoire se résument donc comme suit :

- Automatiser la classification des préoccupations psychosociales exprimées en texte libre.

- S’adapter à l’apparition de nouvelles thématiques et à l’évolution des catégories d’analyse.
- Fournir des explications claires et interprétables pour soutenir les intervenants dans leur prise de décision.

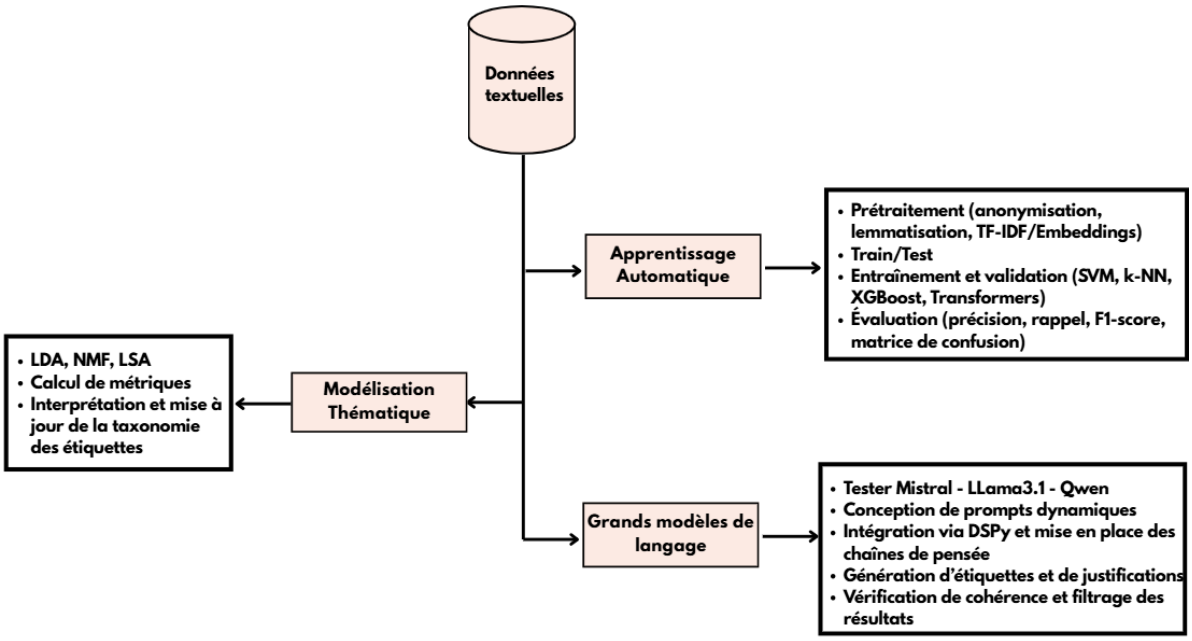
4. METHODOLOGIE :

Dans le but d’atteindre les objectifs cités préalablement, ce mémoire s’appuie sur une méthodologie tripartite complémentaire. La première étape consiste à automatiser la classification des préoccupations psychosociales exprimées en texte libre. Les réponses collectées mensuellement via Lime Survey font l’objet d’un prétraitement comprenant l’anonymisation, le nettoyage et la normalisation linguistique (lemmatisation, suppression des stopwords) afin d’assurer la qualité des données. Ces données sont ensuite vectorisées à l’aide de techniques adaptées, telles que TF-IDF pour les modèles classiques ou des embeddings pour les modèles fondés sur les transformateurs, avant d’être utilisées pour entraîner et évaluer différents algorithmes. Les performances des modèles traditionnels comme les SVM, les forêts aléatoires et les k-NN sont comparées à celles des modèles récents, notamment les transformateurs, à l’aide de métriques standard telles que la précision, le rappel et le F1-score, afin de déterminer l’approche la plus robuste et la plus adaptée aux textes courts et hétérogènes (Tunstall, 2022).

Le processus mis en place vise ensuite à s’adapter à l’apparition de nouvelles thématiques et à l’évolution des catégories d’analyse. Pour ce faire, les corpus mensuels sont soumis à des techniques de modélisation thématique telles que LDA, LSA ou NMF, ce qui permet d’identifier les thèmes dominants et de détecter les signaux faibles indiquant l’émergence de nouvelles préoccupations à l’échelle infrarégionale. Les catégories d’analyse sont mises à jour en conséquence, et ces ajustements sont validés par les experts de terrain afin d’assurer la pertinence et l’utilité des étiquettes utilisées.

Enfin, la stratégie retenue se poursuit en mettant l'accent sur la production des explications claires et interprétables pour soutenir les intervenants dans leur prise de décision. Dans cette optique, des grands modèles de langage comme LLaMA 3.1, LLaMA 2 et Mistral sont intégrés dans le processus au moyen d'une stratégie de prompting. Cette intégration est implémentée à l'aide de DSPy, qui facilite la construction de chaînes de traitement modulaires et la mise en place de chaînes de pensée qui améliorant ainsi la qualité des raisonnements générés et la cohérence des justifications produites. Chaque classification s'accompagne ainsi d'une justification textuelle contextualisée, mettant en évidence les éléments lexicaux ou sémantiques ayant motivé la prédiction, rendant le système non seulement performant mais également explicatif et adaptatif, capable de suivre l'évolution des préoccupations psychosociales sans nécessiter un réentraînement constant.

Les résultats ainsi obtenus sont restitués via une application web développée avec Panel et déployée sur un serveur distant, permettant aux intervenants d'explorer les données, de visualiser l'évolution des préoccupations dans le temps et d'exporter les analyses pour les besoins décisionnels.



5. CONTRIBUTION

L'originalité de ce mémoire repose sur l'exploration et l'évaluation de plusieurs approches pour la classification de préoccupations psychosociales exprimées sous forme de textes courts. Il propose une comparaison structurée entre des méthodes de classification traditionnelles, des techniques de modélisation thématique, et l'utilisation de grands modèles de langage accessibles localement.

L'intérêt de cette étude réside dans la transformation progressive de la tâche de classification en une tâche de génération guidée par des prompts, permettant d'attribuer des étiquettes existantes ou d'en proposer de nouvelles lorsque cela est nécessaire. L'approche mise en place tient compte de contraintes spécifiques au contexte d'intervention, comme la brièveté des textes, l'évolution des thématiques dans le temps, et le besoin d'une certaine lisibilité des sorties pour les professionnels.

Les résultats de cette recherche ont permis de souligner que certains modèles traditionnels, tels que les SVM, conservent une performance compétitive sur des données textuelles courtes, les approches fondées sur les transformateurs offrent une précision supérieure et une meilleure robustesse face à la variabilité lexicale (Fatima azzahrae , Amraoui, Adda, & Lessard, 2024). Par ailleurs, il a été démontré que les techniques de modélisation thématique, notamment LDA avec unigrams, sont particulièrement efficaces pour détecter des tendances émergentes et suivre l'évolution des préoccupations psychosociales dans le temps (Amraoui, Adnane, Adda, & Lessard, 2024), offrant ainsi une lecture longitudinale précieuse des données de vigie. Enfin, le dernier article propose l'intégration des grands modèles de langage (LLaMA 3.1, Mistral, Qwen) pour transformer la classification en une tâche de génération explicative, atteignant une précision de 97,2 %

au niveau des sous-catégories et produisant des justifications textuelles exploitables par les intervenants en santé publique.

L'utilisation de modèles de langage de grande taille dans ce mémoire soulève plusieurs défis techniques, éthiques et sociétaux qu'il convient de reconnaître. Sur le plan technique, leur coût computationnel et énergétique demeure particulièrement élevé.

Dans la continuité de ces travaux, une application web a été développée pour permettre aux intervenants d'utiliser ces outils dans un cadre opérationnel, en facilitant l'analyse et l'organisation des données recueillies. Ce travail se limite à une validation sur un jeu de données spécifique, mais il pose les bases d'un système adaptable à d'autres contextes similaires.

6. STRUCTURE DU MEMOIRE

Ce mémoire suit la structure d'un mémoire par articles. Il est structuré en quatre chapitres, dont un premier consacré à l'état de l'art, suivi de trois chapitres correspondant chacun à un article scientifique. Chaque article se propose d'ouvrir un axe méthodologique spécifique, soutenant ainsi la mise en lumière ou la résolution des problèmes de recherche.

Le premier chapitre est la présentation d'un état de l'art, sous forme d'article, qui regroupe les concepts clés, les définitions obligatoires, et les recherches précédentes en classification puis analyse des problématiques psychosociales qui sont l'objet de ce travail, mais également le cadre théorique et conceptuel censé éclairer et faciliter la compréhension des articles qui le suivent, ce dernier incluant les notions de machine learning traditionnel, la modélisation thématique, et les grands modèles de langage(LLMs), tout en pointant les lacunes relevées dans les travaux antérieurs, que ce mémoire cherche à combler, et qui justifient sa démarche.

Le deuxième chapitre (Fatima azzahrae , Amraoui, Adda, & Lessard, 2024) , résultant du premier article, est centré sur une analyse comparative entre les algorithmes de machine learning traditionnel (SVM, k-NN, XGBoost, etc.), et les approches avancées par les transformers en général et les sentence-transformers (SetFit) en particulier pour la classification des préoccupations psychosociales, montrant ainsi les limites des méthodes traditionnelles ainsi que la supériorité des nouvelles approches.

Le troisième chapitre (Amraoui, Adnane, Adda, & Lessard, 2024), s'inspirant du second article, se penche sur l'application de techniques de modélisation thématique (LDA, NMF, LSA) pour analyser des corpus textuels issus de données d'enquêtes mensuelles. Il permet de faire émerger les nouvelles tendances et de faire un suivi temporel de l'évolution des préoccupations psychosociales.

Enfin, le quatrième chapitre, issu du troisième article, introduit une approche novatrice basée sur les grands modèles de langage, tels que LLama3.1, LLama2 et Mistral, pour transformer la tâche de classification en une démarche explicative et adaptative. Cette méthode propose des étiquettes hiérarchiques (Niveau 1, Niveau 2), des analyses contextuelles et des visualisations interactives adaptées aux besoins spécifiques des intervenants en santé publique.

Une conclusion générale vient clore le mémoire en récapitulant les principaux points discutés et en ouvrant sur les perspectives de recherche futures. Les implications pratiques et éthiques de ce travail sont également discutées, en soulignant notamment les enjeux liés à l'acceptabilité de tels systèmes d'IA dans le champ de la santé publique, de la surveillance et de l'intérêt de poursuivre ces investigations afin de mieux traiter des volumes croissants de témoignages, de réponses ouvertes et de signaux textuels recueillis dans des contextes de veille psychosociale.

CHAPITRE 1 : ÉTAT DE L'ART

1.1 RESUME EN FRANÇAIS DU PREMIER ARTICLE

Cet article, intitulé «Classification de données textuelles : Étude comparative de l'apprentissage automatique, de l'apprentissage profond et des grands modèles de langage», constitue une revue de l'état de l'art et ne fera pas l'objet d'une publication pour l'instant.

En tant que première auteure, j'ai structuré l'analyse des concepts clés, synthétisé les recherches existantes et organisé le cadre théorique et conceptuel afin d'éclairer et de faciliter la compréhension des trois articles suivants.

L'article met l'accent sur les concepts fondamentaux nécessaires à l'étude des préoccupations psychosociales, en explorant les approches de classification traditionnelles : Apprentissage automatique, apprentissage profond, modélisation thématique et l'évolution vers les grands modèles de langage(LLMs). Il identifie également les principales limites des méthodes existantes et les défis spécifiques au domaine psychosocial.

Il propose ensuite une analyse des avancées récentes en intelligence artificielle générative, en mettant en avant l'apport des LLMs et des stratégies de prompt engineering (zero-shot, few-shot, chain-of-thought) pour améliorer la classification et l'interprétation des préoccupations.

Enfin, il établit le cadre théorique qui justifie l'orientation du mémoire, en positionnant les recherches futures comme une réponse aux limites identifiées et en préparant la transition vers les contributions des trois articles suivants.

1.2 ÉTAT DE L'ART :

Classification de données textuelles : Étude comparative de l'apprentissage automatique, de l'apprentissage profond et des grands modèles de langage

Adnane Fatima-Azzahrae[a,*], Adda Mehdi[a,**], Lessard Lily[b]

[a]*Département de Mathématiques, Informatique et Génie, Université du Québec à Rimouski (UQAR), Canada (QC)*
[b]*Département des sciences de la santé, Université du Québec à Rimouski (UQAR), Canada (QC)*

Abstract

Le présent article ambitionne d'offrir un cadre conceptuel et méthodologique de compréhension de l'évolution des techniques du traitement automatique du langage naturel, aussi bien dans leur versant traditionnel que dans leur versant moderne. Il présente dans un premier temps les techniques traditionnelles d'apprentissage suprvisé et de vectorisation comme les classifieurs classiques (SVM, Naïve Bayes, XGBoost,) et les transformées (Bag-of-Words, TF-IDF, word embeddings). Puis, il s'intéresse aux techniques contemporaines de l'apprentissage profond, en particulier au modèle des *Transformers* qui redéfinissent le domaine tant les modèles de langage massive, notamment les modèles de langage massif (LLMs). Par ailleurs, de plus en plus de travaux introduisent les méthodes non supervisées telles que la modélisation de sujets (à savoir LDA, NMF, LSA), qui s'utilisent pour extraire des thèmes d'un corpus n'ayant pas été étiqueté. L'article aborde également la problématique des textes courts, le déséquilibre des classes, et la fiabilité des modèles en termes de génération de texte.

Keywords: Artificial intelligence; Natural language processing; Psychosocial concerns; Classification; apprentissage automatique; Apprentissage profond.

1. Introduction

Au cours des dernières années, les approches en intelligence artificielle (IA) ont connu beaucoup de transformations, en passant des méthodes d'apprentissage automatique classiques, nécessitant souvent une ingénierie des attributs, à des méthodes de l'apprentissage profond [1]. Ces changements ont été facilités par l'augmentation du volume de données disponibles, les avancées en matériel informatique, notamment les unités de traitement graphique (Graphics Processing Units, GPU) et les unités de traitement tensoriel (Tensor Processing Units, TPU), ainsi que par l'amélioration des architectures neuronales [62].

Un domaine comme le traitement automatique du langage naturel (TALN), à l'instar de nombreuses disciplines de l'IA, est marqué par l'émergence de nouvelles approches fondées sur l'apprentissage profond [3]. Si, à ses débuts, seules des représentations de texte de base (bag-of-words, TF-IDF) et des algorithmes de classification traditionnelle étaient au programme, la recherche dans le domaine de l'IA a progressivement amené à construire

* Corresponding author : Adnane Fatima-Azzahrae
** Corresponding author : Adda Mehdi

des modèles neuronaux tels que les réseaux de neurones récurrents (*Recurrent Neural Networks*, RNN) et les réseaux de neurones convolutifs (*Convolutional Neural Networks*, CNN) [2] qui ont fini par évoluer, en intégrant des architectures plus complexes. L'introduction de l'architecture Transformer, reposant entièrement sur des mécanismes d'attention, a permis de surmonter certaines limites des RNN en matière de parallélisation et de traitement de longues dépendances. Les Transformers ont également ouvert la voie à une nouvelle génération de modèles de langage massifs (*Large Language Models* , LLMs) [4] conçus pour exploiter des ensembles de données à grande échelle, affichant aujourd'hui des performances améliorées dans des tâches diverses telles que la classification, la génération de texte ou la compréhension sémantique.

En plus de la classification de texte, d'autres types de tâches peuvent être nécessaires dans l'analyse de données textuelles, notamment celles visant à explorer, structurer ou synthétiser l'information contenue dans un corpus. Dans ce contexte, il est possible de recourir à d'autres types de méthodes non supervisées comme la modélisation des sujets, qui permettent de révéler automatiquement des structures latentes et des thématiques sous-jacentes au sein du corpus, sans nécessiter d'annotations préalables.
Cette démarche méthodologique, appliquée à l'analyse automatique de textes dans le contexte du TALN, s'appuie sur des techniques de classification, de modélisation thématique et de génération de texte à l'aide des LLMs. Elle met en évidence le potentiel de ces approches pour traiter des corpus non étiquetés et met en avant les aspects suivants :

— Une comparaison entre les méthodes classiques et neuronales pour la classification de textes courts.
— Une exploration des différentes méthodes de modélisation thématique (LDA, NMF, LSA)[5] appliquées à des corpus non annotés.
— L'intégration des LLMs [4] générer des étiquettes et approfondir l'analyse sémantique à l'aide de diverses techniques.

Afin de garantir la rigueur de l'analyse, une stratégie de sélection explicite a été mise en œuvre. Celle-ci s'est appuyée sur des critères d'inclusion et d'exclusion définis en amont, une sélection de bases de données scientifiques reconnues, et une requête systématique par mots-clés.
Les critères d'inclusion retenus pour cette revue sont présentés dans le tableau 1, qui précise les conditions d'éligibilité des études en termes de période, de type de publication, de nature des données, de thématique et de langue.

Table 1. Critères d'inclusion des études sélectionnées

Critère	**Définition**
Période de publication	Inclusion large couvrant les contributions fondatrices en apprentissage automatique sans limite stricte de date ; accent particulier sur les publications entre 2020 et 2025 pour les approches en apprentissage profond et LLMs.
Type de publication	Articles évalués par les pairs dans des revues scientifiques ou actes de conférence indexés.
Type de données analysées	Données textuelles courtes : questionnaires, forums, réseaux sociaux, journaux de bord, etc.
Domaine thématique	Préoccupations psychosociales, santé mentale, bien-être individuel
Langue des documents	Anglais

Plusieurs critères d'exclusion ont été définis afin d'assurer la rigueur de la sélection. Les études ont été exclues lorsqu'elles ne faisaient pas l'objet d'une évaluation par les pairs, les publications dont le texte intégral n'était pas accessible, celles rédigées dans une langue autre que l'anglais, ainsi que les études ne présentant pas de lien explicite avec la classification ou l'analyse de données textuelles courtes.

Nous avons utilisé plusieurs bases de données et bibliothèques virtuelles. Scopus a été l'une des principales sources en raison de sa couverture exhaustive de revues scientifiques, techniques et médicales. IEEE Xplore a fourni un accès aux publications en informatique et en ingénierie, tandis que Google Scholar a été utilisé pour une recherche plus générale incluant des thèses et des rapports techniques. Finalement Mendeley a été utilisé non seulement comme

outil de gestion de références, mais aussi pour découvrir des publications pertinentes via ses fonctionnalités de recommandation.

La stratégie de recherche a consisté à utiliser des mots-clés spécifiques et des combinaisons de termes pour garantir une couverture exhaustive du sujet. Les termes de recherche incluaient "classification de texte", "apprentissage automatique", "apprentissage profond", "*Transformers*", "TALN " et "LLMs".

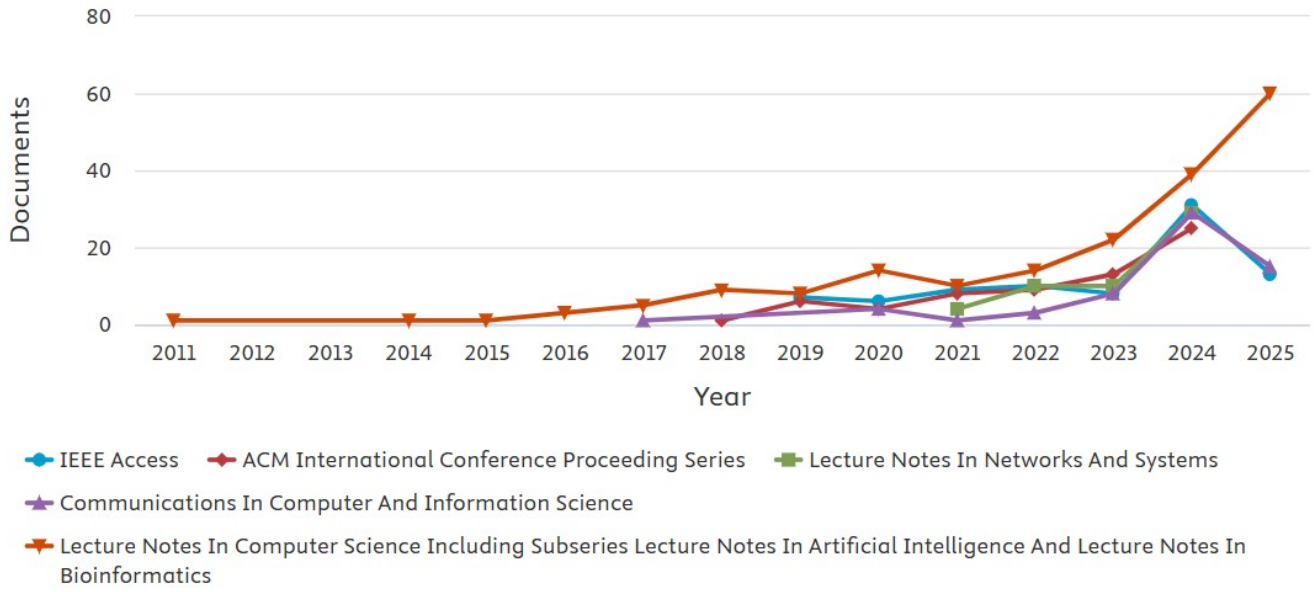


Figure 1. Évolution des publications scientifiques dans différentes sources indexées par Scopus

La Figure 1 illustre l'évolution du nombre de publications scientifiques sur la thématique étudiée au fil des années, en se basant sur les principales sources indexées dans Scopus. On observe une croissance significative à partir de 2020, avec une accélération marquée en 2023 et 2024.

Ce document est organisé de manière à offrir une progression logique, depuis les concepts généraux jusqu'aux applications et méthodologies plus pointues. Après cette introduction, la section 2 et 3 proposent un état de l'art détaillé sur les méthodes d'apprentissage automatique et de traitement automatique du langage naturel, en rappelant notamment l'émergence de l'apprentissage profond et des architectures transformer. La section4 approfondit les principes et les limites de la modélisation thématique (LDA, NMF, LSA), tandis que la section6 décrit plus précisément les caractéristiques des modèles de langage massifs (LLMs) et leurs impacts sur la classification de textes courts. Enfin, la conclusion générale met en perspective l'ensemble de ces notions, en soulignant comment elles se déclinent au sein des trois travaux de recherche auxquels cet article sert de cadre conceptuel et méthodologique.

2. L'apprentissage automatique en TALN

L'apprentissage automatique est une approche computationnelle qui repose sur la capacité d'un modèle à apprendre des motifs et des régularités à partir de données, sans être explicitement programmé pour chaque tâche [14]. Avec l'émergence de nouvelles approches[6], il a connu des transformations dans la manière d'interroger des tâches de classification textuelle. Cette première section vise à expliciter les concepts de base et techniques de la littérature mobilisés dans les approches de ce domaine.

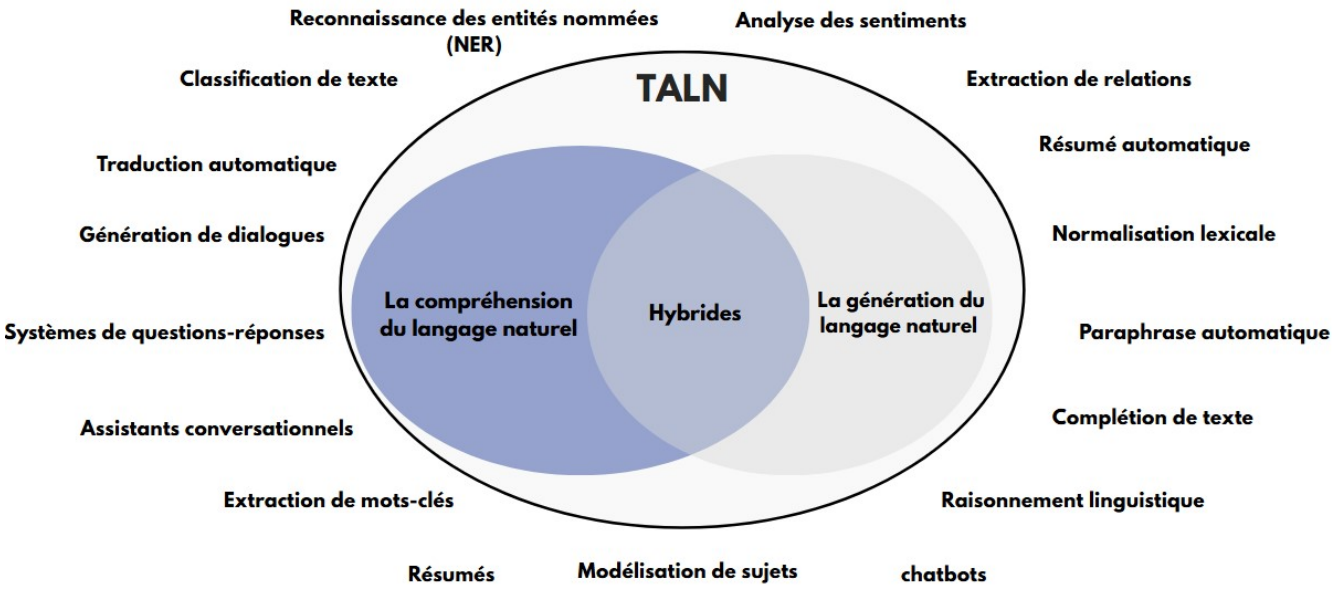


FIGURE 2. Taxonomie fonctionnelle des tâches en TALN

Afin de poser un cadre théorique, nous allons examiner les fondements classiques de l'apprentissage automatique [7], notamment les algorithmes supervisés SVM, k-NN, Naive Bayes, et les techniques de vectorisation comme Bag-of-Words, TF-IDF ou n-grams, ayant trouvé de nombreuses applications en TALN [9].

2.1. Cartographie des tâches du TALN

Le traitement automatique du langage naturel est une branche qui combine la linguistique informatique et l'apprentissage automatique, ce dernier incluant l'apprentissage profond comme sous-branche, afin de permettre l'interprétation, la manipulation et la compréhension du langage humain. Les tâches qui relèvent du TALN peuvent être organisées selon leur finalité en trois grandes catégories fonctionnelles : la compréhension, la génération et les tâches hybrides. La compréhension du langage naturel regroupe les tâches qui visent à analyser et interpréter un texte, comme la classification de texte[7], la détection d'intention, la reconnaissance des entités nommées (NER) [100] et l'analyse des sentiments, l'extraction de relations ou la modélisation de sujets. À l'inverse, la génération du langage naturel se concentre sur la production de texte cohérent et contextuellement pertinent. Elle peut inclure des applications comme la traduction automatique, le résumé automatique et la génération de réponses aux questions, en mobilisant des modèles génératifs de type LLMs. Entre ces deux pôles, certaines tâches dites hybrides combinent des capacités d'analyse et de génération, notamment les chatbots [8], les systèmes de questions-réponses, la complétion de texte ou la réécriture automatique. Ces tâches mobilisent des compétences mixtes, consistant à comprendre le contexte en entrée pour produire une sortie linguistiquement et sémantiquement adéquate. Par ailleurs, les tâches du TALN trouvent des applications dans une grande diversité de domaines, notamment en médecine, où elles permettent d'extraire des informations pertinentes à partir de comptes rendus cliniques ou radiologiques [53], facilitant le traitement de grandes masses de données textuelles non structurées. En robotique, il contribue à la traduction d'instructions langagières en commandes opérationnelles, renforçant les capacités d'interaction entre humains et machines[52]. Le secteur de la construction mobilise également le TALN pour analyser des documents techniques, modéliser des exigences ou interpréter des contrats en langage naturel. Dans ce même contexte, des applications en psychologie et en ressources humaines exploitent les capacités d'analyse textuelle pour détecter les préoccupations de la population et les classifier [83].

2.1.1. *Introduction à l'apprentissage automatique et ses méthodes*

Après avoir cartographié ces tâches et mis en évidence leurs différentes catégories, il est essentiel d'examiner les méthodes qui permettent leur mise en œuvre. L'apprentissage automatique représente un sous-domaine de l'IA, il repose sur la mise en œuvre d'algorithmes capables d'extraire des structures et des régularités à partir de données sans nécessiter des règles explicites préprogrammées [14]. Son objectif est de modéliser divers problèmes en apprenant à partir d'un ensemble d'exemples, de manière à pouvoir généraliser à de nouveaux cas. Les différentes méthodes en apprentissage automatique se regroupent principalement dans l'une des quatre catégories suivantes : D'abord l'apprentissage supervisé [14] qui recourt à des ensembles de données étiquetées où chaque observation se voyant assigner une sortie attendue, ce qui le rend adapté aux problèmes de classification, comme la prédiction de catégories, ou de régression (cas des valeurs continues). Par ailleurs, l'apprentissage non supervisé s'appuie sur des ensembles de données non étiquetées, permettant d'accéder à une structure ou à des relations non triviales dans ces données [15], ce qui inclut le regroupement des données en ensembles homogènes selon leurs similarités et la réduction de la dimensionnalité pour simplifier les variables tout en conservant l'essentiel de l'information . Une troisième catégorie, l'apprentissage semi-supervisé, combine les avantages des deux méthodes précédentes. Elle exploite un petit sous-ensemble de données étiquetées, combiné à un grand volume de données non étiquetées, pour entraîner les modèles de manière plus efficace [16]. Enfin, nous retrouvons l'apprentissage par renforcement, qui repose, sur l'interaction avec un environnement. Les modèles s'y affinent en recevant des signaux de récompense ou de pénalité en fonction des actions réciproques entreprises chaque fois et en étant régulièrement nourris des résultats subséquents de leurs actions [17]. C'est ce qui leur permet d'évoluer dans la direction la plus adaptée au fur et à mesure. À ces catégories classiques s'ajoute aujourd'hui une approche générationnelle incarnée par les LLMs. Ces modèles, tels que GPT, BERT ou T5, ont pour particularité de produire du contenu textuel en sortie, qu'il s'agisse de répondre à des questions, de reformuler un texte, de générer une suite plausible ou de traduire. Bien qu'ils puissent être entraînés de façon supervisée ou auto-supervisée, leur finalité les distingue en tant que modèles génératifs, orientés vers la production autonome de texte.

2.2. *Prétraitement des données textuelles en TALN*

La classification constitue la tâche centrale abordée dans cet article. En TALN, elle permet d'attribuer une étiquette à un texte en fonction de son contenu [6]. Toutefois, avant de pouvoir classifier un texte, il est nécessaire de le transformer en une représentation exploitable par un modèle d'apprentissage automatique [21]. Cette transformation repose sur une série d'étapes de prétraitement, illustrées dans la Fig. 3, qui incluent notamment le nettoyage du texte, la normalisation, la tokenisation, ainsi que l'extraction et la vectorisation des caractéristiques. Il convient toutefois de souligner que les opérations de prétraitement à mettre en œuvre dépendent fortement de la nature des techniques de modélisation employées. Par exemple, les approches classiques nécessitent souvent un prétraitement rigoureux (suppression des stopwords, lemmatisation, etc.), alors que les modèles d'apprentissage profond ou les LLMs intègrent en amont des mécanismes de traitement du langage naturel, rendant certaines de ces étapes facultatives, voire superflues dans certains cas.

Plusieurs étapes peuvent être mobilisées à ce stade [20]. Il convient, dans un premier temps, de supprimer les caractères spéciaux, d'harmoniser la casse (généralement en convertissant le texte en minuscules), de corriger l'orthographe lorsque cela s'avère pertinent, et d'anonymiser certaines données sensibles si nécessaire. Ensuite, la normalisation du texte intervient, une étape qui consiste à ajuster les représentations numériques des données textuelles dans une échelle commune, facilitant ainsi des comparaisons fiables et standardisées entre les différentes entrées [22]. De plus, le texte est découpé en unités plus petites (ce que l'on appelle la segmentation), en séparant les phrases, les mots, ou parfois les sous-unités de mots [23]. On élimine également dans un objectif d'efficacité les mots vides les plus fréquents (le, la, de, etc.) dont il est souvent démontré qu'ils n'apportent pas véritablement de matière pour la caractérisation des objets.
Enfin, pour éviter une trop grande variation lexicale, des techniques de transformation sont appliquées. L'élagage consiste à supprimer les suffixes des mots pour en extraire une racine approximative, souvent sans tenir compte du contexte grammatical, ce qui peut générer des formes incorrectes ou la lemmatisation, elle repose sur des dictionnaires linguistiques et des règles grammaticales pour ramener un mot à sa forme canonique, ou lemme, en fonction de son contexte [21].

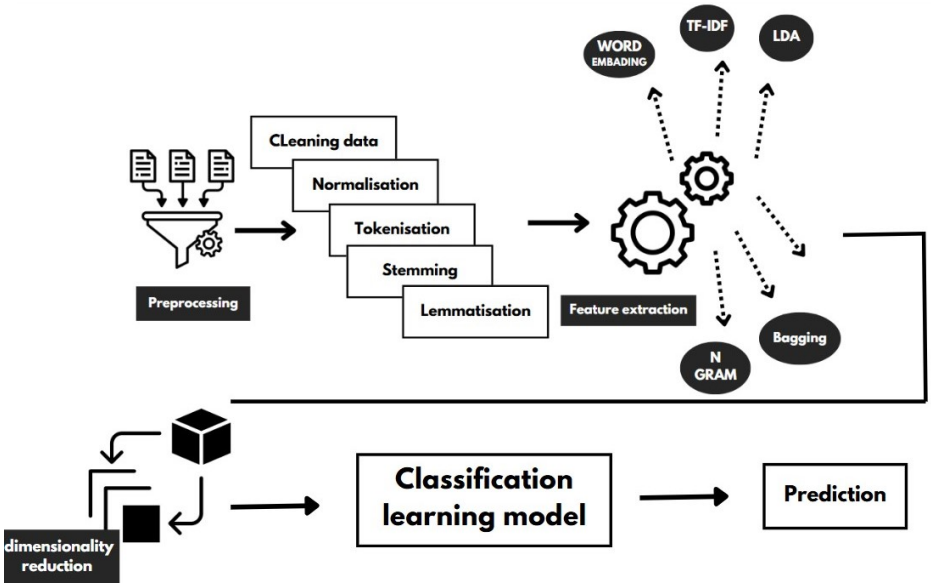


FIGURE 3. Flux de traitement des données en apprentissage automatique (inspiré de [102])

2.3. Techniques d'extraction et de vectorisation des caractéristiques

Les modèles d'apprentissage automatique sont conçus pour traiter des données numériques, il est donc nécessaire de transformer les textes, généralement non structurés, en représentations numériques exploitables. Cette transformation repose sur deux étapes complémentaires : l'extraction des caractéristiques pertinentes du texte, suivie de leur vectorisation. L'extraction vise à identifier les éléments linguistiques porteurs d'information (mots, expressions, structures syntaxiques, etc.), tandis que la vectorisation consiste à convertir ces éléments en représentations numériques, appelées vecteurs de caractéristiques. Ces représentations peuvent ensuite être utilisées par des algorithmes d'apprentissage supervisé ou non supervisé. Plusieurs techniques sont couramment utilisées dans ce processus, parmi les plus simples :

— Bag of Words (BOW) : une représentation simple d'un texte sous la forme d'un vecteur à partir de la présence ou non de mots, sans notion d'ordre. Au cours de ce processus, des listes de mots sont créées et organisés sous forme de matrice. Ils ne sont pas structurés en phrases et ne suivent aucune grammaire, elle se traduit donc par un schéma ignorant complètement le contexte sémantique mais prenant en considération la fréquence de chaque mots. Voici un exemple illustratif :

  Soit les deux préoccupations suivantes :

  — Phrase 1 (P1) : "Je ressens du stress et de l'anxiété à cause de ma charge de travail."
  — Phrase 2 (P2) : "L'insécurité financière m'empêche de dormir la nuit."

  Après prétraitement (segmentation, suppression des mots vides, lemmatisation), nous obtenons le vocabulaire unique :

$$V = \{\text{stress, anxiété, charge, travail, insécurité, financière, dormir, nuit}\}$$

  Nous représentons alors chaque préoccupation sous forme de vecteur individuel. Dans cette représentation vectorielle, chaque composante du vecteur correspond à un mot du vocabulaire commun V. La valeur 1 indique que le mot correspondant est présent dans la phrase, tandis que la valeur 0 signifie qu'il est absent. :

$$\text{BoW}(P1) = \begin{bmatrix} 1 & 1 & 1 & 1 & 0 & 0 & 0 & 0 \end{bmatrix}$$
$$\text{BoW}(P2) = \begin{bmatrix} 0 & 0 & 0 & 0 & 1 & 1 & 1 & 1 \end{bmatrix}$$

Toutefois, lorsque le vocabulaire s'étend à plusieurs milliers de termes, les modèles Bag-of-Words rencontrent des limitations. En effet, cette représentation génère des vecteurs de grande dimension, souvent très creux, dont la majorité de ses composantes sont égales à zéro, ce qui augmente la complexité computationnelle, la consommation mémoire et peut nuire à la performance des modèles en apprentissage automatique.

— TF-IDF (Term Frequency–Inverse Document Frequency) : il s'agit d'une méthode de pondération utilisée pour représenter les textes sous forme vectorielle. Elle permet de quantifier l'importance d'un mot dans un document donné, en tenant compte de sa fréquence d'apparition dans ce document (Term Frequency) et de sa rareté à l'échelle de l'ensemble du corpus (Inverse Document Frequency) [26]. Cette mesure accorde ainsi un poids plus élevé aux termes caractéristiques d'un document, tout en atténuant l'influence des mots fréquents mais peu informatifs Le TF-IDF d'un mot t dans un document d est défini comme :

$$\text{TF-IDF}(t, d) = \text{TF}(t, d) \times \text{IDF}(t) \tag{1}$$

avec :

$$\text{TF}(t, d) = \frac{f_{t,d}}{\sum_i f_{i,d}} \tag{2}$$
$$\text{IDF}(t) = \log \frac{N}{1 + \text{DF}(t)} \tag{3}$$

où N est le nombre total de documents et $\text{DF}(t)$ est le nombre de documents contenant le mot t.
Considérons un mini-corpus composé de deux préoccupations liées à la santé psychosociale :

— Doc1 : *« Je ressens du stress et de l'anxiété depuis la perte de mon emploi. »*
— Doc2 : *« Mon anxiété me pousse à l'isolement et trouble mon sommeil. »*

Après prétraitement (suppression des mots vides, mise en minuscules, lemmatisation légère), on extrait le vocabulaire suivant :

$$V = \{\text{ressentir, stress, anxiété, perte, emploi, pousser, isolement, troubler, sommeil}\}$$

Chaque document contient 5 mots utiles. La fréquence d'apparition (TF) de chaque mot est donc 0.20 pour les mots présents, et 0 pour les absents.
En termes de fréquence inverse de document (IDF), on considère ici que :

— Les mots présents dans les deux documents (comme *anxiété*) ont une IDF faible ;
— Les mots présents dans un seul document ont une IDF plus élevée.

On obtient alors la matrice simplifiée suivante :

Mot	**TF (Doc1)**	**TF (Doc2)**	**IDF**	**TF-IDF**
ressentir	0.20	0	élevé	0.20 (Doc1)
stress	0.20	0	élevé	0.20 (Doc1)
anxiété	0.20	0.20	**faible**	0.05 (Doc1 et Doc2)
perte	0.20	0	élevé	0.20 (Doc1)
emploi	0.20	0	élevé	0.20 (Doc1)
pousser	0	0.20	élevé	0.20 (Doc2)
isolement	0	0.20	élevé	0.20 (Doc2)
troubler	0	0.20	élevé	0.20 (Doc2)
sommeil	0	0.20	élevé	0.20 (Doc2)

On observe que le terme *anxiété*, bien que présent dans les deux documents, reçoit un poids plus faible car il est moins discriminant. En revanche, des termes spécifiques à chaque document, comme *isolement* ou *emploi*, reçoivent un poids plus élevé, car ils permettent de mieux distinguer les préoccupations exprimées.

— Représentation vectorielle des mots (*Word Embeddings*) : Désigne un ensemble de techniques permettant de représenter les mots sous forme de vecteurs continus dans un espace à plusieurs dimensions, et apprenant à structurer l'espace lexical en rapprochant les termes partageant des contextes similaires [98].

Parmi les techniques les plus connues, on retrouve :

1. Word2Vec [27], qui repose sur deux architectures :
   — Continuous Bag-of-Words (CBOW) : prédit un mot cible w_t en fonction de son contexte $(w_{t-n}, ..., w_{t-1}, w_{t+1}, ..., w_{t+n})$.
   — Skip-gram : prédit le contexte d'un mot donné.
   Dans CBOW, la fonction objectif est définie par :

$$J = -\sum_{t=1}^{T} \log P(w_t | w_{t-n}, ..., w_{t+n})$$

   où la probabilité est modélisée par la fonction softmax :

$$P(w_O|w_I) = \frac{\exp(v_{w_O}^T v_{w_I})}{\sum_{w \in V} \exp(v_w^T v_{w_I})}$$

2. GloVe (Global Vectors for Word Representation), qui s'appuie sur les statistiques globales de cooccurrence des mots pour apprendre une représentation dense.
   La fonction d'optimisation utilisée est :

$$J = \sum_{i,j} f(X_{ij})(v_i^T v_j + b_i + b_j - \log X_{ij})^2$$

   où X_{ij} représente le nombre de cooccurrences entre les mots w_i et w_j.

3. FastText, une extension de Word2Vec qui utilise des sous-mots (n-grammes de caractères) pour représenter chaque mot, permettant une meilleure prise en charge des langues morphologiquement riches.

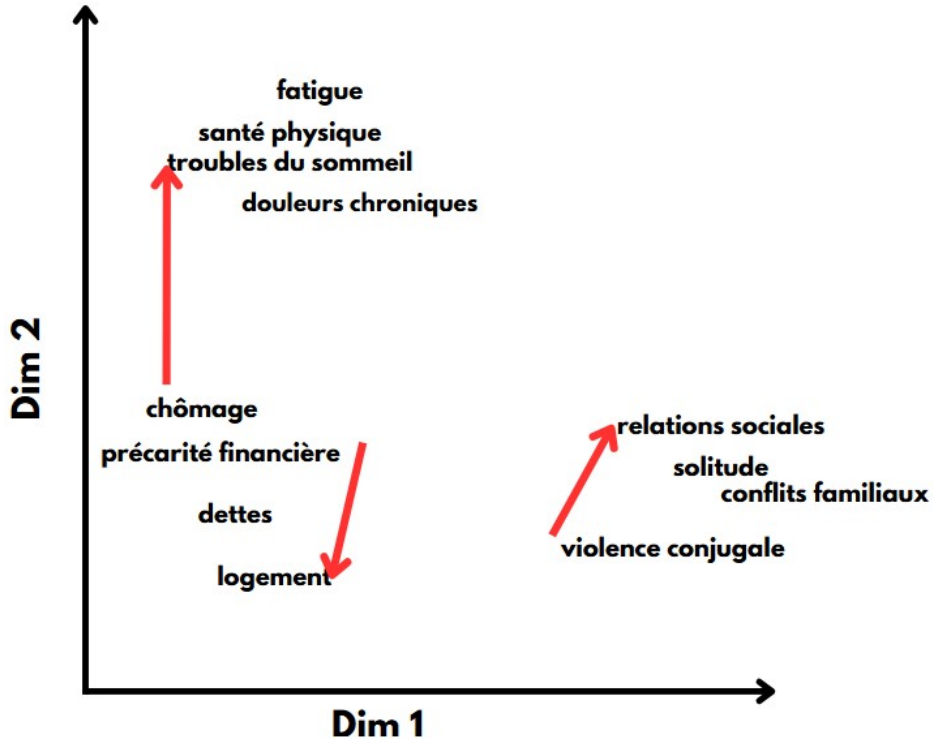


FIGURE 4. Illustration de la structuration sémantique des mots à l'aide des word embeddings

La figure 4 met en évidence plusieurs exemples illustratifs de relations, par exemple le lien entre le chômage et les troubles du sommeil :

$$v(\text{chômage}) + v(\text{stress}) \approx v(\text{troubles du sommeil})$$

Le chômage entraîne souvent du stress et de l'incertitude, ce qui peut perturber le sommeil et mener à des troubles chroniques.

— N-grams : une séquence continue de n éléments (généralement des mots ou des caractères) qui apparaissent dans le même ordre dans un texte donné. En capturant des séquences de mots voisins, les n-grams permettent de conserver un certain degré important de contexte [28]. Cette technique est couramment utilisée pour améliorer les performances dans des applications sensibles aux associations lexicales.
Exemple d'application des N-Grams : "Je ressens une grande solitude depuis la perte de mon emploi."

Unigrammes (1-gram) :

"Je", "ressens", "une", "grande", "solitude",
"depuis", "la", "perte", "de", "mon", "emploi"

Bigrammes (2-grams) :

"Je ressens", "ressens une", "une grande", "grande solitude",
"solitude depuis", "depuis la", "la perte", "perte de",
"de mon", "mon emploi"

Trigrammes (3-grams) :

"Je ressens une", "ressens une grande", "une grande solitude",
"grande solitude depuis", "solitude depuis la", "depuis la perte",
"la perte de", "perte de mon", "de mon emploi"

— 1-gram (BoW) : Détecte les termes clés (*"solitude"*, *"emploi"*), mais sans lien entre eux.
— 2-gram et 3-gram : Permettent de repérer des expressions significatives et d'améliorer la compréhension des préoccupations.
— N-grammes syntaxiques : Peuvent être utilisés pour analyser les relations cause-effet (ex. *"perte de mon emploi"* → *"solitude"*).

2.4. Techniques de réduction de la dimensionnalité

La réduction de dimension est une méthodes fondamentale en sciences de la donnée et apprentissage automatique. Elle vise à diminuer le nombre de variables tout en préservant une part importante de l'information [30]. Cette démarche permet de réduire la complexité des modèles, d'améliorer leur efficacité computationnelle ou encore de réduire le bruit, tout en maintenant les performances des algorithmes. Les techniques de réduction de la dimensionnalité sont choisies en fonction des caractéristiques des données et des objectifs analytiques. On distingue généralement deux approches principales : une réduction fondée sur la sélection de caractéristiques [94], qui vise à conserver les variables les plus informatives du jeu de données initial. Et une réduction fondée sur la transformation des données, qui consiste à projeter les données dans un nouvel espace de dimensions réduites, souvent à l'aide de combinaisons linéaires ou non linéaires des variables initiales [29].

2.4.1. Réduction basée sur la sélection de caractéristiques

Cette approche consiste à identifier et à retenir les caractéristiques les plus pertinentes parmi l'ensemble des variables décrivant le phénomène étudié [95]. L'objectif est de conserver uniquement les variables les plus influentes tout en éliminant celles qui sont redondantes ou non pertinentes.

Soit un ensemble initial de caractéristiques $\mathcal{F} = \{f_1, f_2, ..., f_N\}$ de taille N. Chaque caractéristique f_i représente une variable explicative observée sur un ensemble de m exemples, et $Y = (y_1, y_2, ..., y_m)$ désigne le vecteur des sorties associées.

La sélection de caractéristiques vise à trouver un sous-ensemble optimal $\mathcal{F}^* \subset \mathcal{F}$ de taille $M < N$ tel que :

$$\mathcal{F}^* = \arg\max_{\mathcal{Z} \subset \mathcal{F}} \text{Ev}(\mathcal{Z}) \tag{4}$$

où $\text{Ev}(\mathcal{Z})$ est une fonction d'évaluation mesurant la pertinence de l'ensemble $\mathcal{Z}$, elle peut être évaluée avec :

— La corrélation entre la i^e caractéristique f_i (c'est-à-dire la i^e colonne de la matrice des données) et la variable cible Y :

$$C(f_i) = \frac{\sum_{k=1}^{m}(x_{ki} - \bar{x}_i)(y_k - \bar{y})}{\sqrt{\sum_{k=1}^{m}(x_{ki} - \bar{x}_i)^2}\sqrt{\sum_{k=1}^{m}(y_k - \bar{y})^2}} \tag{5}$$

— Le critère de Fisher :

$$F(f_i) = \frac{\sum_{c=1}^{C} n_c(\bar{x}_c - \bar{x})^2}{\sum_{c=1}^{C} n_c\sigma_c^2} \tag{6}$$

Dans cette section, on évoque trois méthodes de sélection de caractéristiques :

— Méthodes Filter : Ces méthodes sélectionnent les caractéristiques en fonction de critères statistiques indépendants du modèle d'apprentissage. On peut citer le test du χ^2, les informations mutuelles et le score de Fisher [94].

— Méthodes *Wrapper* : Elles évaluent l'impact des caractéristiques en testant différents sous-ensembles via un modèle d'apprentissage [95]. La Table 2 présente une synthèse des principales méthodes *Wrapper* utilisées en sélection de caractéristiques.

Table 2. Synthèse des méthodes *Wrapper* en sélection de caractéristiques

Méthode Wrapper	**Principe**
SFS (Sequential Forward Selection)	Ajoute les caractéristiques une par une en maximisant la performance du modèle.
SBS (Sequential Backward Selection)	Supprime les caractéristiques une par une en minimisant la perte de performance.
***Stepwise* (Méthode bidirectionnelle)**	Combine SFS et SBS en ajoutant ou supprimant les caractéristiques selon un critère d'amélioration.
Recherche exhaustive	Teste toutes les combinaisons possibles de caractéristiques pour trouver l'optimum.
Méthodes heuristiques (Algorithmes génétiques)	Utilisent des mécanismes stochastiques pour explorer l'espace des solutions et optimiser la sélection.

— Méthodes Embedded : Ces méthodes intègrent la sélection de caractéristiques directement dans le processus d'apprentissage du modèle. On retrouve :

  — LASSO (Régression L1) Utilise une régularisation L_1 pour imposer des coefficients nuls à certaines caractéristiques :

$$\min_{\beta} ||Y - X\beta||^2 + \lambda||\beta||_1 \tag{7}$$

  — Forêts Aléatoires (Random Forest Feature Importance) : Un arbre de décision évalue l'importance des caractéristiques via la réduction de l'impureté de Gini.

2.4.2. Réduction basée sur une transformation des données

Également appelée extraction de caractéristiques, cette approche consiste à remplacer l'ensemble initial de données par un nouvel ensemble réduit, construit à partir des caractéristiques initiales. Cette transformation permet de représenter les données dans un espace de dimension inférieure tout en conservant les informations essentielles.

— L'Analyse en Composantes Principales (PCA) est une méthode linéaire qui transforme les données en un espace de variables non corrélées appelées composantes principales [29]. Ces composantes sont ordonnées de manière à ce que la première composante principale capture la plus grande variance des données, suivie par la deuxième, et ainsi de suite.
Soit un ensemble de données $\mathbf{X} = \{x_1, x_2, ..., x_m\}$ de taille m, où chaque observation $x_i \in \mathbb{R}^N$ est un vecteur de N caractéristiques. L'ACP repose sur la diagonalisation de la matrice de covariance $\mathbf{S}$:

$$\mathbf{S} = \frac{1}{m} \sum_{i=1}^{m} (\mathbf{x}_i - \bar{\mathbf{x}})(\mathbf{x}_i - \bar{\mathbf{x}})^T \tag{8}$$

où $\bar{\mathbf{x}}$ est la moyenne des observations.
Les composantes principales sont données par les valeurs propres et vecteurs propres de $\mathbf{S}$:

$$\mathbf{SW} = \mathbf{W}\Lambda \tag{9}$$

où $\mathbf{W}$ contient les vecteurs propres triés selon l'ordre décroissant des valeurs propres Λ.

— Analyse Discriminante Linéaire (LDA) est une méthode supervisée qui, au contraire de la PCA, cherche à maximiser la séparabilité entre les classes. En projetant les données dans un espace de dimension inférieure, elle veille à conserver les différences qui existent entre catégories [30]. L'objectif est donc de trouver une projection qui maximise le critère de Fisher décrit préalablement.

— Projection Aléatoire est une méthode qui consiste à projeter les données dans un espace de dimension inférieure en appliquant une transformation linéaire basée sur une matrice aléatoire [32].
Elle repose sur la propriété de Johnson-Lindenstrauss qui garantit que les distances euclidiennes sont approximativement préservées lorsqu'on projette les données dans un sous-espace de dimension réduite M :

$$\mathbf{Z} = \mathbf{XR} \tag{10}$$

où $\mathbf{R} \in \mathbb{R}^{N \times M}$ est une matrice aléatoire gaussienne.

— T-Distributed Stochastic Neighbor Embedding (t-SNE) est une méthode non linéaire utilisée principalement pour la visualisation des données en réduisant leur dimensionnalité à 2 ou 3 dimensions tout en préservant les relations locales entre les points de données [31], ce qui la rend très utile pour identifier des structures complexes et des regroupements.
t-SNE cherche à minimiser la divergence de Kullback-Leibler entre les distributions de probabilités dans l'espace initial et l'espace réduit :

$$KL(P\|Q) = \sum_i \sum_j p_{ij} \log \frac{p_{ij}}{q_{ij}} \tag{11}$$

où :

— p_{ij} représente la similarité entre les points x_i et x_j dans l'espace original.
— q_{ij} représente la similarité entre les points projetés dans l'espace réduit.

2.5. Méthodes classiques d'apprentissage automatique

Les premières approches de classification s'appuient à la fois sur des méthodes purement statistiques et sur des algorithmes de l'apprentissage automatique[3]. D'un côté, les méthodes statistiques reposent sur des hypothèses probabilistes explicites quant à la distribution des données tel que la normalité des variables ou l'indépendance conditionnelle. D'un autre côté, certains algorithmes de l'apprentissage profond adoptent plutôt une approche fondée sur la recherche algorithmique et l'optimisation, sans nécessairement imposer de distribution probabiliste spécifique.

— Les machines à vecteurs de support (SVM) sont des modèles de classification supervisée qui cherchent l'hyperplan optimal séparant les différentes classes dans un espace de caractéristiques [33]. Pour une séparation linéaire, SVM tente de maximiser la marge [38] c'est-à-dire la distance entre les points de données des différentes classes. La formulation mathématique de ce problème d'optimisation est :

$$\min_{\mathbf{w},b} \quad \frac{1}{2}\|\mathbf{w}\|^2 \tag{12}$$

sous la contrainte :

$$y_i(\mathbf{w} \cdot \mathbf{x}_i + b) \geq 1, \quad \forall i \tag{13}$$

où :
— $\mathbf{w}$ est le vecteur des poids définissant l'orientation de l'hyperplan,
— b est le biais ajustant la position de l'hyperplan,
— $\mathbf{x}_i$ représente les vecteurs de caractéristiques,
— $y_i \in \{+1, -1\}$ sont les labels des classes.

Pour les cas non-linéairement séparables, les SVM génèrent des noyaux qui permettent de recourir à des transformations de données dans des espaces features de plus grande dimension où la séparation linéaire devient possible [38].

$$K(x_i, x_j) = \phi(x_i) \cdot \phi(x_j) \tag{14}$$

où $\phi(x)$ est une fonction de transformation des données.
Quelques exemples de noyaux courants :
— Noyau linéaire : $K(x_i, x_j) = x_i^T x_j$
— Noyau polynomial : $K(x_i, x_j) = (x_i^T x_j + c)^d$
— Noyau gaussien (RBF) :

$$K(x_i, x_j) = \exp\left(-\frac{\|x_i - x_j\|^2}{2\sigma^2}\right) \tag{15}$$

— L'algorithme des k-plus proches voisins (k-Nearest Neighbors, k-NN) implémente une méthode non paramétrique utilisée, l'algorithme assigne une classe à une observation en fonction des classes majoritaires parmi ses k plus proches voisins dans l'espace des caractéristiques [36]. La proximité est généralement mesurée par une distance métrique, telle que la distance euclidienne. Elle est définie entre deux vecteurs $\mathbf{X} = (x_1, x_2, \ldots, x_n)$ et $\mathbf{Y} = (y_1, y_2, \ldots, y_n)$ dans un espace à n dimensions comme :

$$\text{Dist}_E(\mathbf{X}, \mathbf{Y}) = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2} \tag{16}$$

On peut aussi faire appel à la distance Cosinus et la distance Minkowski qui peuvent être très utiles dans de nombreuses situations.

Les avantages de k-NN incluent sa simplicité de mise en œuvre et son efficacité pour des ensembles de données de petite à moyenne taille. Toutefois, il est sensible aux dimensions élevées et peut être coûteux en termes de calcul pour de grands ensembles de données.

— XGBoost (Extreme Gradient Boosting) est une implémentation avancée de l'algorithme de boosting par gradient qui a démontré des performances remarquables au niveau des applications réelles [35]. XGBoost fonctionne en construisant une séquence d'arbres de décision, chaque nouvel arbre corrigeant les erreurs des arbres précédents. L'objectif est de minimiser une fonction de perte définie par :

$$\mathcal{L}(\phi) = \sum_i l(y_i, \hat{y}_i) + \sum_k \Omega(f_k) \tag{17}$$

où

- l est la fonction de perte différenciable,
- Ω est la régularisation des arbres f_k,
- et ϕ représente les paramètres de tous les arbres.

Les avantages de XGBoost incluent sa capacité à gérer des données manquantes, son efficacité de calcul grâce à l'optimisation matérielle, et ses capacités de régularisation qui réduisent le surapprentissage.

— Naïve Bayes : ces classificateurs sont basés sur l'application du théorème de Bayes avec une forte hypothèse d'indépendance entre les caractéristiques [35].
La formule de base du classificateur Naïve Bayes est :

$$P(C_k \mid \mathbf{x}) = \frac{P(C_k) \prod_{i=1}^{n} P(x_i \mid C_k)}{P(\mathbf{x})} \tag{18}$$

où

- $P(C_k \mid \mathbf{x})$ est la probabilité postérieure de la classe C_k donnée les caractéristiques $\mathbf{x}$,
- $P(C_k)$ est la probabilité préalable de la classe C_k,

— $P(x_i \mid C_k)$ est la probabilité conditionnelle de x_i donnée C_k,
— $P(\mathbf{x})$ est la probabilité des caractéristiques $\mathbf{x}$.

Les avantages de Naïve Bayes incluent sa simplicité et son efficacité computationnelle [36], ce qui le rend particulièrement adapté aux grandes bases de données textuelles.

3. L'apprentissage profond en TALN

Les pratiques de l'apprentissage traditionnel ont montré leur efficacité dans la résolution de divers problèmes. Cependant, ces modèles rencontrent des difficultés lorsqu'ils sont appliqués à des données complexes et non structurées [33]. À la différence de ces méthodes, les modèles d'apprentissage profond peuvent apprendre à partir de données brutes sans recours à l'extraction manuelle des caractéristiques [6]. Cette intégration a donc permis d'améliorer le traitement automatique du langage naturel, notamment dans la capture des structures et des relations sémantiques. Cette section décrit les principes fondamentaux de l'apprentissage profond, ses principales architectures [20] et son impact sur les applications du traitement automatique du langage naturel.

3.1. Apprentissage des séquences en TALN par les RNN et LSTM

L'apprentissage profond repose sur les réseaux de neurones artificiels (ANN), qui s'inspirent en partie de la structure du cerveau humain [6]. Contrairement aux algorithmes classiques, les réseaux neuronaux profonds (DNN) utilisent plusieurs couches cachées pour capturer les relations complexes et extraire des représentations hiérarchiques des données. Le tableau 3 résume et offre une vue d'ensemble des éléments essentiels en apprentissage profond[37].

Les réseaux neuronaux récurrents (RNN) sont conçus pour modéliser des séquences temporelles et capturer des relations de dépendance dans les données [34]. Corrélativement aux modèles traditionnels qui traitent l'entrée de manière isolée, leur fonctionnement s'explique par l'utilisation d'une mémoire partagée pour transmettre de l'information d'une étape de la séquence à une autre. Cela leur permet d'apprendre les dépendances locales dans les données séquentielles. Cependant, les RNN classiques rencontrent des difficultés à modéliser les dépendances à long terme dans les séquences. Ce problème résulte de la disparition ou de l'explosion des gradients, rendant l'entraînement des séquences longues moins efficace. À cet effet les modèles Long Short-Term Memory (LSTM) ont été développés pour surmonter ces limitations [11], ces modèles étendent les RNN en introduisant des cellules mémoire capables de stocker et de gérer les informations sur de longues périodes.

3.2. Introduction aux transformers

Les *Transformers* sont des modèles de transduction séquence à séquence qui reposent sur un mécanisme d'attention pour traiter les données en parallèle[10]. Contrairement aux réseaux récurrents (RNN) [34] et aux réseaux de neurones convolutifs (CNN) que cet article a brièvement évoqué, l'architecture des *Transformers* est composée d'un encodeur (figure 5) et d'un décodeur, chacun structuré en plusieurs blocs identiques. L'encodeur transforme la séquence d'entrée en une représentation latente tandis que le décodeur génère la sortie en utilisant cette représentation ainsi que les sorties précédemment générées. L'absence de connexions récurrentes dans cette architecture permet d'exploiter une parallélisation efficace des calculs, réduisant ainsi le temps d'entraînement sur de grands ensembles de données.

Le principe fondamental des *Transformers* repose sur le mécanisme d'auto-attention, qui pondère les relations entre les éléments d'une séquence en fonction de leur importance contextuelle [20]. Contrairement aux architectures récurrentes, qui traitent les séquences en tenant compte uniquement des éléments précédents, l'auto-attention permet d'établir des relations entre tous les éléments indépendamment de leur position.

Pour ce faire, chaque élément d'entrée est projeté sous la forme de trois vecteurs distincts : les requêtes (Q), les clés (K) et les valeurs (V). L'attention est obtenue en calculant le produit scalaire des requêtes et des clés [112], puis

TABLE 3. Concepts fondamentaux de l'apprentissage profond et leur déscription.

Concepts	Description
Perceptron multicouche (MLP)	— Un réseau composé de couches entièrement connectées constitue la base des réseaux profonds utilisés en apprentissage profond, mais il reste limité pour le traitement des séquences et des données non structurées.
Fonctions d'activation	— Elles introduisent la non-linéarité, essentielle pour capturer des relations complexes, et permettent une meilleure transformation des données à chaque couche, avec des exemples courants tels que ReLU, tanh et sigmoid.
Backpropagation	— Algorithme d'optimisation basé sur la règle de la chaîne, il ajuste les poids d'un réseau neuronal en propageant les erreurs depuis la sortie jusqu'aux couches précédentes, permettant ainsi une mise à jour efficace des paramètres via la descente de gradient.
Optimisation	— Les algorithmes d'optimisation tels que SGD et Adam améliorent la convergence en minimisant la fonction de coût : SGD repose sur la descente de gradient stochastique, tandis qu'Adam combine cette approche avec des taux d'apprentissage adaptatifs, facilitant ainsi la gestion des gradients instables (explosion ou disparition). D'autres techniques, comme le momentum, accélèrent la convergence en conservant une fraction des mises à jour précédentes, tandis que RMSprop ajuste dynamiquement le taux d'apprentissage pour stabiliser l'optimisation, notamment dans les réseaux récurrents.

en appliquant une normalisation par la dimension des représentations, avant de pondérer les valeurs correspondantes :

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V \tag{19}$$

où d_k est la dimension des clés. L'opération de *softmax* permet de normaliser les scores d'attention, garantissant que leur somme est égale à 1.

Une extension de ce mécanisme est le multi-head attention, qui applique plusieurs instances de l'auto-attention en parallèle [103]. Chaque tête effectue une transformation linéaire indépendante sur les entrées avant d'appliquer l'auto-attention. Les sorties des différentes têtes sont ensuite concaténées et projetées à nouveau pour générer une représentation enrichie :

$$\text{MultiHead}(Q, K, V) = \text{Concat}(\text{head}_1, ..., \text{head}_h)W^O \tag{20}$$

où chaque tête d'attention suit la formulation :

$$\text{head}_i = \text{Attention}(QW_i^Q, KW_i^K, VW_i^V) \tag{21}$$

avec W_i^Q, W_i^K, W_i^V représentant les matrices de projection spécifiques à chaque tête. L'intérêt de l'attention multi-tête est de permettre au modèle de capturer plusieurs types de relations contextuelles en parallèle, ce qui améliore sa capacité de généralisation sur des séquences complexes.

Outre le mécanisme d'auto-attention, l'architecture des *Transformers* repose sur plusieurs éléments structurants visant à stabiliser l'apprentissage et à améliorer la représentation des séquences :

— Normalisation des couches et connexions résiduelles pour faciliter l'entraînement des réseaux profonds, chaque sous-couche du Transformer est entourée de connexions résiduelles et de normalisations de couches. Deux variantes existent :

  — Post-layer normalization (Post-LN) : la normalisation est appliquée après les connexions résiduelles.

  — Pre-layer normalization (Pre-LN) : la normalisation est appliquée avant chaque sous-couche.

— Encodage positionnel : contrairement aux modèles récurrents, les *Transformers* ne possèdent pas de structure séquentielle implicite. Pour conserver l'ordre des éléments d'une séquence, un ncodage positionnel est ajouté aux représentations des entrées. Il peut être calculé à l'aide de fonctions sinusoïdales ou appris comme un ensemble de paramètres du modèle :

$$P(i, 2j) = \sin\left(\frac{i}{10000^{2j/d}}\right), \quad P(i, 2j+1) = \cos\left(\frac{i}{10000^{2j/d}}\right) \tag{22}$$

où i est la position dans la séquence et j l'indice de la dimension. Cet encodage permet au modèle de différencier les positions des tokens sans recourir à des mécanismes récurrents.

L'architecture des *Transformers* peut être utilisée de trois manières principales selon la tâche ciblée :

— Encoder-Decoder : L'architecture complète du Transformer, telle qu'introduite initialement, est utilisée. Cette configuration est particulièrement adaptée aux modèles de type sequence-to-sequence, comme la traduction automatique.

— Encodeur uniquement : Seule la partie encodeur est exploitée, produisant une représentation compacte de la séquence d'entrée. Cette approche est couramment employée pour les tâches de classification ou d'étiquetage de séquences.

— Décodeur uniquement : Seule la partie décodeur est utilisée, en supprimant le module d'attention croisée entre encodeur et décodeur. Cette configuration est souvent appliquée à la génération de séquences, comme les modèles de langage autoregressifs.

Avec les *Transformers*, les recherches se sont orientées vers comment tirer parti de ces puissantes architectures au service de besoins spécifiques tout en maximisant les ressources de leur apprentissage. Cela a conduit au développement de techniques comme le fine-tuning [12] et SetFit [13], qui adaptent des modèles pré-entraînés à des tâches spécifiques, même avec un nombre limité d'exemples annotés.

3.2.1. Fine-tuning

Le fine-tuning, technique clé dans le domaine de l'apprentissage profond, consiste à ajuster un modèle pré-entraîné sur un grand corpus de données génériques à une tâche donnée à partir d'un ensemble complet de données utiles [12]. Ce processus en définitive comporte deux étapes principales : la première consiste à préentraîner le modèle sur un

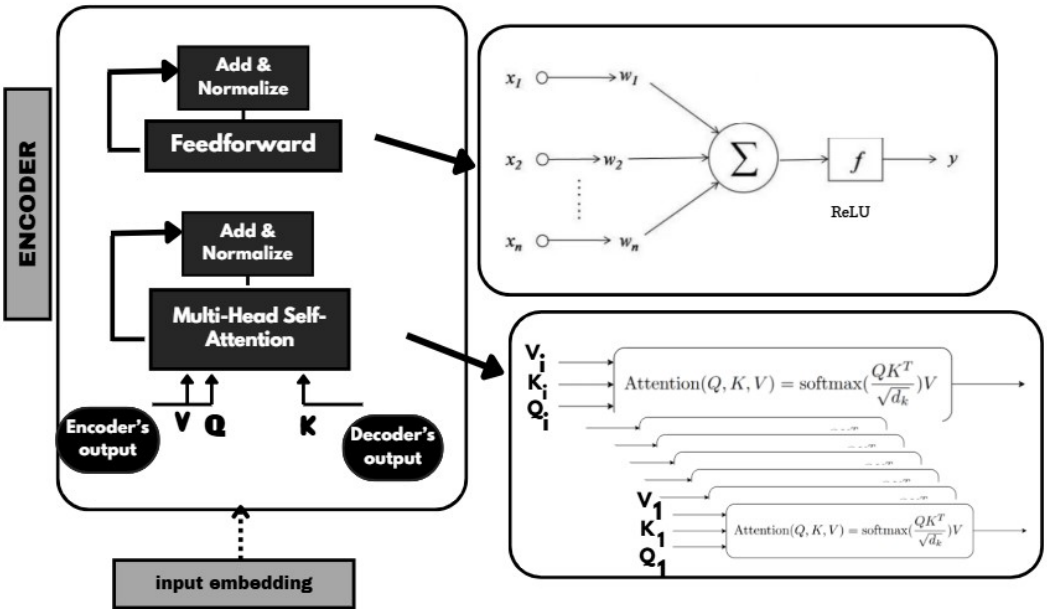


FIGURE 5. Architecture du transformer : fonctionnement de l'encodeur avec mécanisme multi-head self-attention

large corpus de données génériques défini afin de le doter de représentations riches et variées des données textuelles. La deuxième étape ajuste les poids du modèle pour s'adapter aux spécificités d'un ensemble de données ciblé.

Le fine-tuning dans ce cas, permet d'exploiter pleinement les données spécifiques, maximisant ainsi l'information disponible pour le modèle, et une grande flexibilité pour s'adapter à diverses tâches spécifiques. Cependant, cette technique présente des inconvénients notables tels que le coût computationnel élevé et l'éventuel risque de surapprentissage.

Une étude intéressante, "Fine-Tuning Large Neural Language Models for Biomedical Natural Language Processing" [97], explore les défis et les bénéfices de cette approche dans un contexte biomédical. Les auteurs démontrent que le fine-tuning améliore significativement la performance des modèles préentraînés sur des jeux de données spécialisés, mais que cette amélioration varie en fonction de la taille du modèle et des stratégies d'optimisation employées. En particulier, ils montrent que les modèles, peuvent être instables si leur entraînement n'est pas calibré. Pour pallier ce problème, l'étude recommande des ajustements spécifiques comme le gel des premières couches des modèles et l'utilisation de taux d'apprentissage différenciés par couche, ce qui stabilise le processus d'adaptation et améliore la généralisation des modèles sur des corpus médicaux limités.

3.2.2. Le set-fit

Pour contourner les limites du fine-tuning, notamment son coût computationnel accru, le SetFit (Set-based Fine-Tuning) propose une alternative basée sur la sélection d'un sous-ensemble de données, préalablement choisi en se basant sur des méthodes statistiques ou des techniques d'apprentissage actif pour garantir que l'ensemble retenu capture la diversité des exemples présents dans les données d'origine [13]. Cette approche est particulièrement adapté lorsque les ressources sont limitées ou lorsqu'un résultat rapide est nécessaire, mais elle reste conditionnée à une sélection pertinente de données représentatives qui évite de perdre des informations pertinentes, ce qui s'avère difficile dans des jeux de données complexes.

4. La modélisation thématique en TALN

La modélisation thématique est une approche du traitement automatique du langage naturel (TALN) qui repose sur l'apprentissage non supervisé pour analyser de grands corpus textuels. Contrairement aux méthodes supervisées

nécessitant des données annotées, elle permet d'identifier automatiquement les thèmes sous-jacents d'un ensemble de documents en inférant des regroupements basés sur la distribution et la cooccurrence des mots [41]. Cette technique repose sur des modèles probabilistes ou algébriques qui détectent des structures latentes, facilitant ainsi l'organisation et l'analyse de textes sans classification préalable [40]. Elle est utilisée pour pour repérer les tendances émergentes et les structures latentes de corpus textuels [41].
Dans un contexte où les données textuelles non étiquetées sont de plus en plus abondantes, la modélisation thématique constitue un outil pour extraire les régularités lexicales et organiser des contenus hétérogènes [43]. En analysant la distribution des termes et en regroupant les textes selon les sujets qu'ils abordent, ces modèles permettent de structurer les données et d'identifier des tendances émergentes sans intervention humaine directe. Dans le cadre du TALN, la modélisation thématique s'inscrit comme un complément aux techniques supervisées de classification. Elle permet d'explorer et de catégoriser les données sans contrainte d'étiquetage préalable, tout en facilitant l'analyse sémantique des textes [40].

4.1. Approches statistiques de modélisation thématique

La modélisation thématique repose sur des méthodes statistiques permettant d'inférer automatiquement la structure latente d'un corpus textuel en identifiant les thèmes qui le composent. Ces méthodes s'appuient sur des techniques probabilistes et algébriques pour analyser la distribution des mots et leur cooccurrence dans les documents, facilitant ainsi l'organisation et l'interprétation de grands ensembles de données textuelles. Parmi les méthodes les plus courantes, on distingue :

Latent dirichlet allocation (LDA) : Ce modèle probabiliste part de l'hypothèse que chaque document est composé d'un mélange de plusieurs thèmes et que chaque thème est caractérisé par une distribution spécifique de mots [5]. LDA est largement utilisé pour son efficacité à capturer des structures thématiques dans des corpus volumineux. Cependant, il requiert une calibration précise des hyperparamètres, tels que le nombre de thèmes, et peut être sensible à la qualité des données textuelles[45].

LDA modélise la probabilité conjointe des mots w, des thèmes z, et des distributions de thèmes θ comme suit :

$$P(w, z, \theta \mid \alpha, \beta) = P(\theta \mid \alpha) \prod_{n=1}^{N} P(z_n \mid \theta) P(w_n \mid z_n, \beta)$$

où :

— α : hyperparamètre pour la distribution de Dirichlet sur θ (distribution des thèmes dans les documents),

— β : hyperparamètre pour la distribution de Dirichlet sur les mots,

— z_n : thème assigné au mot w_n,

— w_n : mot dans le document.

L'estimation des paramètres est généralement effectuée via des méthodes comme la variational inference ou le Gibbs sampling.
Latent semantic analysis (LSA) : Basée sur la décomposition en valeurs singulières (SVD), cette technique réduit la dimensionnalité des données en capturant les relations entre termes et documents[44]. Bien qu'elle soit rapide et simple à mettre en œuvre, elle souffre d'une incapacité à modéliser la probabilité des thèmes, ce qui peut limiter son interprétabilité.

LSA utilise la décomposition en valeurs singulières (SVD) de la matrice des termes et documents A :

$$A = U\Sigma V^T$$

où :

— $A \in \mathbb{R}^{m \times n}$: matrice des fréquences des mots (m est le nombre de mots, n le nombre de documents),

— U : matrice orthogonale représentant les mots,

— Σ : matrice diagonale contenant les valeurs singulières,

— V^T : matrice orthogonale représentant les documents.

Pour réduire la dimensionnalité, seules les k-plus grandes valeurs singulières de Σ sont retenues, ce qui revient à approximer A par :

$$A_k = U_k \Sigma_k V_k^T$$

Non-negative matrix factorization (NMF) : En décomposant la matrice des fréquences des mots en deux matrices non négatives, NMF permet de représenter les thèmes sous forme de regroupements de mots liés. Cette méthode est particulièrement adaptée aux données textuelles contenant des valeurs exclusivement positives, mais elle peut manquer de robustesse face aux variations linguistiques [45].

La décomposition par NMF s'exprime comme suit :

$$A \approx WH$$

où :

— $A \in \mathbb{R}_+^{m \times n}$: matrice des fréquences des mots,

— $W \in \mathbb{R}_+^{m \times k}$: matrice des contributions des mots aux thèmes,

— $H \in \mathbb{R}_+^{k \times n}$: matrice des contributions des thèmes aux documents,

— k : nombre de thèmes.

L'objectif est de minimiser une fonction de coût, souvent basée sur la norme Frobenius :

$$\min_{W,H} \|A - WH\|_F^2$$

sous les contraintes $W \geq 0$ et $H \geq 0$.

5. Optimisation des hyperparamètres en modélisation thématique

La performance et l'interprétabilité des modèles de modélisation thématique, tels que LDA, NMF et LSA, reposent sur le choix des hyperparamètres ainsi que sur les outils et métriques d'optimisation [46]. Ces outils facilitent l'optimisation des modèles et contribuent à assurer la cohérence des thèmes extraits.

5.1. Nombre de thèmes (K)

Le choix du nombre de thèmes (K) est une décision déterminante. Plusieurs outils et métriques peuvent être utilisés pour identifier une valeur optimale :

— **Perplexité** : Cette mesure probabiliste est souvent utilisée avec LDA [5]. Elle évalue dans quelle mesure un modèle généré est capable de prédire des données de test. Une perplexité plus faible indique une meilleure généralisation :

$$\text{Perplexité} = \exp\left(-\frac{\sum_{d\in D}\log P(d)}{\sum_{d\in D} N_d}\right)$$

où $P(d)$ est la probabilité du document d donnée par le modèle, et N_d est la longueur du document.

— **Score de cohérence des thèmes** : Cette métrique, particulièrement utile pour l'évaluation humaine, mesure la similarité sémantique entre les mots d'un même thème [47, 56]. Les outils comme Gensim fournissent des fonctions intégrées pour calculer la cohérence.

— **Validation croisée** : En divisant les données en ensembles d'entraînement et de test, des outils comme scikit-learn permettent d'évaluer l'impact de K sur la performance globale [56].

5.2. Taille des n-grammes

La configuration des n-grammes est importante pour capturer des expressions multi-mots pertinentes. Les outils suivants sont souvent utilisés pour tester différentes tailles de n-grammes :

— CountVectorizer et TfidfVectorizer de scikit-learn : Ils permettent de générer des vecteurs pour des n-grammes spécifiques et de comparer leur impact sur les résultats.

— NLTK et spaCy : Ces bibliothèques aident à prétraiter et générer des n-grammes avant leur passage dans le modèle.

Table 4. Hyperparamètres spécifiques aux modèles de modélisation thématique

Modèle	Hyperparamètres et outils
Latent dirichlet allocation (LDA)	— **Grid search** : Exploration des combinaisons de α et β avec des outils comme Gensim ou hyperopt. — **Auto-tuning intégré** : Ajustement automatique des paramètres avec des frameworks tels que *Mallet*.
Non-negative matrix factorization (NMF)	— **Initialisation** : Test des méthodes comme random ou nndsvd (Non-negative Double Singular Value Decomposition) via *scikit-learn*. — **Nombre d'itérations** : Analyse des courbes de convergence produites avec Matplotlib.
Latent semantic analysis (LSA)	— **Seuil de SVD** : Utilisation de TruncatedSVD dans scikit-learn pour tester différentes dimensions et visualiser les performances.

6. Introduction aux modèles de langage massif (LLMs)

La classification à l'aide des modèles d'apprentissage profond et la modélisation thématique[46], bien qu'efficaces dans de nombreux scénarios, présentent des limitations dans des contextes où les données évoluent rapidement. Ces défis incluent l'apparition de nouvelles thématiques, la nécessité d'adapter en continu les catégories et l'absence fréquente d'annotations, réduisant ainsi l'efficacité des modèles traditionnels.
Un autre problème pertinent est celui du déséquilibre des classes [19], certaines catégories sont significativement sous-représentées dans la base de données, provoquant ainsi des biais d'apprentissage pour les algorithmes classiques qui sont incapables de généraliser de manière satisfaisante pour les catégories problématiques. Ces observations motivent l'utilisation de méthodes adaptées aux données non étiquetées, permettant un meilleur équilibre et ouvre donc de nouvelles perspectives aux les modèles de langage massif.

Les LLMs sont des modèles de l'IA basé sur l'apprentissage profond [4], capables de traiter et générer du texte avec un haut niveau de cohérence. Leur large pré-entraînement sur d'immenses corpus leur de s'adapter à diverses applications telles que la classification automatique, la traduction, la génération de texte, la synthèse de documents et l'assistance conversationnelle, tout en s'adaptant dynamiquement aux spécificités des données traitées [60].

6.1. Fonctionnement architecture et des LLMs

Les LLMs reposent sur l'architecture des *Transformers* [4], des mécanismes avancés d'attention et un entrainement sur d'énormes corpus de données. Pour mieux contextualiser ce processus, les LLMs, dont les plus populaires 'Claude', 'GPT' et leurs successeurs [4], passent par plusieurs étapes. La collecte de données consiste à rassembler un volume important de textes provenant de différentes sources (sites web, livres, articles, bases de données). Ces données sont ensuite nettoyées et normalisées, puis stocké dans une base NoSQL pour pouvoir les utiliser lors de l'entraînement.

Au cours de l'entraînement, le modèle apprend à comprendre la langue en réalisant de larges tâches non supervisées. Par exemple, il prédit le mot suivant dans une phrase (modélisation autorégressive) ou il remplit des mots masqués (modélisation masquée) afin de mieux cerner le contexte. Enfin, le modèle est affiné sur un ensemble de données plus restreint, mais spécialisé pour la tâche cible. Ces étapes permettent d'adapter le modèle aux caractéristiques des corpus traités.

La Figure 6 présente la composition générale des Grands Modèles de Langage, mettant en évidence les différentes couches qui structurent ces architectures avancées. Lorsqu'un texte est soumis en entrée, il passe par plusieurs transformations successives au sein du modèle. Les couches d'intégration assurent la conversion des tokens en représentations numériques exploitables par le réseau de neurones. Ensuite, les couches de rétroaction et récurrentes, comprenant plusieurs couches entièrement connectées, permettent au modèle de capturer des relations entre les éléments du texte, facilitant ainsi l'apprentissage de dépendances contextuelles à court et long terme. Les couches d'attention, cœur des architectures *Transformers*, jouent un rôle central en pondérant l'importance des différentes parties du texte, optimisant ainsi la compréhension et la génération de texte. À l'issue de ces traitements, le modèle produit une sortie textuelle, adaptée à la tâche spécifique (classification, génération, résumé, etc.)

6.1.1. Modèles auto-régressifs : le cas de GPT

Le modèle GPT (Generative Pre-trained Transformer) est un exemple typique de modèle auto-régressif [60]. Développée pour la première fois par OpenAI en 2018, la série GPT a introduit le concept fondamental de collecte de données, de pré-formation et de réglage fin des LLM. Dans ce même contexte, le modèle est entraîné à prédire le mot suivant dans une séquence donnée en s'appuyant uniquement sur les mots précédents. Ce processus suit une approche unidirectionnelle, où le contexte est construit à partir des données passées.

La probabilité d'une séquence de mots $(x_1, x_2, \ldots, x_T)$ est définie comme suit :

$$P(x_1, x_2, \ldots, x_T) = \prod_{t=1}^{T} P(x_t \mid x_1, x_2, \ldots, x_{t-1})$$

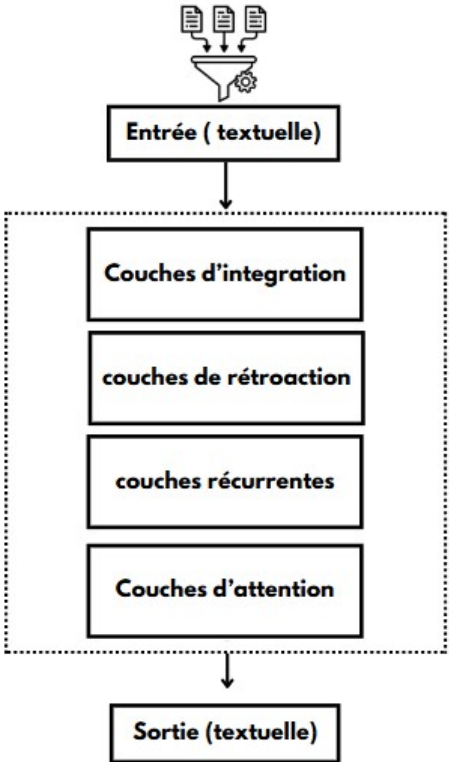


FIGURE 6. Composition générale des LLMS

où x_t est le mot à la position t, et la probabilité conditionnelle $P(x_t \mid x_1, x_2, \ldots, x_{t-1})$ est calculée à l'aide du mécanisme d'attention des *Transformers*.

Les versions récentes, dont GPT-4 ou GPT-o1, améliorent la précision des réponses et atténuent certains biais observés dans les versions antérieures. Grâce à ces améliorations, GPT étend ses capacités en compréhension et génération de texte.

6.1.2. Modèles bidirectionnels en TALN : le cas de BERT

Contrairement à GPT, BERT *(Bidirectional Encoder Representations from Transformers)*, développé par Google en 2018, repose sur une architecture bidirectionnelle permettant une meilleure compréhension contextuelle. Il utilise une architecture bidirectionnelle grâce à une tâche d'entraînement appelée masked language modeling (MLM) [11]. L'objectif est de prédire des mots masqués dans une séquence, permettant au modèle de comprendre les relations contextuelles dans les deux directions (passée et future).

Le modèle est entraîné sur des paires (x, y), où x représente une séquence contenant des mots masqués, et y les mots masqués à prédire. La fonction d'entraînement pour MLM peut être formulée comme suit :

$$\mathcal{L}_{MLM} = -\sum_{i \in M} \log P(x_i \mid x_{\text{masqué}})$$

où M est l'ensemble des positions masquées, et $P(x_i \mid x_{\text{masqué}})$ représente la probabilité prédite par le modèle pour le mot masqué à la position i.

Les mécanismes contextuels des *Transformers* permettent une compréhension très fine des relations entre les mots, grâce à leur capacité à attribuer des poids différents aux mots en fonction de leur pertinence dans un contexte donné.

Cette caractéristique est cruciale pour la classification thématique, où la compréhension des nuances sémantiques est essentielle, et a conduit à des avancées significatives dans les tâches de compréhension du langage.

6.1.3. Modèles encodeur-décodeur (Seq2Seq) : le cas de T5

Le modèle T5 (Text-to-Text Transfer Transformer), développé par Google Research, repose sur une architecture encodeur-décodeur permettant une transformation flexible des entrées textuelles en sorties adaptées à diverses tâches [85]. Contrairement aux modèles purement bidirectionnels (BERT) ou auto-régressifs (GPT), T5 adopte une approche entièrement générative tout en conservant une architecture Transformer standard. Son apprentissage est basé sur une approche guidée, où chaque phase d'entraînement repose sur une association explicite entre une entrée textuelle et la réponse attendue. Le texte d'entrée est converti en une représentation numérique et injecté dans la première partie du réseau. De son côté, la sortie attendue est prétraitée : elle est réajustée en décalant son alignement et en y insérant un marqueur initial. Ensuite, elle est transmise à la seconde partie du modèle, chargée de la génération. Durant l'apprentissage, le modèle cherche à maximiser la probabilité conditionnelle de la séquence cible donnée la séquence d'entrée, ce qui peut être formulé de la manière suivante :

$$P(y_1, y_2, \ldots, y_T \mid x_1, x_2, \ldots, x_N) = \prod_{t=1}^{T} P(y_t \mid y_1, \ldots, y_{t-1}, x_1, \ldots, x_N) \tag{23}$$

où x représente la séquence d'entrée et y la séquence de sortie générée. Cette formulation traduit le processus itératif de génération, où chaque élément de la sortie est prédit en fonction des éléments précédemment générés et du contexte global fourni par l'entrée.

L'utilisation d'un marqueur neutre permet d'assurer une structuration cohérente des données, tandis que l'approche employée favorise une convergence stable du modèle, optimisant ainsi son entraînement en mode supervisé et non supervisé.

6.1.4. Modèles adaptatifs et spécialisés

La généralisation des LLMs a motivé le développement de variantes spécialisées dans des domaines précis, et en particulier. L'objectif est d'exploiter des corpus de grande taille pour améliorer sensiblement les performances en traitement du langage naturel sur des tâches telles que la reconnaissance d'entités spécifiques, l'extraction de relations, la réponse à des questions spécialisées ou encore la génération de résumés scientifiques. Plusieurs modèles ont ainsi vu le jour, notamment :

— **BioBERT** : Dérivé du modèle BERT et pré-entraîné sur des bases de données biomédicales (PubMed, PMC). Il s'illustre dans des tâches telles que la reconnaissance d'entités nommées et le question-réponse en contexte biomédical [59].

— **ClinicalBERT** : Conçu à partir de BERT et adapté spécifiquement aux notes cliniques issues du jeu de données MIMIC-III. Il vise à améliorer l'analyse des dossiers médicaux, notamment pour la détection d'événements cliniques.

— **PMC-LLaMA** : Variante basée sur LLaMA, formée sur PubMed Central (PMC). Conçu pour des tâches multilingues et biomédicales [42] .

— **BioGPT** : Dérivé de GPT-2 et entraîné de manière auto-régressive sur un vaste corpus d'articles biomédicaux (environ 15 millions de documents provenant de PubMed). Il se distingue par sa capacité de génération de textes biomédicaux cohérents, élément crucial pour les tâches de rédaction ou de synthèse scientifique [105].

La spécialisation de ces modèles dans le champ biomédical favorise une compréhension plus fine des terminologies scientifiques et des contextes cliniques, permettant d'atteindre des performances supérieures à celles de modèles généralistes sur des tâches exigeant une forte expertise du domaine. Chacune de ces approches illustre la tendance actuelle consistant à adapter les architectures de *deep learning* existantes à des corpus hyper-spécialisés, renforçant ainsi leur pertinence pour la recherche médicale, la pharmacovigilance ou l'analyse automatisée de la littérature scientifique.

6.2. Stratégie d'optimisation des LLMs

6.2.1. Optimisation de l'architecture des LLMs :

— Réduction de la complexité du mécanisme d'attention : La complexité quadratique du mécanisme d'attention dans les *Transformers* peut être réduite grâce à plusieurs techniques visant à limiter ou à réorganiser les calculs entre tokens. D'abord, les approches clairsemées (sparsity) restreignent la comparaison à un sous-ensemble de paires plutôt que de connecter chaque token à tous les autres, ce qui abaisse la complexité globale [80]. Ensuite, le hashing sensible à la localité (LSH) regroupe les tokens similaires dans des sous-groupes, permettant d'éviter les comparaisons redondantes [81]. Enfin, l'approximation linéaire des matrices d'attention projette les séquences dans un espace de plus faible dimension [82], menant à un allègement significatif de la charge computationnelle.

— Architectures alternatives pour améliorer l'efficacité des modèles : D'autres stratégies repensent la structure même du réseau afin de réduire la complexité ou la consommation de ressources. D'abord, la propagation séquentielle des états propose de remplacer l'attention auto-régressive par un enchaînement d'opérations linéaires et de convolutions[106], conduisant à une complexité quasi linéaire pour les longues séquences. Ensuite, la technique Mixture of Experts (MoE) segmente le réseau en plusieurs experts, dont seul un sous-ensemble est activé pour chaque requête, permettant de préserver un haut potentiel d'expressivité tout en limitant le calcul effectif [109].

— Réduction du nombre de paramètres : Elle constitue une priorité pour limiter le coût d'entraînement et d'inférence. Premièrement, le pruning (élagage) consiste à éliminer les poids ou les connexions les moins pertinents[86], rendant le réseau plus compact sans altérer significativement les performances. Deuxièmement, le partage de paramètres (parameter sharing) autorise certaines couches à réutiliser les mêmes poids [87], freinant ainsi la croissance exponentielle du nombre total de paramètres.

6.2.2. Optimisation de l'inférence :

L'inférence, qui correspond à la phase où un modèle de langage génère une réponse après avoir été entraîné, peut être coûteuse en calcul et en mémoire, notamment pour les modèles de grande taille. Pour accélérer ce processus sans dégrader la qualité des sorties, la quantization réduit la précision, ce qui diminue à la fois le volume mémoire et le temps de calcul [54], la mise en cache (Key-Value Cache) évite de recalculer sans cesse l'attention sur les tokens précédents, et l'optimisation des accès mémoire, illustrée notamment par FlashAttention [48], limite les allers-retours inutiles et fluidifie la production de nouveaux tokens.

6.2.3. Optimisation de la gestion des ressources :

Plusieurs méthodes se concentrent sur la répartition de la charge et la réduction de la consommation mémoire, tant à l'entraînement qu'à l'inférence.

— Gradient Checkpointing : Pour économiser la mémoire lors de l'entraînement, seules certaines activations intermédiaires sont conservées. Les autres sont recalculées à la demande durant la rétropropagation, ce qui permet d'entraîner des modèles plus volumineux sur des GPU moins dotés en VRAM [49].

— Offloading et parallélisation avancée : La mémoire GPU peut être complétée par la CPU ou par un stockage externe. Des outils comme DeepSpeed (ZeRO Offload) répartissent automatiquement les calculs ou les poids pour optimiser les ressources disponibles, réduisant la redondance et améliorant la scalabilité [51].

— Distillation de connaissances : Cette approche consiste à entraîner un modèle plus petit à reproduire les prédictions d'un modèle de grande taille. La distillation génère ainsi des variantes plus légères et moins coûteuses en calcul, tout en conservant des performances satisfaisantes [50].

6.2.4. Synthèse de quelques modèles de LLMs

Bien que les principes fondateurs de ces différents modèles puissent varier, ils se rejoignent tous sur un point : l'échelle du pré-entraînement. En combinant des milliards de mots issus de sources variées, ils développent une compréhension des structures linguistiques et du sens, qui peut être réinvestie avec succès dans des tâches ultérieures comme la classification thématique ou l'extraction d'informations. la table 5 résume quelques modeles et leurs caractéristiques.

TABLE 5. Synthèse des caractéristiques des principales familles de LLMs (Données recueillies des sources officielles des modèles et documentations techniques.

Modèle	Entreprise	Architecture	Taille (paramètres)	Caractéristiques clés
GPT (3.5, 4, Turbo)	OpenAI	Décodeur	175B (GPT-3.5) / 1760B (GPT-4) / 20B (GPT-3.5 Turbo)	Modèles généralistes, génération avancée, GPT-4 multimodal (image, texte).
Llama (1 & 2)	Meta	Architecture complète transformer	65B (Llama) / 70B (Llama-2)	Optimisé pour la sécurité et les performances, alternative open-source.
Mistral & Mixtral	Mistral AI	Décodeur	7B (Mistral) / 84B (Mixtral, MoE)	Mixtral utilise une Mixture of Experts pour optimiser la charge et réduire les coûts de calcul.
Flan-T5 & Flan-Alpaca	Google	Encodeur-Décodeur	Variable selon la version	Modèles instruction-tuned, spécialisés dans la compréhension du texte et le transfert de tâches.
phi (1, 1.5, 2)	Microsoft	transformer	1.3B (phi-1 & phi-1.5) / 2.7B (phi-2)	Optimisé pour des performances élevées avec un faible nombre de paramètres, battant des modèles beaucoup plus grands.
Claude (Claude 1 & 2)	Anthropic	Décodeur	130B	Focus sur l'alignement et la sécurité, performances proches de GPT-3.5 avec une gestion améliorée des biais.
Cohere	Cohere AI	Décodeur	Variable	Modèle optimisé pour le TALN en entreprise, utilisé dans des applications de recherche documentaire et analyse de texte.
Falcon	TII (Émirats Arabes Unis)	Décodeur	40B (Falcon-40B)	Alternative open-source, spécialisée dans la génération efficace avec un bon rapport coût/performance.
PaLM & T5	Google	Encodeur-Décodeur	540B (PaLM-2), variable (T5)	Optimisé pour le dialogue et la compréhension approfondie du langage, utilisé notamment pour les assistants IA avancés.

7. Le prompting

7.1. Principes du prompting en LLMs

Pour ce qui est des LLMs, l'un des principaux atouts réside dans leur capacité à prendre en charge des scénarios où les données annotées sont rares ou font complètement défaut [61]. Grâce au zero-shot et few-shot learning, ces modèles classent des données non étiquetées avec une performance élevée, évitant ainsi les tâches d'annotation manuelle, qui peuvent être longues, coûteuses voire irréalisables dans certaines situations.
Plutôt que de nécessiter un réentraînement complet pour chaque nouvelle tâche, comme dans l'apprentissage traditionnel, les LLMs s'appuient sur un ensemble d'instructions appelées prompts pour orienter leur génération de réponses. Par exemple, un prompt tel que : 'Classez le texte suivant dans l'une des catégories prédéfinies' permet d'effectuer une classification immédiate sans modification du modèle.

Au cours des dernières avancées en TALN, il est apparu que la performance et la flexibilité des LLMs, pouvaient notamment être considérablement améliorées à travers l'usage de différents types de prompts adéquats, que ce soit en précision, efficacité ou structure.

7.2. Concepts et applications du zero-shot et few-shot learning

Les approches de zero-shot learning et few-shot learning transforment l'utilisation des LLMs en leur permettant de résoudre des tâches sans nécessiter de données annotées ou avec seulement quelques exemples.
Dans le zero-shot learning, les modèles s'appuient uniquement sur des instructions bien formulées, exploitant les relations contextuelles apprises durant leur pré-entraînement sur des corpus variés. Cela leur permet de généraliser à des tâches inédites, comme l'attribution de préoccupations à des thématiques prédéfinies, sans disposer d'étiquettes spécifiques [64]. Cette approche réduit fortement la dépendance aux données annotées, bien qu'elle soit sensible à la clarté des instructions et puisse montrer des limites sur des tâches spécialisées.
Ceci à été pleinement illustré dans de nombreuses étude dans le domaine de santé, notamment à travers l'étude du HealthPrompt [99] un cadre de TALN clinique basé sur les prompts appelé HealthPrompt. Les auteurs démontrent que, grâce à une conception soignée des prompts, les modèles de langage préentraînés peuvent être utilisés efficacement pour des tâches de classification de textes cliniques en l'absence de données d'entraînement spécifiques.

Le Few-shot learning, quant à lui, complète cette approche en intégrant quelques exemples dans le prompt pour fournir un contexte supplémentaire au modèle. Grâce à ces exemples [63], les LLMs parviennent à s'adapter rapidement à des domaines spécifiques ou à des thématiques émergentes, même lorsque les données disponibles sont limitées. Par exemple, fournir au modèle des exemples d'association entre des préoccupations et leurs catégories permet d'améliorer la cohérence et la précision des classifications dans des scénarios complexes. Cette approche réduit considérablement le besoin d'entraînement supplémentaire tout en offrant des performances supérieures au Zero-shot learning, bien qu'elle reste dépendante de la qualité des exemples présentés. Ces deux paradigmes optimisent l'utilisation des LLMs dans des contextes où les données annotées sont limitées.

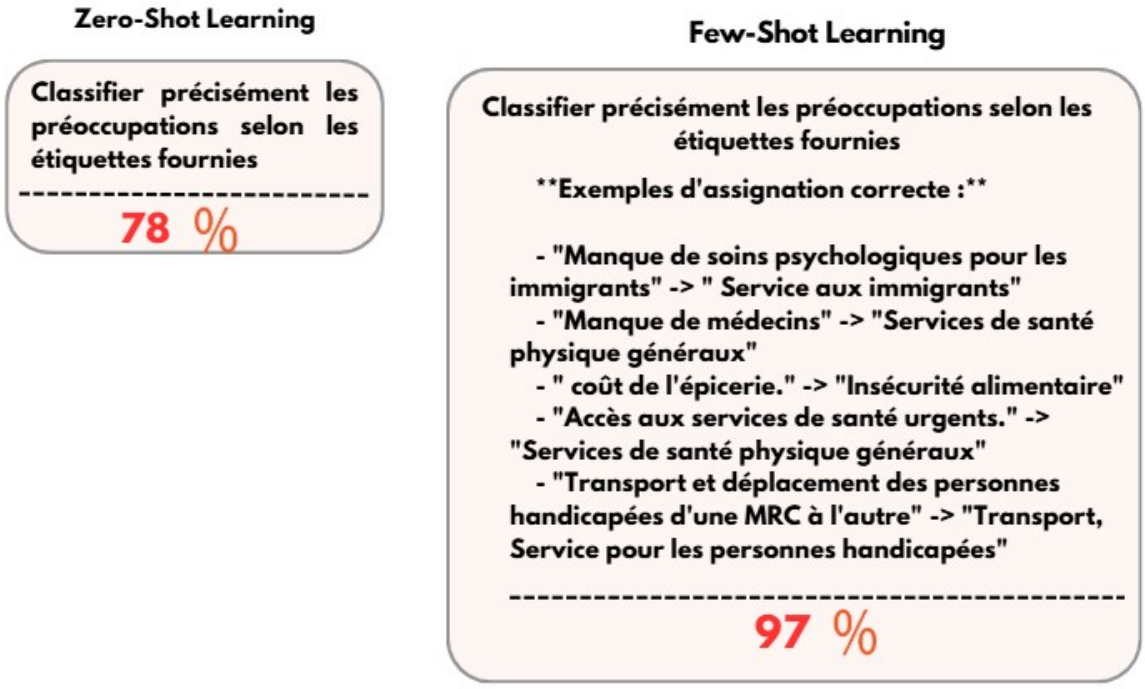


FIGURE 7. Comparaison et résultat du zero-shot et few-shot learning Sur les données des 3 Articles.

7.3. Structuration du raisonnement dans les LLMs avec le Chain-of-Thought

7.3.1. Principe et fonctionnement du Chain-of-Thought

L'un des défis des LLMs est leur capacité à raisonner de manière logique et structurée pour des tâches nécessitant plusieurs étapes de raisonnement. Les approches classiques de prompting standard exigent au modèle de fournir directement une réponse, augmentant ainsi le risque d'erreurs. Le chain-of-thought prompting (CoT) a été introduite pour améliorer la gestion des raisonnements complexes dans les LLMs[67]. Cette approche, proposée par des chercheurs de Google en 2022, incite les modèles à décomposer un raisonnement en plusieurs étapes intermédiaires avant de produire une réponse finale. Cette méthode a prouvé son efficacité dans des domaines tels que la résolution de problèmes mathématiques, le raisonnement logique et l'analyse sémantique avancée. En détail, le CoT fonctionne en renforçant la focalisation de l'attention du modèle sur chaque sous-problème avant d'arriver à une conclusion globale. Cela réduit le risque d'erreurs causées par la gestion simultanée de plusieurs informations complexes.

7.3.2. Variantes et améliorations du CoT

— Self-consistency chain-of-thought : Plutôt que de générer une seule réponse, cette approche produit plusieurs raisonnements distincts, puis sélectionne celui qui revient le plus fréquemment comme étant le plus fiable [79].

— Tree-of-thought (ToT) : Cette approche suit un raisonnement arborescent au lieu de générer un raisonnement linéaire. Le modèle explore plusieurs chemins simultanément et choisit la meilleure approche, ce qui est idéal pour les tâches nécessitant des explorations multiples [89].

— Generated knowledge prompting : Cette approche oblige le modèle à générer des informations contextuelles avant de répondre, améliorant ainsi la pertinence et la cohérence des prédictions.

8. Panorama de travaux récents en classification appliquée au TALN

La littérature scientifique récente témoigne d'un intérêt croissant pour l'automatisation de la classification des préoccupations psychosociales, notamment à partir de données issues de textes libres, qu'il s'agisse de questionnaires ou de contenus publiés en ligne. Plusieurs revues ont abordé ce champ sous différents angles, mais nombreuses sont celles qui présentent des limitations méthodologiques notables, qu'il s'agisse de la restriction à des corpus non cliniques, de l'absence d'analyse critique des performances des modèles sur des données déséquilibrées, ou encore d'une prise en compte limitée des enjeux propres à la santé mentale. Le Tableau 6 propose une synthèse de ces limites. Dans ce contexte, la présente revue vise à dépasser ces insuffisances en proposant une analyse comparative des approches mobilisées, depuis les modèles d'apprentissage automatique traditionnels jusqu'aux LLMs, en s'attachant à évaluer leur pertinence et leurs performances.

Dans un premier temps, les travaux pionniers dans le domaine de la classification de données psychosociales se sont appuyés sur des techniques issues de l'apprentissage automatique traditionnel, notamment Naive Bayes, les machines à vecteurs de support (SVM), k-Nearest Neighbors (k-NN) ou encore XGBoost. Ces approches, bien établies, ont été privilégiées pour leur simplicité de mise en œuvre et leur efficacité dans des contextes à faibles volumes de données. Par exemple, l'étude [107] a montré que des modèles supervisés classiques peuvent efficacement détecter les expressions liées à la dépression à partir de messages publiés sur Twitter.

Cependant, bien que performants dans certains contextes, ces modèles présentent plusieurs limitations structurelles. Une étude récente [83] souligne qu'ils nécessitent une ingénierie manuelle des caractéristiques particulièrement coûteuse en temps et en expertise. De plus, ils peinent à saisir les relations linguistiques complexes, en particulier dans les textes courts et informels caractéristiques des données psychosociales. Leur sensibilité aux jeux de données déséquilibrés constitue également un obstacle important, réduisant leur capacité à identifier des classes minoritaires mais cruciales, comme les signaux faibles de détresse psychologique.

Face à ces limites, la littérature a progressivement intégré des approches issues de l'apprentissage profond [10], capables d'apprendre des représentations complexes directement à partir des données textuelles, sans nécessiter de

TABLE 6. Limites principales de quelques revues récentes sur la classification en NLP

Titre	Année	Limites signalées par les auteurs
Natural Language Processing in Mental Health Applications Using Non-clinical Texts[18]	2017	Cette revue se limite à l'analyse de textes issus du grand public, exclut les notes cliniques, ne propose aucune méta-analyse quantitative des performances et présente une hétérogénéité dans la définition des troubles.
Natural Language Processing Advancements by Deep Learning : A Survey[62]	2020	Cette revue est bien exhaustive sur les avancées générales du deep learning, mais elle n'aborde pas spécifiquement la santé psychologique, propose très peu de comparaisons inter-jeux de données et ne traite pas du déséquilibre des classes.
Deep Learning–based Text Classification : A Comprehensive Review[6]	2021	La revue se concentre exclusivement sur les méthodes d'apprentissage profond, ignore les approches hybrides ou traditionnelles, couvre très marginalement les applications en santé mentale et n'évalue pas de façon critique les protocoles expérimentaux.
Text classification algorithms : A survey[33]	2019	En raison de son champ très large, la revue cite peu de cas psychosociaux, s'appuie presque uniquement sur des données en anglais et ne discute ni des questions éthiques ni de la confidentialité des données.
Large Language Models in Medical and Healthcare Fields : Applications, Advances, and Challenges[72]	2024	Cette revue fournit des mesures de performance peu détaillées et se contente d'une discussion éthique encore largement spéculative.
The potential of GPT-4 to analyse medical notes in three different languages[68]	2025	Étude rétrospective portant sur un échantillon limité, absence d'intégration clinique en temps réel, couverture linguistique restreinte et erreurs notables (20%) dans les diagnostics moins fréquents.
Validation of GPT-4 for clinical event classification : A comparative analysis with ICD codes and human reviewers[69]	2024	Données issues d'un seul centre gastro-entérologique, classes rares insuffisamment représentées, comparaison limitée à un codage manuel local et analyse de coût/latence non abordée.
GPT is an effective tool for multilingual psychological text analysis[70]	2024	Corpus majoritairement occidental, nombre de langues encore restreint, absence d'évaluation sur la stabilité temporelle et possibles biais culturels dans la détection des émotions.
Validating the use of large language models for psychological text classification[104]	2025	Échantillon de convenance sur-représentant les pays anglophones, absence d'expérimentation en conditions réelles, interprétabilité limitée malgré l'utilisation de chaînes de raisonnement et manque de mesures d'impact clinique.
Enhancing suicide attempt risk prediction models with temporal clinical note features [76]	2024	Données issues exclusivement d'anciens combattants américains, généralisation incertaine à d'autres populations, couverture temporelle restreinte et contribution exacte du LLM difficile à isoler dans le modèle hybride.

sélection manuelle des caractéristiques. Cette transition a été marquée par plusieurs études, dont l'entraînement d'un pipeline BERT+Bi-LSTM sur 60000 messages Reddit relatifs au stress et aux idées suicidaires [24], la F1-score atteint 0,93 contre 0,85 pour un auto-encodeur TF-IDF confirmant la capacité de l'attention contextuelle à capturer des expressions idiomatiques ou implicites que les modèles linéaires peinent à détecter. La supériorité de l'apprentissage profond s'étend également aux dossiers médicaux. Une étude [25] évalue un CNN séquentiel sur

500000 notes cliniques de vétérans américains, la valeur prédictive positive pour la tentative de suicide dans le décile le plus à risque passe de 0,028 (algorithme logistique REACH-VET) à 0,55.

Toutefois, l'apprentissage profond n'est pas systématiquement supérieur. Un exemple révélateur est l'étude comparative de modèles appliqués à des sessions de thérapie textuelles [110], où BERT affiche des performances inférieures à celles de Naive Bayes ou SVM, malgré sa complexité architecturale. Ce résultat paradoxal s'explique par la nature hétérogène et déséquilibrée des données, mais aussi par des critères d'évaluation biaisés, centrés sur la satisfaction des patients plutôt que sur des indicateurs objectifs d'efficacité thérapeutique.

En parallèle du classement supervisé, le topic modeling (LDA, NMF) facilite l'exploration non supervisée de préoccupations émergentes. une étude [77] a démontré qu'un modèle ATAM dérivé de LDA détecte les tendances sanitaires dans 1,6 millions de tweets PMC. D'autres travaux ont appliqué ces approches aux préoccupations psychosociales [84] à savoir l'insécurité alimentaire ou au stress et au bien-être, dévoilant plusieurs limites dans ce sens à savoir les nouvelles thématiques émergentes, le déséquilibre de classes et voir même le chevauchement des clusters.

Dans la continuité des progrès offerts par l'apprentissage profond, l'introduction des LLms dévoile de nouvelles perspectives. Les premiers travaux ont exploité les modèles propriétaires accessibles via API, tels que GPT-4. Une étude [73] a montré que GPT-4 peut coder des entretiens médicaux avec une qualité comparable à celle de chercheurs humains, mais présente encore des lacunes sur environ 20 % des diagnostics finaux. De même, une autre analyse [70] sur près de 48 000 messages multilingues démontre que GPT-4 surpasse nettement les lexiques traditionnels pour la détection d'émotions et de contenus à risque, sans nécessiter d'annotation supervisée.

En santé mentale, les LLMs se révèlent particulièrement performants pour la détection des idées suicidaires. Une méta-analyse récente dans le Journal of Medical Systems [74] montre qu'ils dépassent parfois les cliniciens dans la prédiction précoce de comportements à risque. À une échelle plus large, Krause et al [76] ont démontré qu'un modèle hybride combinant des embeddings textuels de dossiers cliniques améliore significativement la valeur prédictive de la tentative de suicide, comparé à l'algorithme logistique REACH-VET (Applied Clinical Informatics).

Dans des contextes où la souveraineté des données est cruciale, des alternatives locales aux LLMs hébergés dans le cloud ont émergé. Des modèles comme Mistral, Qwen ou LLaMA permettent un déploiement local. Une étude comparative [78] menée sur des rapports de thrombectomie a montré que Mixtral surpasse BioMistral en précision (0,99 contre 0,81) tout en réduisant de plus de 60 % le temps d'annotation grâce à une interaction humaine contrôlée. Cependant, ces approches locales ne sont pas exemptes de défis, elles exigent une maintenance continue, notamment pour adapter les modèles aux évolutions des pratiques cliniques. De plus, la question de l'explicabilité reste centrale. Bien que les techniques de raisonnement structuré comme le chain-of-thought soient prometteuses, leur coût computationnel reste prohibitif dans les applications en temps réel, et leur robustesse face à des cas cliniques atypiques demeure insuffisamment documentée.

Afin d'illustrer concrètement la diversité des applications récentes de l'apprentissage automatique dans le champ de la santé psychologique, le Tableau 7 présente une sélection d'études représentatives. Celles-ci couvrent un large spectre d'usages allant de la détection de la dépression sur les réseaux sociaux à l'annotation clinique automatisée, en passant par l'identification de préoccupations émergentes. Elles mobilisent aussi bien des modèles traditionnels que des approches profondes ou des LLMs.

Le tableau 8 suivant propose une synthèse comparative des principaux corpus, protocoles et résultats de classification selon les trois générations d'approches examinées : modèles classiques (ML), réseaux neuronaux profonds (DL), et grands modèles de langage (LLMs) permettant ainsi une vue globale plus détaillée.

TABLE 7. Exemples récents d'applications de l'apprentissage automatique dans différents domaines de la santé psychologique

Domaine d'utilisation	**Applications**	**Type de méthode ML**	**Technique ML utilisée**	**Année**	**Réf.**
Détection de la dépression	Classification automatique du niveau de sévérité à partir de messages sur les réseaux sociaux	Classification	Transformer —BERT, RoBERTa, et DeBERTa	2024	[113]
Prédiction du risque suicidaire post-hospitalisation	Stratification du risque de suicide via rapports de sortie EHR	Classification	LLM fine-tuned(GPT-4)	2025	[96]
Idéation suicidaire en ligne	Identification de posts Reddit exprimant des pensées suicidaires	Classification	CNN-LSTM hybride	2024	[101]
Classification de préoccupations psychosociales	Catégorisation de textes d'enquêtes mensuelles en thèmes de santé psychosociales et de bien-être physique et mental	Classification	NB, SVM et KNN comparés à Sentences transformers, Bert et Distillbert	2024	[83]
Analyse de séances de counseling	Évaluation automatique de conversations thérapeutiques pour la qualité de l'interaction	Classification	BERT vs SVM / NB	2024	[110]
Annotation clinique automatisée	Codage multilingue d'entretiens médicaux par LLM	Classification	GPT-4+ prompt engineering	2025	[104]
Détection d'émotions multiples dans les posts sociaux	Identification simultanée de *plusieurs* émotions dans des tweets	Classification multi-étiquettes	BERT+ transfert de connaissances	2023	[75]
Suicide— vétérans	Prédiction de tentative de suicide à partir de notes cliniques séquentielles	Classification	Réseau neural séquentiel (CNN + embeddings)	2023	[25]
Préoccupations émergentes (insécurité alimentaire)	Découverte de thèmes santé/alimentation dans 1,6M de tweets	Topic modelling/ Clustering	LDA + NMF	2024	[88]

TABLE 8. Corpus, protocoles et performances clés entre ML, DL et LLMs

Cat.	Étude	Corpus (langue)	N	labels	Ratio	Split	Modèle testé	Baseline	Métrique / Score
ML	[21]	E-mails et news (Turc + Anglais)	Emails : 800 et News : 8000	Binary (e-mail) et Multi-class (news, 10 classes)	Imbalanced (news)	75% train / 25% test (ex : 300 train / 100 test e-mails)	SVM	Comparaisons internes entre 16 combinaisons de prétraitements	Micro-F1 max : 0.9888 (EN mail, 500 feats)
ML	[38]	Reuters-21578 (EN), Web, MNIST, Adult (EN) et images faciales	Reuters : ≈13,000 MNIST : 60,000 Adult : 11,221 Web : 49,749	Reuters : 118 Visages : 2	Reuters : déséquilibré et images : ≈équilibré	ModApte	SVM (linéaire, RBF, poly, sigmoïde) + SMO	Naïve Bayes, Decision Tree, NN, FindSim	Reuters : Breakeven point (ex : ≈0.85)
ML	[35]	Avis produits	25,581	2	déséquilibré (8141 good vs 2323 bad)	80% train / 20% test	XGBoost + TF-IDF XGBoost + Word2Vec NB + TF-IDF NB + Word2Vec	4 comparaisons	F1 Score : Max = XGB+W2V : 0.941
ML	[108]	Patient–provider msgs (EN)	Participants : 621 (SM) / 221 (CCVT). Messages : 1,105 (SM) / 1,229 (CCVT)	2 (risk/no-risk)	SM : 82.3% vs 17.7% CCVT : 81.8% vs 18.2%	5-fold CV	Logistic Regression Random Forest SVM PHS-BERT (fine-tuned) + transfert learning	Aléatoire	F1 max : 0.797 (PHS-BERT + transfert + LR personnalisés)
DL	[59]	PubMed abstracts + PMC full text (EN)	NER : 9 datasets RE : 3 QA : 3	Multi-label	NR	Std. split	BioBERT finetuned	BiLSTM-CRF	F1 NER : +0.62 F1 RE : +2.80 MRR QA : +12.24
DL	[97]	PubMed abstracts, QA et corpus d'extraction (EN)	13 datasets / 6 tâches	Variable	BioASQ équilibré ; ChemProt déséquilibré	Std. split	BioBERT, BlueBERT, PubMedBERT, PubMedELECTRA (base et large)		F1 : jusqu'à 93.84 Score BLURB : 82.91 (PubMedBERT-LARGE)
DL	[100]	CoNLL 2002 (ES, NL) CoNLL 2003 (EN, DE) DrugNER (EN)	Variable	PER, LOC, ORG, MISC (CoNLL) + entités médicales	NR	Std. split	HMM, SVM, CRF BiLSTM, BiLSTM-CRF, CNN	SOTA antérieur : CRF et SVM	F1 jusqu'à 91.62
DL	[13]	Monolingue : CR, AGNews, Emotion Multilingue : MARC (EN, FR, DE, ZH, ES, JA)	8, 64, ou full shots/class × plusieurs classes	Variable	Uniforme par configuration (8, 64 ou full / classe)	10 splits aléatoires	SETFIT (RoBERTa, MPNet, MiniLM)	T-FEW, ADAPET, PERFECT, RoBERTa-Large	Accuracy (RAFT), MAE (MARC) ; Score max : 75.3% (SetFitMPNet, N=64)
LLM	[42]	4.8M articles PMC (EN) + QA sets : PubMedQA, MedMCQA, USMLE	PubMedQA : 211K MedMCQA : 182K USMLE : 12K	QA	USMLE : ≈81/19/19 Autres : standard officiel	Full fine-tuning, PEFT (LoRA), Data-efficient finetuning	PMC-LLaMA (LLaMA-7B finetuné)	LLaMA-7B, ChatGPT, InstructGPT	Score max : 50.54% (MedMCQA) 69.53% (PubMedQA) 40.61% (USMLE, ID)
LLM	[96]	Discharge notes (EN) – 2 hôpitaux universitaires USA	11,970 (1 cas pour 5 témoins)	Mort par suicide ou accident (binaire)	1 :5 (cas vs témoins)	Aléatoire sur 9 ans, 2005–2014	GPT-4-1106-preview (zero-shot, API HIPAA)	Modèles cliniques classiques (sociodémographiques et utilisation)	AUC = 0.629 (suicide/accident) HR = 8.86 (Fine et Gray) AUC = 0.74 (suicide seul)
LLM	[73]	20 entretiens semi-directifs de patients atteints d'AABP (EN)	20 transcripts	Thèmes émergents (urinaire, sexuel, santé mentale, etc.)	Non applicable (thématique libre)	Non applicable	GPT-4	Codage humain manuel en 3 phases	Cohen's = 0.401 (accord modéré humain vs GPT-4)
LLM	[111]	100 notes d'admission psychiatrique (Allemagne, DE)	100 patients	2 (suicidal / not)	1 :1 (défini par psychiatres)	Full + 5 stratégies de prompts	Llama-2 (English, Emgerman, Sauerkraut)	humain (résident + psychiatre)	(résident + psychiatre) Score max : Accuracy = 87.5% Sensitivity = 83%, Specificity = 91.8% (Emgerman, prompt P3)
LLM	[114]	Reddit – r/SuicideWatch (EN)	125 utilisateurs annotés (Task B) 162 posts (Task A)	Task A : Evidence (binaire) Task B : Risk level (Low, Moderate, Severe)	Task B : 13 Low, 74 Moderate, 38 Severe	Evaluation sur 125 utilisateurs sélectionnés parmi 209	LLaMA-2, Mistral, MentaLLaMA, Tulu-2, ChatGLM-3, OpenHermes, Zephyr (LLMs open source)	comparaison inter-équipes	Task A : Rappel max = 0.944 Task B : Consistance moyenne max = 0.979

9. Défis et enjeux des LLMs

Les LLMs soulèvent plusieurs défis techniques, éthiques et sociétaux. Tout d'abord, leur coût computationnel et énergétique représente un défi important. L'entraînement et l'inférence de ces modèles exigent une puissance de calcul élevée, en particulier pour les modèles de très grande taille comme LLama3.1 :405b, ce qui engendre des coûts élevés et un impact environnemental significatif. Cet enjeu motive le développement d'architectures plus légères et optimisées pour réduire leur empreinte carbone [91].
Ensuite, les LLMs sont exposés à des biais algorithmiques qui peuvent affecter la fiabilité de leurs réponses et conduire à la diffusion de désinformation [92]. Ces biais résultent principalement de la qualité et de la diversité des données d'entraînement. Si ces données contiennent des informations erronées, biaisées ou incohérentes, le modèle risque de les apprendre et de les reproduire, amplifiant ainsi ces erreurs. De plus, l'un des défis majeurs des LLMs est leur tendance à générer du texte réaliste mais potentiellement inexact. Cette hallucination se produit lorsque le modèle produit des informations non fondées, créant ainsi l'illusion d'une connaissance fiable là où il n'y en a pas. En généralisant à partir des données d'entraînement, les LLMs peuvent parfois extrapoler de manière excessive et générer du contenu plausible mais incorrect, notamment en l'absence de contexte approprié.
En passant à la question de la confidentialité et de la protection des données qui est également un enjeu présent, la capacité des LLMs à mémoriser et à reproduire involontairement des informations sensibles si celles-ci étaient présentes dans leurs données d'entraînement représente un défis majeur. De plus, les utilisateurs peuvent, sans le savoir, partager des données confidentielles lors de leurs interactions avec ces modèles, amplifiant ainsi les risques de fuite d'informations. Un exemple notable illustrant ces risques concerne Samsung, qui a fait face à une fuite de données après l'utilisation de ChatGPT par certains de ses employés en mars 2023. En conséquence, ChatGPT est devenu capable de restituer ces informations, exposant ainsi des secrets industriels sensibles.
Cette problématique est d'autant plus préoccupante que les politiques de confidentialité et de traitement des données varient selon les versions des modèles et les comptes utilisateurs. Certaines interactions peuvent être stockées et exploitées pour améliorer les performances du modèle, tandis que d'autres versions garantissent une confidentialité renforcée en limitant la conservation et l'utilisation des données saisies.
Par ailleurs, l'utilisation de techniques comme le Retrieval-Augmented Generation (RAG), bien qu'elle permette d'améliorer la précision et la pertinence des réponses en intégrant des données externes, peut également compromettre la confidentialité des informations traitées. En accédant à des bases de données non sécurisées ou en interagissant avec des sources sensibles, le modèle peut incorporer et divulguer involontairement des informations confidentielles. Enfin, l'explicabilité et l'interprétabilité des modèles restent des défis ouverts, ce qui rend difficile la compréhension de leurs décisions. Cela peut être problématique dans des applications nécessitant une forte transparence à titre d'exemple notre application lié à la classification des préoccupations psychosociales surlesquelles plusieurs décision seront prises. Face à ces défis, plusieurs techniques sont explorées, notamment l'amélioration du raisonnement avec le chain-of-thought et la conception de modèles plus sûrs, transparents et économes en ressources.

10. Conclusion

L'état de l'art proposé ici a pour but de donner un socle conceptuel et méthodologique solide, permettant d'atteindre une compréhension des fondements et des évolutions des techniques de l'IA et du TALN mises en œuvre dans les trois articles. En s'attachant aux méthodes traditionnelles d'apprentissage automatique, aux méthodes de modélisation thématique non supervisées et à l'émergence des modèles de langage massifs (LLMs), Nous avons présenté un aperçu détaillé des outils et concepts essentiels pour répondre aux défis liés à la classification et à l'analyse de textes courts et complexes.

Les techniques classiques telles que les SVM ou les méthodes de vectorisation ont joué un rôle fondamental dans l'établissement des bases de la classification textuelle. Au même moment, des méthodes de modélisation thématiques telles que la LDA ou la NMF ont donné les moyens d'explorer et de structurer des corpus non étiquetés. L'essor des *Transformers* a conduit au développement des modèles de langage massifs (LLMs), capables de capturer des relations contextuelles complexes et d'être appliqués à diverses tâches. Enfin, des innovations comme le retrieval-augmented generation (RAG) renforcent la robustesse et la transparence des systèmes de génération de texte en intégrant des

ressources externes, illustrant la façon dont ces différentes approches s'articulent pour relever les défis du traitement automatique du langage naturel.

Références

[1] LeCun, Y., Bengio, Y., Hinton, G. (2015). Deep learning. Nature, 521(7553), 436-444

[2] Schmidhuber, J. (2015). Deep learning in neural networks : An overview. Neural Networks, 61, 85-117.

[3] Young, T., Hazarika, D., Poria, S., Cambria, E. (2018). Recent Trends in Deep Learning Based Natural Language Processing. IEEE Computational Intelligence Magazine, 13(3), 55-75.

[4] Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, Amodei, D. (2020). Language Models are Few-Shot Learners. arXiv preprint arXiv :2005.14165

[5] Blei, D. M., Ng, A. Y., Jordan, M. I. (2003). Latent Dirichlet Allocation. Journal of Machine Learning Research, 3, 993-1022.

[6] Minaee, S., Cambria, E., et al. (2021). Deep Learning–based Text Classification : A Comprehensive Review. ACM Computing Surveys, 54(3), 1-40

[7] Aggarwal, C. C., Zhai, C. (2012). A Survey of Text Classification Algorithms. In Mining Text Data (pp. 163-222). Springer.

[8] Gao, J., Galley, M., & Li, L. (2019). Neural approaches to conversational AI. Foundations and Trends in Information Retrieval, 13(2-3), 127-298. https ://doi.org/10.1561/1500000074

[9] Joulin, A., Grave, E., Bojanowski, P. Mikolov, T. (2017). Bag of Tricks for Efficient Text Classification. arXiv preprint arXiv :1607.01759.

[10] Vaswani, A., Shazeer, N., Parmar, N., et al. (2017). Attention is All You Need. In Advances in Neural Information Processing Systems (pp. 5998-6008).

[11] Devlin, J., Chang, M. W., Lee, K., Toutanova, K. (2019). BERT : Pre-training of Deep Bidirectional Transformers for Language Understanding. In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics : Human Language Technologies, Volume 1 (Long and Short Papers) (pp. 4171-4186).

[12] Howard, J., Ruder, S. (2018). Universal Language Model Fine-tuning for Text Classification. In Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics .

[13] Tunstall, L., Reimers, N., Jo, U. E. S., et al. (2022). SetFit : Efficient Few-Shot Learning Without Prompts.

[14] Kotsiantis, S. B., Zaharakis, I., Pintelas, P. (2007). Supervised Machine Learning : A Review of Classification Techniques. Emerging Artificial Intelligence Applications in Computer Engineering, 160, 3-24.

[15] Jain, A. K. (2010). Data Clustering : 50 Years Beyond K-Means. Pattern Recognition Letters, 31(8), 651-666.

[16] Zhu, X. (2005). Semi-Supervised Learning Literature Survey. Computer Science, University of Wisconsin-Madison, 2(3), 4.

[17] Sutton, R. S., Barto, A. G. (2018). Reinforcement Learning : An Introduction. MIT Press.

[18] Calvo, R. A., Milne, D. N., Hussain, M. S., Christensen, H. (2017). Natural Language Processing in Mental Health Applications Using Non-clinical Texts. Natural Language Engineering, 23(5), 649-685.

[19] He, H., Garcia, E. A. (2009). Learning from Imbalanced Data. IEEE Transactions on Knowledge and Data Engineering, 21(9), 1263-1284.

[20] Jima, J. R., Talukder, M. A. R., Malakar, P., Kabir, M. M., Nur, K., Mridha, M. F. (2024). Recent advancements and challenges of NLP-based sentiment analysis : A state-of-the-art review. Natural Language Processing, 100059. https ://doi.org/10.1016/j.nlp.2024.100059

[21] Uysal, A. K., Gunal, S. (2014). The impact of preprocessing on text classification. Information Processing Management, 50(1), 104-112. https ://doi.org/10.1016/j.ipm.2013.08.006

[22] Patro, S., Sahu, K.K., 2015. Normalization : A preprocessing stage. arXiv preprint arXiv :1503.06462.

[23] Aliwy, A.H., 2012. Tokenization as preprocessing for Arabic tagging system. Int. J. Inf. Educ. Technol. 2 (4), 348.

[24] Jamali,A.A., Berger,C., &Spiteri,R.J. (2023). Momentary depressive feeling detection using X (formerly Twitter) data : Contextual language approach. JMIR AI,2, e49531. https ://doi.org/10.2196/49531

[25] Martinez,C., Levin,D., Jones,J., Finley,P.D., McMahon,B., Dhaubhadel,S., Cohn,J., Million Veteran Program, MVP Suicide Exemplar Workgroup, Oslin,D.W., Kimbrel,N.A., & Beckham,J.C. (2023). Deep sequential neural network models improve stratification of suicide attempt risk among US veterans. Journal of the American Medical Informatics Association,31(1),220-230. https ://doi.org/10.1093/jamia/ocad167

[26] Kuang, Q., Xu, X. (2010). Improvement and Application of TF·IDF Method Based on Text Classification. In Proceedings of the 2010 International Conference on Internet Technology and Applications

[27] Bojanowski, P., Grave, E., Joulin, A., Mikolov, T. (2017). Enriching Word Vectors with Subword Information. Transactions of the Association for Computational Linguistics, 5, 135-146.

[28] Kruczek, J., Kruczek, P., Kuta, M. (2020). Are n-gram categories helpful in text classification ? In V. V. Krzhizhanovskaya, G. Závodszky, M. H. Lees, J. J. Dongarra, P. M. A. Sloot, S. Brissos, J. Teixeira (Eds.), Computational Science – ICCS 2020. Springer, Lecture Notes in Computer Science, 12138. https ://doi.org/10.1007/978-3-030-50423-6-24

[29] Zhang, G., Zhou, Y., Bollegala, D. (2024). Evaluating unsupervised dimensionality reduction methods for pretrained sentence embeddings. https ://arxiv.org/abs/2403.12078

[30] Doe, J., Smith, A. (2023). Dimensionality Reduction for Text Classification using Linear Discriminant Analysis. Journal of Natural Language Processing, 15(2), 123-135.

[31] Arora, S., Hu, W., Kothari, P. K. (2018). An analysis of the t-SNE algorithm for data visualization. In Proceedings of the 31st Conference On Learning Theory (Vol. 75, pp. 1455–1462). PMLR.

[32] Hamster, U. A., Lee, J.-U., Geyken, A., Gurevych, I. (2023). Rediscovering Hashed Random Projections for Efficient Quantization of Contextualized Sentence Embeddings.
[33] Kowsari, K., Jafari Meimandi, K., Heidarysafa, M., Mendu, S., Barnes, L., & Brown, D. (2019). Text classification algorithms : A survey. Information, 10(4), 150. https ://doi.org/10.3390/info10040150
[34] Mikolov, T., Karafiát, M., Burget, L., Cernocký, J., Khudanpur, S. (2010). Recurrent Neural Network Based Language Model. Interspeech, 1045-1048.
[35] Rifky, I., Utami, E., Hartanto, A. D. (2022). Comparison of Naïve Bayes Algorithm and XGBoost on Local Product Review Text Classification. Edumatic : Jurnal Pendidikan Informatika, 6(1), 143-149
[36] Paradita, A. X., Agustiana, N., Asriana, Rukmana, P. U., Nelsa, P., Lubis, M. (2024). Comparative Analysis of Naïve Bayes and K-Nearest Neighbors Algorithms for Customer Churn Prediction : A Kaggle Dataset Case Study. Proceedings of the International Conference on Information Science and Technology Innovation (ICoSTEC), 3(1).
[37] Ghojogh, B., Ghodsi, A. (2024). Backpropagation and optimization in deep learning : Tutorial and survey. OSF Preprints. https ://doi.org/10.31219/osf.io/d97b3
[38] Hearst, M.A., Dumais, S.T., Osuna, E., Platt, J., Scholkopf, B., 1998. Support vector machines. IEEE Intell. Syst. Appl. 13 (4), 18–28.
[39] Blei, D. M., Ng, A. Y., Jordan, M. I. (2003). Latent Dirichlet Allocation. Journal of Machine Learning Research, 3, 993-1022.
[40] Wang, C., Blei, D. M. (2011). Collaborative Topic Modeling for Recommending Scientific Articles. Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 448-456.
[41] Paul, M. J., and Dredze, M. (2014). "Discovering Health Topics in Social Media Using Topic Models." PLoS ONE, 9(8), e103408. DOI : 10.1371/journal.pone.0103408.
[42] Wu, C., Zhang, X., Zhang, Y., & Wang, Y. (2023). PMC-LLaMA : Further finetuning LLaMA on medical papers [Preprint]. arXiv. https ://doi.org/10.48550/arXiv.2304.14454
[43] Yang, A. C., et al. (2021). "Social Media Topic Modeling and Online Mental Health Discussions During the COVID-19 Pandemic." Journal of Affective Disorders, 295, 268–275. DOI : 10.1016/j.jad.2021.08.035.
[44] Landauer, T. K., Foltz, P., and Laham, D. (1998). "An Introduction to Latent Semantic Analysis." Discourse Processes, 25, 259–284. DOI : 10.1080/01638539809545028
[45] Zoya, S. L., Shafait, F., and Latif, R. (2021). "Analyzing LDA and NMF Topic Models for Urdu Tweets via Automatic Labeling." IEEE Access, 9, 3112620. DOI : 10.1109/ACCESS.2021.3112620.
[46] Egger, R., Yu, J. (2022). A topic modeling comparison between LDA, NMF, Top2Vec, and BERTopic to demystify Twitter posts. Frontiers in Sociology, 7, 886498. https ://doi.org/10.3389/fsoc.2022.886498
[47] Lau, J. H., Newman, D., Baldwin, T. (2013). Evaluating topic coherence measures. In Neural Information Processing Systems Foundation (NIPS 2013) - Topic Models Workshop.
[48] Dao, T., Fu, D. Y., Ermon, S., Rudra, A., & Ré, C. (2022). FlashAttention : Fast and memory-efficient exact attention with IO-awareness. In Advances in Neural Information Processing Systems (NeurIPS).
[49] Chen, T., Xu, B., Zhang, C., & Guestrin, C. (2016). Training deep nets with sublinear memory cost. arXiv preprint arXiv :1604.06174.
[50] Hinton, G., Vinyals, O., & Dean, J. (2015). Distilling the knowledge in a neural network. Deep Learning Workshop (NIPS).
[51] Rajbhandari, S., Smith, S., Ruwase, O., & He, Y. (2021). ZeRO-Infinity : Breaking the GPU Memory Wall for Extreme Scale Deep Learning. In SC '21 : Proceedings of the International Conference for High Performance Computing, Networking, Storage and Analysis.
[52] Tellex, S., Kollar, T., Dickerson, S., Walter, M. R., Banerjee, A. G., J., Teller, S., & Roy, N. (2020). Robots That Use Language : A Survey. Annual Review of Control, Robotics, and Autonomous Systems, 3, 1–35.
[53] Pons, E., Braun, L. M. M., Hunink, M. G. M., & Kors, J. A. (2016). Natural language processing in radiology : A systematic review. Radiology, 279(2), 329–343. https ://doi.org/10.1148/radiol.16142770
[54] Dettmers, T., Lewis, M., Ganin, Y., & Zettlemoyer, L. (2022). 8-bit optimizers via block-wise quantization. In International Conference on Learning Representations (ICLR).
[55] Meaney, C., Stukel, T. A., Austin, P. C., Moineddin, R., Greiver, M., Escobar, M. (2023). Quality indices for topic model selection and evaluation : A literature review and case study. BMC Medical Informatics and Decision Making, 23, Article 132.
[56] Rudiger, M., Antons, D., Joshi, A. M., and Salge, T. O. (2022). "Topic Modeling Revisited : New Evidence on Algorithm Performance and ¨ Quality Metrics." PLoS One, 17(4), e0266325. DOI : 10.1371/journal.pone.0266325. PMID : 35482786.
[57] R. Bommasani, D. A. Hudson, et al., On the Opportunities and Risks of Foundation Models, Journal of Artificial Intelligence Research, vol. 71, pp. 517–659, 2021.
[58] P. Liu, W. Yuan, J. Fu, Z. Jiang, et al., Pre-train, Prompt, and Predict : A Systematic Survey of Prompting Methods in NLP, ACM Computing Surveys, 2023.
[59] J. Lee, W. Yoon, S. Kim, D. Kim, S. Kim, et J. Kang, BioBERT : a pre-trained biomedical language representation model for biomedical text mining, Bioinformatics, vol. 36, no. 4, pp. 1234–1240, 2020.
[60] Radford, A., Narasimhan, K., Salimans, T., Sutskever, I. (2018). Improving Language Understanding by Generative Pre-Training. OpenAI.
[61] K. Singhal, P. X. Liu, et al., Large Language Models Encode Clinical Knowledge, Nature, vol. 610, pp. 60–67, 2022.
[62] Torfi, A., Shirvani, R. A., Keneshloo, Y., Tavaf, N., and Fox, E. A. (2020). Natural Language Processing Advancements By Deep Learning : A Survey. arXiv :2003.01200.
[63] Wang, W., Yao, Q., Kwok, J. T., Ni, L. M. (2020). Generalizing from a Few Examples : A Survey on Few-Shot Learning. ACM Computing Surveys (CSUR), 53(3), 1-34.
[64] Xian, Y., Schiele, B., Akata, Z. (2017). Zero-Shot Learning—The Good, the Bad and the Ugly. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 4582-4591.
[65] Zhao, P., Zhang, H., Yu, Q., Wang, Z., Geng, Y., Fu, F. Cui, B. (2024). Retrieval-Augmented Generation for AI-Generated Content : A Survey.

[66] Fedus, W., Zoph, B., Shazeer, N. (2022). Switch Transformers : Scaling to Trillion Parameter Models with Simple and Efficient Sparsity. Journal of Machine Learning Research, 23(120), pp. 1–39.

[67] Wei, J., Wang, X., Schuurmans, D., et al. (2022). Chain-of-Thought Prompting Elicits Reasoning in Large Language Models. Advances in Neural Information Processing Systems (NeurIPS), 35.

[68] Saad M.C., HoffmannA.F., Tan A. L. M., Nalbandyan M., Omenn G. S., Kohane I. S. (2025).The potential of Generative Pre-trained Transformer 4 (GPT-4) to analyse medical notes in three different languages : a retrospective model-evaluation study. The Lancet Digital Health, 7(1), e35–e43.

[69] Wang, Y., Huang, Y., Nimma, I. R., Pang, S., Pang, M., Cui, T., & Kumbhari, V. (2024). Validation of GPT-4 for clinical event classification : A comparative analysis with ICD codes and human reviewers. Journal of Gastroenterology and Hepatology, 39(8), 1535–1543. https ://doi.org/10.1111/jgh.16561

[70] Rathje, S., Mirea, D.-M., Sucholutsky, I., Marjieh, R., Robertson, C. E., & Van Bavel, J. J. (2024). GPT is an effective tool for multilingual psychological text analysis. Proceedings of the National Academy of Sciences, 121(34), e2308950121. https ://doi.org/10.1073/pnas.2308950121

[71] Meddeb, A., Ebert, P., Bressem, K. K., Desser, D., Dell'Orco, A., Bohner, G., Nawabi, J. (2024). Evaluating local open-source large language models for data extraction from unstructured reports on mechanical thrombectomy in patients with ischemic stroke. Journal of NeuroInterventional Surgery. Advance online publication.

[72] Wang, D., & Zhang, S. (2024). Large language models in medical and healthcare fields : Applications, advances, and challenges. Artificial Intelligence Review, 57, 299.

[73] Li, K. D., Fernandez, A. M., Schwartz, R., Rios, N., Carlisle, M. N., Amend, G. M., Patel, H. V., & Breyer, B. N. (2024). Comparing GPT-4 and human researchers in health care data analysis : Qualitative description study. Journal of Medical Internet Research, 26, e56500. https ://doi.org/10.2196/56500

[74] Levkovich, I., & Omar, M. (2024). Evaluating BERT-based and large language models for suicide detection, prevention, and risk assessment : A systematic review. Journal of Medical Systems, 48(1), 113. https ://doi.org/10.1007/s10916-024-02134-3

[75] Ameer, I., Bölücü, N., Siddiqui, M. H. F., Can, B., Sidorov, G., & Gelbukh, A. (2023). Multi-label emotion classification in texts using transfer learning. Expert Systems with Applications, 200, 118534. https ://doi.org/10.1016/j.eswa.2022.118534

[76] Krause, K. J., Davis, S. E., Yin, Z., Schafer, K. M., Rosenbloom, S. T., & Walsh, C. G. (2024). Enhancing suicide attempt risk prediction models with temporal clinical note features. Applied Clinical Informatics, 15(5), 1107–1120. https ://doi.org/10.1055/a-2411-5796

[77] Paul, M. J., & Dredze, M. (2011). You Are What You Tweet : Analyzing Twitter for Public Health. In Proceedings of the Fifth International AAAI Conference on Weblogs and Social Media (pp. 265–272). AAAI Press

[78] Meddeb, A., Ebert, P., Bressem, K. K., Desser, D., Dell'Orco, A., Bohner, G., Kleine, J. F., Siebert, E., Grauhan, N., Brockmann, M. A., Othman, A., Scheel, M., & Nawabi, J. (2024). Evaluating local open-source large language models for data extraction from unstructured reports on mechanical thrombectomy in patients with ischemic stroke. Journal of NeuroInterventional Surgery. Advance online publication. https ://doi.org/10.1136/jnis-2024-022078

[79] Wang, X., Kordi, Y., Mishra, S., et al. (2023). Self-Consistency Improves Chain-of-Thought Reasoning in Language Models. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 13758–13767.

[80] Zaheer, M., Guruganesh, G., Dubey, A., Ainslie, J., Alberti, C., Ontanón, S., … & Ahmed, A. (2020). Big Bird : Transformers for Longer Sequences. Advances in Neural Information Processing Systems, 33, 17283–17297.

[81] Kitaev, N., Kaiser, Ł., & Levskaya, A. (2020). Reformer : The Efficient Transformer. In International Conference on Learning Representations (ICLR).

[82] Wang, S., Li, B. Z., Khabsa, M., Fang, H., & Ma, H. (2021). Linformer : Self-Attention with Linear Complexity. In International Conference on Learning Representations (ICLR).

[83] Fatima-Azzahrae, A., Amraoui, R., Adda, M., Lessard, L. Automatic classification of psychosocial concerns : From traditional approach to deep learning.

[84] Amraoui, R., Adnane, F.-A., Adda, M., & Lessard, L. (2024). NLP and topic modeling with LDA, LSA, and NMF for monitoring psychosocial well-being in monthly surveys. Procedia Computer Science, 251, 398–405. https ://doi.org/10.1016/j.procs.2024.11.126

[85] Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., & 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. Retrieved from https ://arxiv.org/abs/1910.10683

[86] Han, S., Pool, J., Tran, J., & Dally, W. (2015). Learning both weights and connections for efficient neural networks. In Advances in Neural Information Processing Systems.

[87] Lan, Z., Chen, M., Goodman, S., Gimpel, K., Sharma, P., & Soricut, R. (2020). ALBERT : A Lite BERT for Self-supervised Learning of Language Representations. In International Conference on Learning Representations (ICLR).

[88] Molenaar, A., Lukose, D., Brennan, L., Jenkins, E. L., & McCaffrey, T. A. (2024). Using Natural Language Processing to Explore Social Media Opinions on Food Security : Sentiment Analysis and Topic Modeling Study. Journal of Medical Internet Research, 26, e47826. https ://doi.org/10.2196/47826

[89] Yao, S., Zhao, Y., Yu, D., et al. (2023). Tree of Thoughts : Deliberate Problem Solving with Large Language Models. Proceedings of the 40th International Conference on Machine Learning (ICML).

[90] Chung, H. W., Hou, L., Longpre, S., et al. (2022). Scaling Instruction-Finetuned Language Models. Proceedings of the 39th International Conference on Machine Learning (ICML). [ICML/Scopus]

[91] Strubell, E., Ganesh, A., McCallum, A. (2019). Energy and Policy Considerations for Deep Learning in NLP. Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (ACL), pp. 3645–3650.

[92] Weidinger, L., Uesato, J., Rauh, M., et al. (2022). Ethical and Social Risks of Harm from Language Models. Proceedings of the 2022 ACM

Conference on Fairness, Accountability, and Transparency (FAccT), pp. 1–12.
[93] Le, T.-D., Noumeir, R., Rambaud, J., Sans, G., & Jouvet, P. (2021). Machine learning based on natural language processing to detect cardiac failure in clinical narratives.
[94] Guyon, I., & Elisseeff, A. (2003). An Introduction to Variable and Feature Selection. Journal of Machine Learning Research, 3, 1157–1182.
[95] Chandrashekar, G., & Sahin, F. (2014). A Survey on Feature Selection Methods. Computers & Electrical Engineering, 40(1), 16–28.
[96] McCoy, T. H., & Perlis, R. H. (2025). Applying large language models to stratify suicide risk using narrative clinical notes. Journal of Medical Artificial Intelligence, 1(1), 100109. https ://doi.org/10.1016/j.xjmad.2025.100109
[97] Tinn, R., Cheng, H., Gu, Y., Usuyama, N., Liu, X., Naumann, T., Gao, J., & Poon, H. (2023). Fine-tuning large neural language models for biomedical natural language processing. Briefings in Bioinformatics, 24(5)
[98] Kowsari, K., Jafari Meimandi, K., Heidarysafa, M., Mendu, S., Barnes, L., & Brown, D. (2019). Text classification algorithms : A survey. Information, 10(4), 150.
[99] Chen, M., Luo, Y., & Xu, H. (2022). HealthPrompt : A prompting-based framework for zero-shot clinical text classification. arXiv preprint arXiv :2203.05061.
[100] Yadav, V., & Bethard, S. (2018). A survey on recent advances in named entity recognition from deep learning models. Proceedings of the 27th International Conference on Computational Linguistics, 2145-2158. https ://aclanthology.org/C18-1182/
[101] Ahadi, S. A., Jazayeri, K., & Tebyani, S. (2024). Detecting Suicidality from Reddit Posts Using a Hybrid CNN–LSTM Model. Journal of Universal Computer Science, 30(13), 1872–1904. https ://doi.org/10.3897/jucs.119828
[102] Ekolle, Z., & Kohno, R. (2023). GenCo : A generative learning model for heterogeneous text classification based on collaborative partial classifications. https ://doi.org/10.20944/preprints202306.0078.v1
[103] Lin, T., Wang, Y., Liu, X., & Qiu, X. (2021). A survey of transformers. https ://arxiv.org/abs/2106.04554
[104] Bunt, H. L., Goddard, A., Reader, T. W., & Gillespie, A. (2025). Validating the use of large language models for psychological text classification. Frontiers in Social Psychology, 3. https ://doi.org/10.3389/frsps.2025.1460277
[105] Luo, R., Sun, L., Xia, Y., Qin, T., Zhang, S., Poon, H., & Liu, T.-Y. (2022). BioGPT : Generative pre-trained transformer for biomedical text generation and mining. Briefings in Bioinformatics, 23(6), bbac409. https ://doi.org/10.1093/bib/bbac409
[106] Gu, A., Goel, K., & Ré, C. (2022). Efficiently Modeling Long Sequences with Structured State Spaces. In International Conference on Learning Representations (ICLR).
[107] Fanny, Y., Muliono, Y., & Tanzil, F. (2018). A comparison of text classification methods k-NN, Naïve Bayes, and Support Vector Machine for news classification. Jurnal Informatika : Jurnal Pengembangan IT, 3(2), 157–160. https ://doi.org/10.30591/jpit.v3i2.828
[108] Burkhardt, H. A., Ding, X., Kerbrat, A., Comtois, K. A., & Cohen, T. (2023). From benchmark to bedside : Transfer learning from social media to patient-provider text messages for suicide risk prediction. Journal of the American Medical Informatics Association, 30(6), 1068–1078. https ://doi.org/10.1093/jamia/ocad062
[109] Shazeer, N., Mirhoseini, A., Maziarz, K., Davis, A., Le, Q., Hinton, G., & Dean, J. (2017). Outrageously Large Neural Networks : The Sparsely-Gated Mixture-of-Experts Layer. In International Conference on Learning Representations (ICLR).
[110] Ahmed, S., Khurshid, S., Imran, M., Siddiqui, M. S., Hina, S., & Ahmed, M. (2024). Analysis of Mental Health Counseling Conversation Using Natural Language Processing. Journal of Computer Science, 20(3), 303-309.
[111] Wiest, I. C., Verhees, F. G., Ferber, D., Zhu, J., Bauer, M., Lewitzka, U., Pfennig, A., Mikolas, P., & Kather, J. N. (2024). Detection of suicidality from medical text using privacy-preserving large language models. BJPsych Features, Cambridge University Press. https ://doi.org/10.1192/bjp.2024.134
[112] MIslam, S., Elmekki, H., Elsebai, A., Bentahar, J., Drawel, N., Rjoub, G., & Pedrycz, W. (2023). A comprehensive survey on applications of transformers for deep learning tasks. https ://arxiv.org/abs/2306.06161
[113] Qasim, A., Mehak, G., Hussain, N., Gelbukh, A., & Sidorov, G. (2025). Detection of depression severity in social media text using transformer-based models. Information, 16(2), 114. https ://doi.org/10.3390/info16020114
[114] Chim, J., Tsakalidis, A., Gkoumas, D., Atzil-Slonim, D., Ophir, Y., Zirikly, A., Resnik, P., & Liakata, M. (2024, March). Overview of the CLPsych 2024 shared task : Leveraging large language models to identify evidence of suicidality risk in online posts. In A. Yates, B. Desmet, E. Prud'hommeaux, A. Zirikly, S. Bedrick, S. MacAvaney, K. Bar, M. Ireland, & Y. Ophir, Proceedings of the 9th Workshop on Computational Linguistics and Clinical Psychology (CLPsych 2024) (pp. 177–190). Association for Computational Linguistics. https ://aclanthology.org/2024.clpsych-1.15/

CHAPITRE 2
CLASSIFICATION AUTOMATIQUE DES PREOCCUPATIONS PSYCHOSOCIALES : DE L'APPROCHE TRADITIONNELLE A L'APPRENTISSAGE PROFOND

2.1 RESUME EN FRANÇAIS DU DEUXIEME ARTICLE

Cet article, intitulé « Automatic Classification of Psychosocial Concerns: From Traditional Approach to Deep Learning », a été accepté pour publication dans sa version finale en 2024 par les éditeurs de la revue Procedia Computer Science (Elsevier). Il est désormais accessible en ligne sous le DOI : 10.1016/j.procs.2024.11.125.

En tant que première auteure, j'ai contribué à l'essentiel de la recherche sur l'état de l'art, à la conception méthodologique et à la mise en œuvre des expérimentations portant sur l'ensemble des étapes de prétraitement des données, les modèles d'apprentissage automatique K-NN et Naives Bayes, et les différents modèles de type transformer testés et combinés aux techniques de fine-tuning et de few-shot learning. Ma collègue et co-auteure, Rkia Amraoui, a quant à elle pris en charge l'évaluation approfondie des algorithmes SVM et XGBoost, ainsi que l'implémentation de l'approche SetFit. Nous avons toutes deux participé au développement de l'application automatisant la classification des préoccupations psychosociales, sous la direction et la supervision du professeur Mehdi Adda, qui a également fixé les grandes consignes méthodologiques et validé l'ensemble du travail. Enfin, la professeure Lily Lessard a collaboré à la structure globale du projet et à la révision du contenu scientifique.

Ce travail s'inscrit dans le cadre du projet Vigie Psychosociale, et son objectif est de faciliter la détection et l'automatisation de la gestion des préoccupations psychosociales au

sein des différentes MRC dans une région du Québec. Une version abrégée de cet article a été présentée à la 14th International Conference on Current and Future Trends of Information and Communication Technologies in Healthcare (ICTH 2024), à Leuven (Belgique), du 28 au 30 octobre 2024, ce qui a permis de recueillir des retours constructifs avant sa publication finale.

2.2 CLASSIFICATION AUTOMATIQUE DES PREOCCUPATIONS PSYCHOSOCIALES : DE L'APPROCHE TRADITIONNELLE A L'APPRENTISSAGE PROFOND

Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 251 (2024) 390–397

Procedia Computer Science

www.elsevier.com/locate/procedia

The 14th International Conference on Current and Future Trends of Information and Communication Technologies in Healthcare (ICTH 2024)
October 28-30, 2024, Leuven, Belgium

Automatic Classification of Psychosocial Concerns: From Traditional Approach to Deep Learning

Adnane Fatima-Azzahrae[a,*], Amraoui Rkia[a], Adda Mehdi[a], Lessard Lily[b]

[a]*Département de Mathématiques, Informatique et Génie, Université du Québec à Rimouski (UQAR), Canada (QC)*
[b]*Département des sciences de la santé, Université du Québec à Rimouski (UQAR), Canada (QC)*

Abstract

The advent of artificial intelligence (AI) technologies presents promising prospects for analyzing short texts, significantly impacting the psychosocial health sector. The classification and categorization of texts represent arduous and time-consuming tasks, necessitating systematic automation to optimize the processing of traditional manual workflows. This paper presents a comparative study of various machine learning (ML) techniques in natural language processing (NLP). These techniques, designed to replace manual data categorization effectively, primarily rely on traditional algorithms such as K-Nearest Neighbors (k-NN), Support Vector Machines (SVM), and Extreme Gradient Boosting (XGBoost), as well as deep learning approaches, including fine-tuning, SetFit, and few-shot learning based on transformers.
A detailed analysis of different evaluation metrics revealed that the SetFit approach, integrating the sentence-transformer model, outperformed the best traditional models, with an average accuracy of 70.74% compared to 68.69 % achieved by the SVM model.

Keywords: Artificial intelligence; Natural language processing; Psychosocial concerns; Classification; Machine learning; Deep learning.

1. Introduction

In contemporary societies, psychosocial concerns are prevalent, reflecting the interplay between psychological factors and social influences. These concerns include workplace stress, social isolation, discrimination, housing shortages, and the psychological impacts of crises such as COVID-19, which profoundly affects well-being and psychosocial health. Traditionally, these concerns are manually classified, which is challenging due to their

* Corresponding author: Adnane Fatima-Azzahrae
E-mail address: fatimaazzahrae.adnane@uqar.ca

1877-0509
10.1016/j.procs.2024.11.125

unstructured, subjective, and contextual nature.
This study aims to explore machine learning and deep learning techniques for precise, rapid, and cost-effective classification of psychosocial concerns. The goal is to develop a desktop application for nonexperts that automatically classifies and labels these concerns, enabling faster issue identification and targeted interventions tailored to the needs of various Regional County Municipalities (RCMs) in Quebec. Machine learning, a broad field of AI, includes supervised, unsupervised, semi-supervised and reinforcement learning. Among these, text classification is fundamental, involving the assignment of a label or category to a given observation based on its features. The process begins with several text preprocessing steps: cleaning, normalization, tokenization, stemming, and lemmatization, which convert unstructured text into interpretable features. Following preprocessing, feature extraction techniques such as Bag of Words (BOW), Term Frequency-Inverse Document Frequency (TF-IDF) [7], word embeddings, and n-grams transform texts into numerical vectors for model processing. To further enhance the efficiency of the model, dimensionality reduction techniques such as PCA [13], LDA, random projection, and t-SNE are applied to simplify the models, reduce noise, and improve computational efficiency.

The classification process involves data preprocessing, algorithm training, evaluation, and model comparison. Traditional algorithms require manual feature extraction followed by classification using classical ML methods [1, 3]. In contrast, deep learning models integrate feature extraction into the learning process [8], using complex transformations to relate raw data features to outcomes. This approach more effectively captures the nuances of textual data, improving classification accuracy.

Expanding on this, transformer-based models have revolutionized NLP tasks utilizing self-attention mechanisms to process text sequences efficiently and in parallel [8]. Techniques such as fine-tuning, which adapts a pre-trained model to a specific task using task-specific data, leverage prior knowledge for improved performance even with limited data [5]. Few-shot learning, which includes methods such as meta-learning and transfer learning, trains models to generalize from a small number of examples, addressing the challenge of limited data availability [6]. SetFit, a specific few-shot learning approach, uses pre-trained sentence transformers adapted to new tasks with minimal data, making it ideal for scenarios with limited labeled data. These approaches are interconnected through their ability to maximize the use of existing pre-trained models and adapt them to new, specific tasks with minimal additional data, significantly enhancing the versatility and efficiency of NLP models.

In this study, we compare traditional algorithms such as SVM [3], k-NN [2], XGBoost [4], and Naïve Bayes, which are statistical methods relying on the assumption of an underlying probabilistic model that describes the data, with these advanced deep learning models. Their effectiveness in managing psychosocial concerns is evaluated by examining algorithm type, classification accuracy, and processing time.

The article is structured as follows. Section 2 provides a review of existing short text classification methods and covers the relevant literature related to our study. In Section 3, we detail the approach and methodologies employed in our research. Section 4 presents the experimental results and discusses the implications of our results and their significance. Finally, Section 5 concludes the article and suggests directions for future research.

2. Literature Review

In the complex web of human concerns and social interactions, AI has become a powerful tool for understanding and addressing psychosocial issues. The automatic classification of these concerns has advanced significantly with the development of AI and NLP. This review explores a range of methodologies, from traditional machine learning to deep learning techniques, each contributing uniquely to the detection and classification of psychosocial concerns.
Psychosocial issues, whether they involve stress, anxiety, or challenges related to social and professional environments, play a central role in individual well-being. Several technological approaches, beyond NLP, have emerged to better detect and assess these concerns. For instance, the mStress study [22] utilizes wearable sensors and smartphones to collect physiological and subjective data for continous stress monitoring. While offering valuable automation, this method has limitations, including sensor intrusiveness, which affects user comfort and data quality, and the difficulty in distinguishing stress signals from other physical activities, leading to inaccurate results. Moreover, the high energy consumption of these devices limits their scalability for prolonged use.
These challenges underscore the need for non-invasive, scalable alternatives. Text data analysis offers a promising solution, drawing on sources such as self-reports, surveys, and social media. NLP models and deep learning

techniques, like transformers and few-shot learning, enable precise, automated classification of psychosocial concerns. These methods not only overcome the limitations of sensor-based approaches but also provide greater scalability and deeper insights by capturing the nuances of text.
NLP methods have proven effective in specific contexts, such as evaluating counseling sessions. The study [20] uses machine learning models such as Naive Bayes, SVM, and BERT to analyze text-based communications between patients and therapists. Although promising, the results reveal significant limitations, particularly BERT's lower performance compared to simpler models like Naive Bayes and SVM, due to the small size and uneven quality of the data, a recurring issue in mental health where confidentiality complicates access to large datasets. Furthermore, the use of metrics based on patient satisfaction rather than the actual effectiveness of interventions introduces bias in evaluating interactions, highlighting a lack of robustness in the criteria used to measure therapeutic quality.
Given these limitations, it is crucial to thoroughly understand the functioning and appropriate use of each machine learning model to ensure their correct application in the right situations. The article [10] provides a comprehensive overview of text classification methods, categorizing them into traditional machine learning approaches and deep learning techniques. While traditional models such as k-NN, SVM, XGBoost, and Naive Bayes have been foundational for text classification, they exhibit significant limitations. These models often require manual feature selection and struggle to capture complex relationships between words in short texts. For instance, the article [19] developed a hybrid classification method to evaluate and predict psychosocial risk levels among public schooll teachers in Colombia using epidemiological data. Their approach combined Support Vector Machines (SVM) with a hill-climbing optimization algorithm, enhancing the model's performance by reducing the dimensionality of the dataset and improving prediction accuracy. These findings underscore the importance of optimizing machine learning models and understanding each model's strengths and limitations, particularly in complex datasets where multiple factors are involved.ed.
Moreover, The presence of imbalanced data and small datasets, has prompted researchers to explore advanced deep learning approaches. The article [8] introduced the groundbreaking Transformer architecture, which significantly enhanced the ability of models to capture the contextual nuances of text. Transformer-based models, such as BERT, have set new standards for understanding and processing natural language. The article [21], assume that in a specific linguistic pattern, psychosocial features can provide more information for short text classification. It examines the BERT method for this purpose and includes a section on fine-tuning BERT for improved performance.

Upon critically examining these methodologies, it is evident that many existing approaches face significant challenges, particularly when dealing with imbalanced datasets, small sample sizes, and the contextual nuances inherent to psychosocial concerns. Traditional models, while effective in certain cases, often require manual feature selection and lack the flexibility to manage complex textual relationships. However, transformer-based models like BERT, though promising, struggle with small datasets and tend to overfit [20]. These challenges, underscore the importance of refining existing models to better capture such nuances, and our study aims to address these limitations. By leveraging advanced techniques, such as fine-tuning transformers on smaller datasets and employing strategies to address data imbalance, our aim is to improve classification accuracy and scalability in the context of psychosocial concerns.

3. Methodology

3.1. Preprocessing and Model Optimization Strategies

This section details the methodology adopted to establish the classification process for qualitative concerns collected by the 'Vigie'. The fundamental objective of this application is to automate classification for healthcare personnel, thereby meeting various requirements throughout the process.
To achieve this, several steps were undertaken. First, at the interface level, functionalities such as sorting, filtering, adding, and deleting data were implemented. Additionally, automatic spell correction was integrated, ultimately opting for Speller, Figure 1.

Preoccup	Preoccup_corrigee
L'épuisement des enseigants jumélé au manque de personn...	L'épuisement des enseignants jumelé au manque de per
Difficulté à subvenir à leur besoin primaire au niveau fiancier	Difficulté à subvenir à leur besoin primaire au niveau financier

Fig. 1. Automatic Spell Correction.

Subsequently, we utilized SQL databases for automatic abbreviation management, enabling the conversion of abbreviations into complete phrases to ensure more effective learning by the classification model, given that the nature of the data initially relies on numerous abbreviations. In the context of data privacy management, an approach to anonymization and named entity recognition was explored to ensure the protection of sensitive information. This approach aims to mask or remove personally identifiable information present in the collected data while preserving the integrity and analytical value of the data. Anonymization was achieved using advanced techniques, such as pseudonymization, where personal data are replaced with pseudonyms, and generalization, where specific details are replaced with more general information. To ensure that the approach was well-founded, several techniques for named entity recognition were evaluated. Among these, SpaCy was chosen for its balance of speed and accuracy. BERT, though highly accurate, was not selected due to its higher computational demands.

In conclusion, to ensure optimal performance of the classification model, it is essential to follow a series of data preprocessing steps. Various combinations of techniques were tested and are detailed in Figure 2.

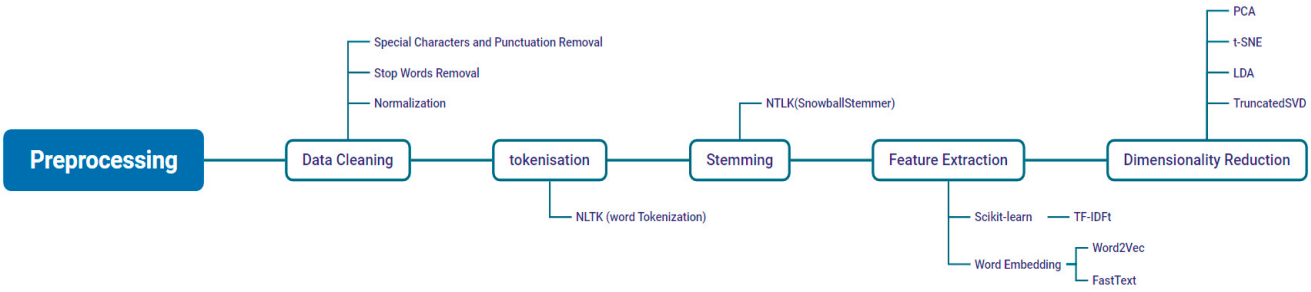


Fig. 2. Preprocessing Pipeline and Model Optimization.

During the preprocessing stage, we analyzed the distribution of token lengths to understand the variability in text lengths within our dataset. The analysis showed that most tokens ranged between 0 and 60, with a concentration of around 30 tokens. This understanding was critical for determining the appropriate preprocessing techniques to apply, ensuring uniformity in subsequent analyses.

Normalization techniques were followed by stemming, which was chosen over lemmatization for its ability to efficiently group word forms by their roots [16]. Stemming was more appropriate for our dataset, where capturing the essential meaning of short, context-specific concerns was more important than maintaining grammatical accuracy. While lemmatization can be useful in certain tasks, it introduced unnecessary distinctions in our case, which did not improve the model's performance. Stemming, by simplifying the model and reducing the risk of overfitting, allowed us to focus on the core meaning of the text. TF-IDF was then used to quantify the importance of each word in the dataset, taking into account both its frequency in individual texts and its rarity across the entire corpus.[7]. Finally, dimensionality reduction using Truncated SVD preserved essential information while efficiently managing the high-dimensional data generated by TfidfVectorizer.

3.2. Comparative Evaluation of Traditional Algorithms

To better understand the characteristics of our data and ensure optimal preparation for model training, an exploratory analysis of the data was conducted. This analysis aimed to identify any potential imbalances in the data, which could influence the performance of classification algorithms.

The class imbalance posed a particular problem because the number of nearby neighbors was insufficient to effectively apply techniques such as SMOTE [11] and ADASYN [12]. To address this issue, we initially attempted to augment the data with synonyms using NLP techniques such as WordNet, thereby increasing data diversity. However, more traditional oversampling methods, like RandomOverSampler, were ultimately adopted. Although this method poses a risk of overfitting due to duplication, this risk was mitigated by closely monitoring the overfitting curve.
To transform categorical labels into numerical representations suitable for ML algorithms, we employed LabelEncoder, a simple and effective method for classification targets. OneHotEncoder, which converts each category into a distinct binary column, was deemed irrelevant for this study due to the absence of multicategorical columns. Subsequently, the combination of LabelEncoder and TfidfVectorizer was used to ensure adequate representation of the data, optimizing model training. To maintain the proportional distribution of classes in each subset, we used cross-validation, specifically stratified k-fold cross-validation. This method involves dividing the dataset into multiple subsets and training the models with different combinations of these subsets while testing them on the remaining parts.
To optimize model hyperparameters, we used GridSearch to test a range of values and identify optimal configurations while minimizing overfitting. We specifically tuned parameters such as the learning rate, number of layers, and early stopping criteria. The hyperparameters for each algorithm are shown in Table 1.

Table 1. Explored Hyperparameters for Different Algorithms

Method	**Description**
K-NN	{'num neighbors': 3, 'weights': 'distance'}
SVM	{'C': 10, 'gamma': 0.1, 'kernel': 'rbf'}
Naives Bayes	{'alpha': 0.1}
XGBoost	{'eval metric': 'merror', 'max depth': 3, 'num class': 43, 'objective': 'multi:softmax', 'reg alpha': 0.01, 'reg lambda': 0.01}

3.3. Transition to Deep Learning

Although traditional models such as K-NN, SVM, and Naive Bayes have been useful in certain tasks [10], their reliance on syntactic features limited their ability to capture contextual nuances critical for classifying psychosocial concerns. These models tend to focus on surface-level patterns, which makes them less suited for tasks requiring an understanding of more complex, multi-faceted concerns. In this study, the goal is to automate the classification of psychosocial concerns while maintaining contextual accuracy. To address this, deep learning techniques, including fine-tuning transformer models [5], SetFit for sentence transformers, and Few-Shot Learning [6], were explored. These models were chosen because they are capable of capturing deeper semantic relationships, which are essential for understanding the complexity and diversity of the concerns in our dataset. By leveraging these advanced techniques, we aim to develop a classification process that better handles contextual intricacies. The workflow for these methods is illustrated in Figure3.

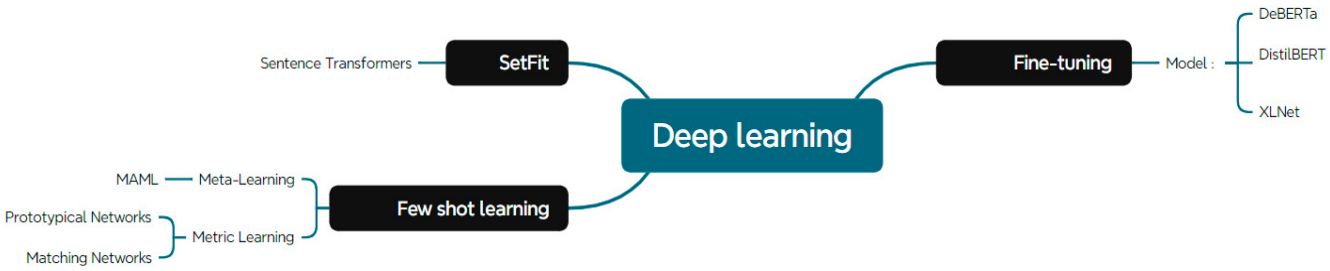


Fig. 3. Techniques and Models for Deep Learning: Few-Shot Learning, SetFit, and Fine-Tuning

In this study, we implemented the basic versions of various models using libraries like Hugging Face Transformers, PyTorch, and TensorFlow, which support advanced computations and pre-trained models. For methods such as

Fine-Tuning, SetFit, and Few-Shot Learning, the data was first preprocessed as described in Section 3.1. Afterward, we prepared the datasets and tokenized the texts using specific tokenizers, such as DebertaTokenizer, DistilBertTokenizer, and XLNetTokenizer. SetFit, on the other hand, relied on a sentence transformer to process the data. This Transformers handled the automatic extraction of features from the text and the models were initialized with pre-trained transformers and meta-learning models for Few-Shot Learning.

To fine-tune the models, we explored various settings using Optuna [14], a tool that helps find the best parameters. We adjusted factors such as learning rate, batch size, and the number of training epochs. This process involved running 20 trials to identify the optimal configurations for transformer models. The results of these tests are shown in Table 2

Table 2. Explored Hyperparameters for Different Transformer Models

Method	**Description**
DeBERTa	{learning_rate: 9.875875111798571e-05, batch_size: 32, num_epochs: 6, weight_decay: 0.052439306268401924}
DistilBERTa	{learning_rate: 8.023469704287049e-05, batch_size: 16, num_epochs: 7, weight_decay: 0.15315293911284625}
XLNet	{learning_rate: 5.319751979763443e-05, batch_size: 32, num_epochs: 7, weight_decay: 0.15383733793168392}
Few-Shot Learning Bert(MAML)	{inner_lr: 4.767808275790072e-05, meta_lr: 0.0004396663860878128, num_inner_steps: 6, num_epochs: 10, k_shot: 7, query_size: 17}

4. Results and Discussion

Evaluation metrics play a crucial role in assessing the performance of classification models. They enable the quantification of the model's prediction quality and facilitate comparisons between different models. The key metrics used include precision, recall, F1-score, and the classification report, each providing a distinct perspective on the model's performance.

Table 3. Precision, Recall, and F1 score for different models.

Models	**Accuracy**	**Weighted**		
		Precision	**Recall**	**F1-score**
k-NN + TF-IDF	0.6404	0.6379	0.6404	0.6326
k-NN + Word2vec	0.3119	0.3908	0.3119	0.3398
SVC + TF-IDF	0.6869	0.6912	0.6869	0.6805
SVC + Word2vec	0.1504	0.4200	0.1504	0.2001
Naive Bayes(MultinomialNB) + TF-IDF	0.6360	0.6659	0.6360	0.6454
XGBoost + TF-IDF	0.6592	0.6656	0.6592	0.6583
XGBoost + Word2vec	0.3561	0.3707	0.3561	0.3571
Fine-Tuning DeBERTa	0.7011	0.7125	0.7011	0.7067
Fine-Tuning DistilBERTa	0.7068	0.7206	0.7068	0.7136
Fine-Tuning XLNet	0.6870	0.7013	0.6870	0.6940
SetFit (Sentence-Transformers)	0.7074	0.7250	0.7074	0.7160
BERT + Few-Shot Learning MAML	0.6777	0.6859	0.6777	0.6749

The results summarized in Table 3 show a clear trend, models utilizing TF-IDF consistently outperform those using Word2Vec. For example, SVC + TF-IDF achieved an accuracy of 68.69 %, while SVC + Word2Vec only achieved 25.04 %. This superior performance of TF-IDF can be attributed to its strength in handling sparse, context-specific textual data, typical of psychosocial concerns, where key terms must be emphasized based on their document frequency. In contrast, Word2Vec's dense embeddings, though beneficial for capturing semantic similarities, may

oversimplify the nuanced context of these concerns, resulting in a loss of critical information essential for accurate classification.

Moreover, While deep learning models, particularly transformers, outperformed traditional machine learning methods, the improvements, though notable, did not fully meet expectations. One potential explanation lies in the quality of the baseline labels used for training. If these labels lack sufficient granularity or precision, they may hinder the models' ability to fully exploit the rich information present in transformer-based embeddings. For instance, the models achieved accuracies of (70.11%) (DeBERTa) and (70.74%)(SetFit), respectively, yet they may still struggle to capture the complex subtleties inherent in the dataset. Psychosocial concerns often involve multifaceted issues expressed in indirect or ambiguous language, and a misalignment between the labels and the data's true structure can prevent the models from reaching their full potential. Revising these labels to ensure greater relevance and granularity would likely enhance classification accuracy.

Although The few-shot learning approach, specifically using MAML, was intended to address the challenges of small and imbalanced datasets by allowing the model to generalize from limited examples. However, its performance (67.77%) fell short of expectations, possibly due to the nuanced and context-specific nature of psychosocial concerns. Vague or poorly defined concerns, such as "Population plus fatiguée," lack the detail needed for accurate classification, making it difficult for models like MAML to establish generalizable patterns. While transformers excel at managing subtle contextual shifts and dependencies, even they can struggle with underspecified inputs. Additionally, the dataset's class imbalance likely exacerbated MAML's limitations, as few-shot learning methods tend to focus disproportionately on more frequent classes, neglecting minority ones. Transformers, with their advanced feature extraction and capacity to learn complex patterns, handle class imbalances more effectively.

From a practical standpoint, these results suggest that while advanced models like transformers (DeBERTa, DistilBERT, and sentencetransformer) provide notable improvements, there is still a need to refine baseline labels and develop techniques that can better handle the specific challenges of imbalanced and nuanced data. Incorporating more sophisticated preprocessing techniques or combining multiple models may further enhance classification accuracy, allowing healthcare professionals to use more precise tools for addressing psychosocial concerns.

5. Conclusion

The rapid advancements in ML for textual data classification hold promises for the field of psychosocial health. By utilizing advanced ML techniques, the developed application can streamline the management of psychosocial concerns, enabling efficient processing and analysis. This, in turn, will enhance community resilience and help reduce health disparities. However, despite these promising developments, our study highlights several key limitations. The class imbalance in the dataset poses a significant challenge, impacting the models' ability to generalize effectively. While various resampling techniques were tested, the results remained suboptimal. Additionally, although deep learning models have demonstrated strong performance, their high computational cost limits their applicability in resource-constrained environments.

To address these challenges, it is crucial to conduct a thorough validation of the labels in collaboration with experts from Vigie-Psychosociale, ensuring that the labels accurately reflect real-world concerns. Moreover, our analysis has revealed the need to adopt a multi-label classification approach. While some concerns can be categorized under a single label, others require multiple labels, a complexity that current models struggle to handle effectively. Furthermore, it is important to recognize that the evolving nature of concerns expressed by the population, rooted in real data, changes over time. Traditional models, which rely on training with static data, are not well-suited to handle the emergence of new labels as they appear. Looking ahead, we aim for the model to autonomously suggest new labels in response to emerging topics, leading us to consider the integration of large language models (LLMs). LLMs not only enable the generation of labels, but also possess the flexibility to adapt to the continuous evolution of psychosocial concerns. Their ability to contextualize complex issues ensures that the model remains relevant as new challenges arise within the psychosocial context.

References

[1] Peterson, L. E. (2009). K-nearest neighbor. Scholarpedia, 4(2), 1883.

[2] Zhang, Z., Wang, J., & Wang, H. (2020). Improved K-nearest neighbors algorithm based on a heuristic method for imbalanced data classification. IEEE Access, 8, 55505-55517.

[3] Caragea, C., Silvescu, A., & Mitra, P. (2014). Combining Bag-of-Words and Graph Dependencies in a Support Vector Machine for Categorizing Text with Rich Features. Proceedings of the 23rd ACM International Conference on Information and Knowledge Management (CIKM), 211-220.

[4] Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining.

[5] Howard, J., & Ruder, S. (2018). Universal language model fine-tuning for text classification. Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics.

[6] Zhou, P., Zhang, J., Zong, H., Zhang, C., & Yang, H. (2022). A Comprehensive Survey on Few-Shot Learning for Natural Language Processing.

[7] Kuang, Q., & Xu, X. (2010). Improvement and Application of TF·IDF Method Based on Text Classification. In Proceedings of the 2010 International Conference on Internet Technology and Applications

[8] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., & Polosukhin, I. (2017). Attention is all you need. Advances in Neural Information Processing Systems.

[9] Yang, R., Sun, L., Yu, P. S., & He, L. (2022). A survey on text classification from traditional to deep learning. ACM Transactions on Intelligent Systems and Technology, 13(2), Article 31.

[10] Kowsari, K., Jafari Meimandi, K., Heidarysafa, M., Mendu, S., Barnes, L., & Brown, D. (2019). Text classification algorithms: A survey. Information, 10(4), 150.

[11] Fernández, A., García, S., & Herrera, F. (2018). SMOTE for Learning from Imbalanced Data: Progress and Challenges, Marking the 15-year Anniversary. Journal of Artificial Intelligence Research, 61, 863-905.

[12] He, H., Bai, Y., Garcia, E. A., & Li, S. (2008). ADASYN: Adaptive synthetic sampling approach for imbalanced learning. In 2008 IEEE international joint conference on neural networks (IEEE world congress on computational intelligence) (pp. 1322-1328). IEEE.

[13] Abdi, H., & Williams, L. J. (2010). Principal component analysis. Wiley Interdisciplinary Reviews: Computational Statistics, 2(4), 433-459.

[14] Akiba, T., Sano, S., Yanase, T., Ohta, T., & Koyama, M. (2019). Optuna: A next-generation hyperparameter optimization framework.In Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining (pp. 2623-2631).

[15] Kowsari, K., Jafari Meimandi, K., Heidarysafa, M., Mendu, S., Barnes, L., & Brown, D. (2019). Text Classification Algorithms: A Survey. Information, 10(4), 150. MDPI.

[16] Singh, J., Gupta, V. (2017). A systematic review of text stemming techniques. Artificial Intelligence Review, https://doi.org/10.1007/s10462-016-9498-2.

[17] Min, B., Ross, H., Sulem, E., Veyseh, A. P. B., Nguyen, T. H., Sainz, O., Agirre, E., Heintz, I., & Roth, D.(2021). Recent Advances in Natural Language Processing via Large Pre-trained Language Models: A Survey.

[18] Dien, T. T., Loc, B. H., & Thai-Nghe, N. (2019). Article classification using natural language processing and machine learning. In 2019 International Conference on Advanced Computing and Applications (ACOMP) (pp. 1-6). IEEE.

[19] Mosquera, R., Gómez, C., Parra Osorio, L., & Carrión García, A. (2018). Classification system for the predicting of psychosocial risk level in public-school teachers based on Artificial Intelligence. In XVIII Conferencia de la Asociación Española para la Inteligencia Artificial (CAEPIA 2018).

[20] Ahmed, S., Khurshid, S., Imran, M., Siddiqui, M. S., Hina, S., & Ahmed, M. (2024). Analysis of Mental Health Counseling Conversation Using Natural Language Processing. Journal of Computer Science, 20(3), 303-309.

[21] Hu, Y., Ding, J., Dou, Z., & Chang, H. (2022). Short-Text Classification Detector: A Bert-Based Mental Approach. Computational Intelligence and Neuroscience, 2022, Article ID 8660828.

[22] Raij, A., Blitz, P., Ali, A. A., Fisk, S., French, B., Mitra, S., Nakajima, M., Nguyen, M. H., Plarre, K., Rahman, M., Shah, S., Shi, Y., Stohs, N., al'Absi, M., Ertin, E., Kamarck, T., Kumar, S., Scott, M., Siewiorek, D., & Smailagic, A. (2010). mStress: Supporting Continuous Collection of Objective and Subjective Measures of Psychosocial Stress on Mobile Devices. In Proceedings of the Wireless Health 2010 Conference.

[23] Siino, M., Tinnirello, I., & La Cascia, M. (2023). Is text preprocessing still worth the time? A comparative survey on the influence of popular preprocessing methods on Transformers and traditional classifiers. Information Systems, 121, 102342.

CHAPITRE 3
NLP ET MODELISATION THEMATIQUE AVEC LDA, LSA ET NMF POUR LE SUIVI DU BIEN-ETRE PSYCHOSOCIAL DANS DES ENQUETES MENSUELLES

3.1 RESUME EN FRANÇAIS DU TROISIEME ARTICLE

Cet article, intitulé « NLP and Topic Modeling with LDA, LSA, and NMF for Monitoring Psychosocial Well-being in Monthly Surveys », a été accepté pour publication dans sa version finale en 2024 par les éditeurs de la revue Procedia Computer Science (Elsevier). Il est désormais accessible en ligne sous le DOI : 10.1016/j.procs.2024.11.126.

En tant que co-auteure, j'ai principalement pris en charge l'implémentation et l'optimisation du modèle LDA, tout en participant à la comparaison globale et à l'interprétation finale des résultats de la modélisation thématique. La première autrice, Rkia Amraoui, s'est chargée du prétraitement complet des données ainsi que de l'implémentation de LSA et NMF, permettant une comparaison approfondie avec LDA. Le professeur Mehdi Adda, auteur correspondant, a assuré la direction scientifique du projet et la supervision des étapes de validation, tandis que la professeure Lily Lessard a contribué au cadrage du projet dans le contexte de la santé publique et à la révision finale de l'article.

Ce travail s'inscrit dans la continuité des recherches menées dans le cadre du projet Vigie Psychosociale, visant à améliorer le suivi et l'analyse des problématiques psychosociales au sein des MRC dans une région du Québec. Plus spécifiquement, l'article propose une approche combinant NLP et modélisation thématique (LDA, LSA, NMF) afin d'identifier et de suivre l'évolution des préoccupations exprimées mensuellement par la population. Les résultats démontrent l'apport de ces méthodes pour repérer des thématiques,

fournissant ainsi aux décideurs des indicateurs pour adapter les services de santé et de soutien. Une version abrégée de cet article a été présentée à la 14th International Conference on Current and Future Trends of Information and Communication Technologies in Healthcare (ICTH 2024), à Leuven (Belgique), du 28 au 30 octobre 2024, ce qui a permis de recueillir des retours constructifs avant sa publication finale.

3.2 NLP ET MODELISATION THEMATIQUE AVEC LDA, LSA ET NMF POUR LE SUIVI DU BIEN-ETRE PSYCHOSOCIAL DANS DES ENQUETES MENSUELLES

Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 251 (2024) 398–405

Procedia Computer Science

www.elsevier.com/locate/procedia

The 14th International Conference on Current and Future Trends of Information and Communication Technologies in Healthcare (ICTH 2024)
October 28-30, 2024, Leuven, Belgium

NLP and Topic Modeling with LDA, LSA, and NMF for Monitoring Psychosocial Well-being in Monthly Surveys

Amraoui Rkia[a,*], Adnane Fatima-Azzahrae[a], Adda Mehdi[a], Lessard Lily[b]

[a]*Département de Mathématiques, Informatique et Génie, Université du Québec à Rimouski (UQAR), Canada (QC)*
[b]*Département des sciences de la santé, Université du Québec à Rimouski (UQAR), Canada (QC)*

Abstract

This article presents an approach for assessing psychosocial concerns using Natural Language Processing (NLP) and topic modeling on text data collected in monthly surveys. We processed a dataset containing more than 10,000 entries from two regional public health department, focusing on psychosocial concerns expressed by the population. Using NLP techniques and topic models like Latent Dirichlet Allocation (LDA), Latent Semantic Analysis (LSA), and Non-Negative Matrix Factorization (NMF), we identified and analyzed recurring themes. The study found that LDA with unigrams performed best, yielding a coherence score of 0.59, while NMF was less effective. Key emerging themes included emotional well-being, stress, and social isolation, which evolved over time, especially during the COVID-19 pandemic. The results demonstrate that these methods can identify emerging issues and provide valuable information for decision making.

Keywords: Psychosocial well-being; Topic modeling; Natural language processing; Latent Dirichlet Allocation; Latent Semantic Analysis; Non-Negative Matrix Factorization; Text analysis; Textual data; psychosocial concerns

1. Introduction

In a constantly evolving world, psychosocial well-being is crucial to the health and quality of life of individuals and society[4]. It encompasses various complex dimensions, including emotions, interpersonal relationships, work stress[31], social integration[32], and the psychological impact of global health crises like the COVID-19 pandemic[5]. During the pandemic, a regional public health department created a monthly monitoring tool to track community psychosocial recovery. The goal was to adjust the provision of health and social services as needed. Mon-

* Corresponding author.
E-mail address: Rkia.Amraoui@uqar.ca

1877-0509
10.1016/j.procs.2024.11.126

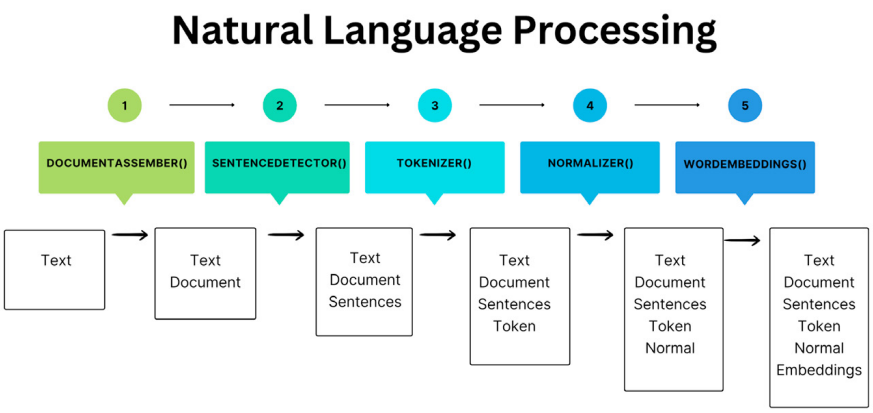


Fig. 1: NLP Pipeline

itoring these dimensions involves analyzing large volumes of unstructured textual data, a task that is both expensive and time-consuming when performed manually by professionals. For example, the regional public health department is tasked with manually categorizing psychosocial concerns and tracking their evolution, a laborious process that underscores the need for automated solutions.

To address this challenge, we explore the application of machine learning tools to assist health network decision-makers in monitoring and analyzing societal needs more efficiently. Machine learning (ML) tools are particularly suited for this task because of their ability to process and analyze large volumes of unstructured data quickly, which would otherwise be impractical to manage manually[6, 7]. NLP techniques, combined with topic modeling, enable the structuring of unstructured text data by automating tasks such as text cleaning, normalization, and tokenization [8, 9]. This facilitates the extraction of meaningful semantic features [11, 12], as illustrated in Figure 1.

Topic modeling, particularly methods like LDA and NMF, extracts latent themes from large data sets, offering a concise way to identify emerging psychosocial trends [1, 14, 15]. These methods constitute a favored approach to understanding psychosocial trends and are now being used in various domains, including psychosocial research. For example, recent studies have applied NLP and topic modeling to analyze mental health discussions on social media and understand the psychological impacts of global crises[24, 25, 27]. While these methods have been applied in various fields, their use in the context of psychosocial health remains underexplored.

The central objective of our study is to address the question: How can topic modeling be used effectively to monitor psychosocial well-being in a community through textual data analysis?In this sense, we investigate assuming that the application of topic models, namely LDA, LSA, and NMF, reveals significant trends and useful information at the decision-maker level. Key variables include model parameters (for example, priors, learning rate, iterations), number of topics, and coherence scores, which influence the quality and interpretability of generated topics. The number of topics determines the granularity of the themes identified in the text. Coherence scores are used to measure the interpretability and semantic consistency of the generated topics. These variables were selected because they have a direct impact on the performance and relevance of the model output in reflecting meaningful psychosocial themes.

This study introduces the application of topic modeling to monitor psychosocial well-being, providing a comprehensive comparison of LDA, LSA, and NMF models, with particular emphasis on hyperparameter tuning and coherence score analysis. It provides insights into tracking psychosocial trends from unstructured data and introduces a methodology to integrate NLP techniques into public health monitoring processes.

This article is organized as follows. Section 2 reviews the literature. Section 3 outlines the methodology. The results are detailed in Section 4, followed by a discussion in Section 5. Section 6 concludes with the main findings and recommendations for future research.

2. Literature Review

Topic modeling methods, including LDA and NMF, have been applied to extract useful patterns from large volumes of text data, revealing underlying themes and structures. For instance, Blei et al. [1] introduced LDA as a powerful method for discovering thematic structures in text, while Griffiths and Steyvers [2] and Alsumait et al. [3] explored topic modeling to detect and track emerging topics in scientific literature and text streams.

In the health domain, topic modeling has been applied to understand patient needs, monitor health trends, and detect disease symptoms. Paul and Dredze [20] used topic modeling to analyze health-related social media posts, demonstrating its utility in capturing public health concerns. Jelodar et al. [13] applied a neural network-based approach to identify COVID-19 topics from online discussions, showcasing the relevance of these methods in critical health contexts.

Recently, there has been a growing interest in using topic modeling to explore psychosocial concerns, particularly in mental health. For example, Molenaar et al. [16] used sentiment analysis and topic modeling to explore public opinion on food security on social media, focusing on its psychosocial implications. Zhang et al.[17] reviewed the application of topic modeling to detect mental illnesses, emphasizing its potential to identify underlying psychological conditions from text data.

In a similar vein, Saha et al.[24] explored social networks as a passive sensor to identify mental health concerns through topic modeling, specifically addressing psychological health trends. Resnik et al.[25] investigated supervised topic modeling approaches to detect depression-related language on Twitter, demonstrating the effectiveness of the model in the psychosocial domain. Guntuku et al.[26] provided a critical review of topic modeling applications for stress and mental health problems detected through online platforms.

In addition, Yang et al.[27] applied topic modeling to mental health discussions during the COVID-19 pandemic, highlighting its value in extracting mental health concerns from social media posts. These studies suggest that topic modeling is effective not only for analyzing general health trends but also for capturing more specific psychosocial issues such as stress, depression, and other mental health concerns.

In summary, topic modeling has demonstrated significant utility within the health domain, particularly for the identification and monitoring of themes pertaining to patient concerns, mental health, and psychosocial well-being. However, challenges remain, such as selecting the optimal number of topics and ensuring the privacy and ethical use of sensitive health data. The number of topics significantly affects the interpretability and quality of the results. If too few topics are selected, important nuances in the data may be lost as diverse themes are merged together. In contrast, selecting too many topics can result in overly granular themes, where meaningful patterns are fragmented across multiple topics, making interpretation more difficult. This balancing act is critical, as the choice of the number of topics influences the model's ability to uncover coherent and relevant themes from the data. Researchers often rely on metrics such as coherence scores to evaluate the quality of topics, but these measures do not always align perfectly with human interpretability, making this a persistent challenge in topic modeling research [1, 2].

In the following sections, we describe the key steps of our methodology in detail. First, we outline the data collection process, followed confidentialily and privacy considerations taken. Next, we describe the tools and techniques employed for text processing and analysis, focusing on data loading and preprocessing methods that were customized for our dataset. Finally, we explain the topic modeling techniques used.

3. Methodology

To effectively monitor psychosocial well-being, we employed a comprehensive methodology that supports the development of an AI-assisted text categorization application to provide insights into psychosocial well-being trends. Figure 2 outlines the process from data collection to topic modeling and visualization of results, highlighting key steps such as customized preprocessing and the use of LDA, NMF and LSA models.

In the following subsections, we detail the key steps of our methodology. The first two subsections outline the data collection process and the tools and techniques used for text processing and analysis, with a focus on the customized data loading and preprocessing methods for our dataset. The final subsection explains the topic modeling techniques employed, providing a rationale for choosing LDA, NMF, and LSA models, and discussing the potential use of more advanced models like BERT and GPT in future research.

3.1. Data Collection

The first stage, as shown in Figure 2, involves collecting textual data. This data is gathered monthly from two regional public health departments as a part of their surveillance mission, which tracks psychosocial concerns in communities through a sentinel network of citizens and community partners. Concerns data is qualitative textual data. Data is provided to our team in .csv or .xlsx file formats. The tables contain several columns, among which are 'anonymous id', 'MRC' (a territory grouping some ten municipalities), 'Date-vigie' (date), 'Preoccup' (concerns), 'Categorie-preoccup-nv1' (first-level categorization), and 'Categorie-preoccup-nv2' (second-level categorization). Our team worked with these anonymized documents and is committed to maintaning data confidentiality. Our team is not involved in the analysis, only their categorization.

3.2. Measurement Instruments

We made extensive use of Python text processing and analysis tools. we selected Pandas (v1.3.3) for its efficient data structures and ability to handle large datasets. We used NLTK (v3.6.5), SpaCy (v3.1.3), and Gensim (v4.1.2), covers SpaCy for advanced tokenization and lemmatization, as well as Gensim for a good LDA implementation. was chosen for its versatility in text vectorization and topic modeling, particularly with LDA and NMF. For visualizations, we used Matplotlib (v3.4.3), Seaborn (v0.11.2), and Bokeh (v2.3.3), along with WordCloud (v1.8.1) for generating thevisual word cloud summaries[10].

3.3. Data Loading and Preprocessing

The preprocessing of the textual data involved several custom steps to suit the characteristics of the data set. Stopwords were removed using a French stopwords list tailored to include additional terms specific to the dataset, such as 'etc,' 'chez,' 'plus,' and 'auprès.' Tokenization was performed using SpaCy's French model, ensuring the accurate segmentation of words and phrases common in French-language psychosocial concerns. Special handling of abbreviations was implemented by replacing them with their full forms using a SQLite database. In addition, spelling corrections were made with a French spell checker. Numerical values and punctuation were systematically removed to focus on the semantic content of the texts. Finally, lemmatization was applied to reduce words to their base forms, enhancing the coherence of topic models.

3.4. Topic Modeling

The processed data are then passed into the topic modeling stage. In selecting LDA, NMF, and LSA for our study, we drew on their proven track record in uncovering hidden patterns and themes within large text datasets. Specifically, LDA excels at modeling probabilistic topic structures[1], while NMF excels at identifying distinctive patterns and relationships [15]. Meanwhile, LSA effectively distills the underlying semantic structure of documents through dimensionality reduction [14].

Although more advanced models such as BERT and GPT have demonstrated impressive capabilities in text analysis [28, 29], their computational demands and requirement for substantial labeled data for fine-tuning were beyond the scope of our current dataset. Moreover, LDA, NMF, and LSA offer the added value of interpretable results that can be easily understood by decision-makers, a critical consideration in our application [30]. Future research may explore the integration of BERT or GPT to further enhance topic modeling capabilities.

4. Results

This section presents the results obtained by training LDA, LSA, and NMF on the monthly collected textual data, with a focus on the evolution of identified themes over time and their implications for health decision-makers. We determined the optimal hyperparameters for each model through a grid search for LDA and a random search for both LSA and NMF. LDA performed best with an online learning method and a learning decay of 0.9, among other parameters. LSA's best configuration involved a randomized algorithm with a tolerance of 1e-06, and NMF showed optimal

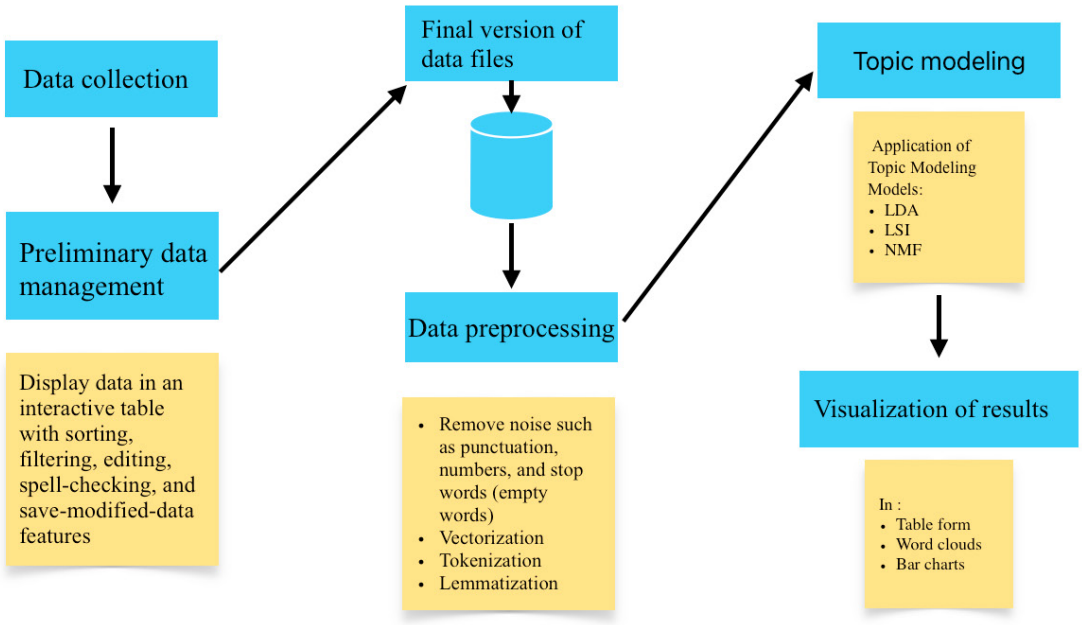


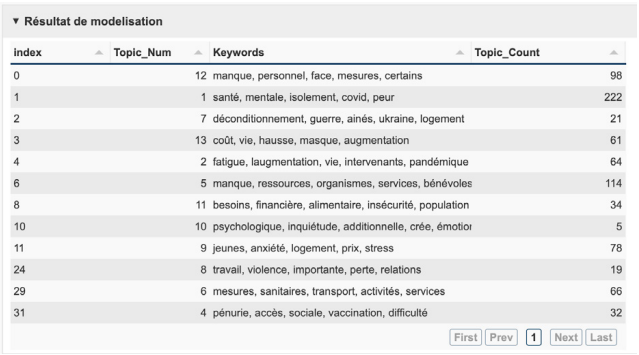
Fig. 2: Implementation Process of Topic Models

Table 1: Comparison of performance metrics for the best configurations across different models.

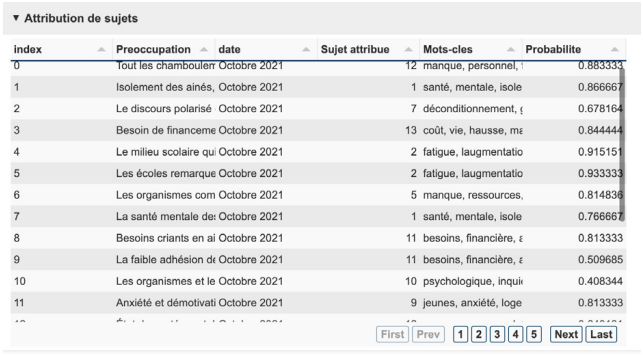
Model	N-gram Type	Number of Topics	Coherence Score	Perplexity Score (LDA)
LDA	Unigram	15	0.59	1218.7
LDA	Unigram	20	0.59	1319.1
LDA	Bigrams	20	0.54	728.4
LDA	Bigrams	25	0.53	754.8
LDA	Trigrams	15	0.44	160.0
LDA	Trigrams	20	0.44	174.5
LSA	Unigram	25	0.47	-
LSA	Unigram	30	0.47	-
LSA	Bigrams	25	0.48	-
LSA	Bigrams	35	0.47	-
LSA	Trigrams	15	0.46	-
NMF	Unigram	30	0.35	-
NMF	Bigrams	10	0.36	-
NMF	Bigrams	25	0.35	-
NMF	Trigrams	15	0.36	-

performance with a coordinate descent solver and an L1-ratio of 0.5. Table 1 compares the performance of these models using coherence and perplexity metrics across different numbers of topics and n-gram types. LDA provided the most interpretable topics, with coherence scores indicating meaningful groupings of concerns. The identified topics (Figure 3a) include prevalent psychosocial themes such as emotional well-being, stress, social isolation, and housing issues, which are key areas of interest for health authorities. Figure 3b shows the distribution of these topics in the data set, revealing how frequently certain concerns are raised by the population.

The high coherence score for LDA with unigrams (0.59) suggests that single-word topics best capture the core concerns. In contrast, the performance of trigrams was weaker, indicating that more complex n-grams may introduce noise rather than clarity in this context. These findings suggest that simple word combinations provide more meaningful insights, perhaps because the psychosocial concerns expressed by the population tend to be succinct and centered around specific key terms such as ”stress,” ”support,” and ”housing.” However, LDA’s performance with bigrams (coherence score of 0.54) shows that some level of context (e.g., two- word phrases) can enhance the model’s ability to capture nuanced concerns, such as ”financial stress” or ”mental health.” This finding emphasizes the importance of selecting the appropriate n-gram configuration based on the type of data analyzed and the level of detail required for decision-making.



(a) Identified topics as keywords.



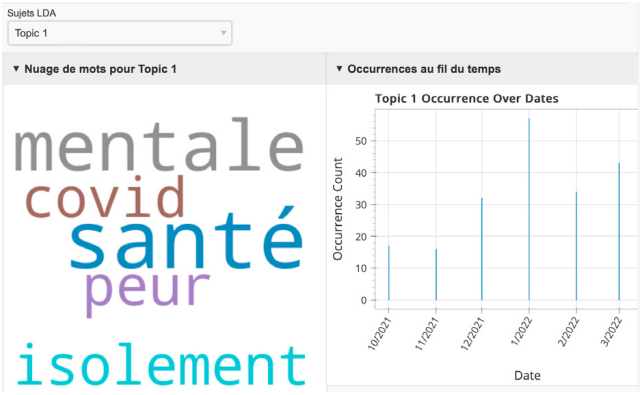
(b) Distribution of topics.

Fig. 3: Identified topics as keywords and distribution of topics.

Figure 4 visualizes the evolution of these topics over time, providing insight into shifting psychosocial concerns. For example, the prominence of topics related to mental health and isolation increased significantly during specific months, likely reflecting the ongoing impact of the COVID-19 pandemic and associated public health measures.

This information is crucial for health stakeholders, as it allows timely interventions and allocation of resources. By tracking how concerns such as housing or financial stress fluctuate, health authorities can adjust service offerings to address emerging needs in real time. The word cloud (Figure 4) offer an intuitive way to understand the most frequent concerns raised by the population. Health decision-makers can use these visualizations to prioritize interventions based on the recurrence of terms like ”stress” or ”support.”

In summary, the results of LDA, LSA, and NMF highlight key psychosocial concerns, and their evolution over time provides actionable insights for stakeholders. The coherence and perplexity scores demonstrate that LDA with unigrams captures the most relevant themes, supporting its application in real-time psychosocial monitoring.



(a) Word cloud and occurrence histogram over time for each topic.

(b) Word cloud across all topics.

Fig. 4: Word cloud and occurrence histogram over time.

5. Discussion

The results show that the three topic modeling models (LDA, LSA, and NMF) successfully identified themes in the textual data. The main objective of the study was achieved, namely determining and tracking topics to monitor psychosocial well-being, with LDA performing the best, particularly with bigrams and a topic number of 20. Our findings align with previous studies, such as [16], which demonstrated the effectiveness of NLP techniques in analyzing social media opinions on food security, revealing varying sentiments over time. Moreover, [19] recently compared the performance of LDA and NMF for the topic modeling of tweets in Urdu. They used an automatic labeling method to evaluate the quality of topics produced by the two algorithms and found that LDA had the best performance in producing coherent and relevant topics, while NMF was more suitable for identifying more varied or diverse topics. In response, we ensured thorough documentation of preprocessing practices and hyperparameters to improve model performance in similar contexts. This study supports topic modeling as a viable solution for analyzing psychosocial well-being, aiding health professionals in decision-making based on identified needs.

Despite the promising results, certain limitations emerged. The coherence and perplexity scores used to evaluate model performance show variability, especially with higher-order n-grams (bigrams, trigrams). For instance, LDA performed well with bigrams, but showed decreased coherence with trigrams. These metrics, while widely used, have known limitations in capturing the full quality of topics. The coherence score, for example, primarily measures the semantic similarity of words within a topic, but may not always reflect the general interpretability of the topics [18].

Compared to prior studies on topic modeling in health data, such as [20] and [13], our work provides a novel approach by focusing on psychosocial well-being in a local population, monitored over time. While existing studies have primarily addressed acute health crises or broader public health trends, our application offers a more nuanced, longitudinal view of psychosocial concerns. This could provide valuable insights for local health organizations looking to adapt their services based on evolving community needs.

However, unlike the BERT-based approaches used in recent studies for topic extraction, we opted for models like LDA, NMF, and LSA because of their balance between interpretability and computational efficiency. BERT's contextual embeddings might offer richer topics, but come at the cost of reduced interpretability and higher resource demand, as noted by [17]. Future studies could explore a hybrid approach, combining traditional topic models with transformer-based models to balance interpretability and topic richness.

6. Conclusion

This study applied topic modeling techniques, including LDA, LSA, and NMF, to monitor psychosocial well-being through text data analysis. LDA, in particular, provided the most interpretable and relevant themes, such as emotional well-being, stress, and social isolation, offering valuable insights for health authorities to adjust their services based on emerging concerns.

Future research should explore more advanced models like BERT and GPT for richer topic extraction, while considering a model-driven architecture (MDA) approach to better align the technical implementation with decision-makers' needs. Despite promising results, further improvements can be made by refining the performance of the model and generalizing the approach to other contexts.

References

[1] Blei, D. M., Ng, A. Y., and Jordan, M. I. (2003). "Latent Dirichlet Allocation." *Journal of Machine Learning Research*, 3, 993–1022.

[2] Griffiths, T. L., and Steyvers, M. (2004). "Finding Scientific Topics." *Proceedings of the National Academy of Sciences*, 101(Suppl 1), 5228–5235. DOI: 10.1073/pnas.0307752101.

[3] Alsumait, L., Barbara, D., and Domeniconi, C. (2008). "On-Line LDA: Adaptive Topic Models for Mining Text Streams with Applications to Topic Detection and Tracking." In *Proceedings of the 8th IEEE International Conference on Data Mining (ICDM 2008)*, Pisa, Italy, December 15–19. DOI: 10.1109/ICDM.2008.140.

[4] Ryff, C. D. (2014). "Psychological Well-Being Revisited: Advances in the Science and Practice of Eudaimonia." *Psychotherapy and Psychosomatics*, 83(1), 10–28. DOI: 10.1159/000353263. PMID: 24281296; PMCID: PMC4241300.

[5] O'Connor, R. C., Wetherall, K., Fazel, S., McClelland, H., Melson, A. J., Niedzwiedz, C. L., O'Carroll, R. E., O'Connor, D. B., Platt, S., Scowcroft, E., Watson, B., Zortea, T., Ferguson, E., and Robb, K. A. (2021). "Mental Health and Well-Being During the COVID-19 Pandemic:

Longitudinal Analyses of Adults in the UK COVID-19 Mental Health & Well-Being Study." *British Journal of Psychiatry*, 218(6), 326–333. DOI: 10.1192/bjp.2020.212. PMID: 33081860; PMCID: PMC7684009.

[6] Miner, G. D., Elder, J., Fast, A., Hill, T., Nisbet, R., and Delen, D. (2012). *Practical Text Mining and Statistical Analysis for Non-Structured Text Data Applications*. Academic Press. ISBN: 9780123870117.

[7] Feldman, R., and Sanger, J. (2007). *The Text Mining Handbook: Advanced Approaches in Analyzing Unstructured Data*. Cambridge University Press. DOI: 10.1017/CBO9780511546914.

[8] Jurafsky, D., and Martin, J. H. (2008). *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition* (2nd ed.). Prentice Hall.

[9] Manning, C. D., Raghavan, P., and Schütze, H. (2008). *Introduction to Information Retrieval*. Cambridge University Press. ISBN: 9780521865715.

[10] Bird, S., Klein, E., and Loper, E. (2009). *Natural Language Processing with Python: Analyzing Text with the Natural Language Toolkit*. O'Reilly Media, Inc. ISBN: 9780596516499.

[11] Mikolov, T., Chen, K., Corrado, G., and Dean, J. (2013). "Efficient Estimation of Word Representations in Vector Space." arXiv preprint arXiv:1301.3781. DOI: 10.48550/arXiv.1301.3781.

[12] Pennington, J., Socher, R., and Manning, C. D. (2014). "Glove: Global Vectors for Word Representation." In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. DOI: 10.3115/v1/D14-1162.

[13] Jelodar, H., Wang, Y., Orji, R., and Huang, S. (2020). "Deep Sentiment Classification and Topic Discovery on Novel Coronavirus or COVID-19 Online Discussions: NLP Using LSTM Recurrent Neural Network Approach." *IEEE Journal of Biomedical and Health Informatics*, 24(10), 2733–2742. DOI: 10.1109/JBHI.2020.3001216.

[14] Landauer, T. K., Foltz, P., and Laham, D. (1998). "An Introduction to Latent Semantic Analysis." *Discourse Processes*, 25, 259–284. DOI: 10.1080/01638539809545028.

[15] Lee, D. D., and Seung, H. S. (1999). "Learning the Parts of Objects by Non-Negative Matrix Factorization." *Nature*, 401(6755), 788–791. DOI: 10.1038/44565.

[16] Molenaar, A., Lukose, D., Brennan, L., Jenkins, E. L., and McCaffrey, T. A. (2024). "Using Natural Language Processing to Explore Social Media Opinions on Food Security: Sentiment Analysis and Topic Modeling Study." *Journal of Medical Internet Research*, 26, e47826. DOI: 10.2196/47826. PMID: 38512326.

[17] Zhang, T., Schoene, A. M., Ji, S., et al. (2022). "Natural Language Processing Applied to Mental Illness Detection: A Narrative Review." *npj Digital Medicine*, 5, 46. DOI: 10.1038/s41746-022-00589-7.

[18] Rüdiger, M., Antons, D., Joshi, A. M., and Salge, T. O. (2022). "Topic Modeling Revisited: New Evidence on Algorithm Performance and Quality Metrics." *PLoS One*, 17(4), e0266325. DOI: 10.1371/journal.pone.0266325. PMID: 35482786.

[19] Zoya, S. L., Shafait, F., and Latif, R. (2021). "Analyzing LDA and NMF Topic Models for Urdu Tweets via Automatic Labeling." *IEEE Access*, 9, 3112620. DOI: 10.1109/ACCESS.2021.3112620.

[20] Paul, M. J., and Dredze, M. (2014). "Discovering Health Topics in Social Media Using Topic Models." *PLoS ONE*, 9(8), e103408. DOI: 10.1371/journal.pone.0103408.

[21] Huang, L., Dou, Z., Hu, Y., and Huang, R. (2019). "Textual Analysis for Online Reviews: A Polymerization Topic Sentiment Model." *IEEE Access*, 7, 91940–91945. DOI: 10.1109/ACCESS.2019.2920091.

[22] Purpura, A. (2018). "Non-Negative Matrix Factorization for Topic Modeling." In *Proceedings of the Biennial Conference on Design of Experimental Search & Information Retrieval Systems*. URL: https://api.semanticscholar.org/CorpusID:52076520.

[23] Wiemer-Hastings, P., Wiemer-Hastings, K., and Graesser, A. (2004). "Latent Semantic Analysis." In *Proceedings of the 16th International Joint Conference on Artificial Intelligence*, 1–14.

[24] Saha, K., et al. (2020). "Social Media as a Passive Sensor in Psychological Health: A Topic Modeling Approach." *Journal of Medical Internet Research*, 22(6), e20389. DOI: 10.2196/20389.

[25] Resnik, P., et al. (2015). "Beyond LDA: Exploring Supervised Topic Modeling for Depression-Related Language in Twitter." In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics*, 99–109. DOI: 10.3115/v1/N15-1010.

[26] Guntuku, S. C., et al. (2019). "Detecting Stress and Mental Health Issues Through Social Media: A Critical Review." *Current Opinion in Psychology*, 36, 89–94. DOI: 10.1016/j.copsyc.2019.06.009.

[27] Yang, A. C., et al. (2021). "Social Media Topic Modeling and Online Mental Health Discussions During the COVID-19 Pandemic." *Journal of Affective Disorders*, 295, 268–275. DOI: 10.1016/j.jad.2021.08.035.

[28] Devlin, J., Chang, M. W., Lee, K., and Toutanova, K. (2019). "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding." In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics*, 4171–4186.

[29] Radford, A., et al. (2019). "Language Models are Unsupervised Multitask Learners." *OpenAI Blog*.

[30] Chaney, A. J. B., and Blei, D. M. (2012). "Visualizing Topic Models." In *Proceedings of the International AAAI Conference on Web and Social Media*, 6(1), 419–422.

[31] Lazarus, R. S., and Folkman, S. (1991). "Stress and Coping: A Theory and Research Overview." In *Handbook of Stress: Theoretical and Clinical Aspects*, edited by A. Monat and R. S. Lazarus, 3–30. John Wiley & Sons, Inc.

[32] Evans, G. W., and Kantrowitz, E. (2003). "Socioeconomic Status and Health: The Potential Role of Environmental Risk Exposure." *Annual Review of Public Health*, 24, 303–331. DOI: 10.1146/annurev.publhealth.24.100901.140923.

CHAPITRE 4
CLASSIFICATION MULTI-ETIQUETTE DES PREOCCUPATIONS PSYCHOSOCIALES EVOLUTIVES A L'AIDE DE MODELES DE LANGAGE DE GRANDE TAILLE BASES SUR LE PROMPTING.

4.1 RESUME EN FRANÇAIS DU QUATRIEME ARTICLE

Cet article, intitulé « Multi-label classification of evolving psychosocial concerns using prompt-based large language models », a été accepté pour publication dans sa version finale en 2025 par les éditeurs de la revue Procedia Computer Science (Elsevier). Il a été présenté lors de la 6e Journée scientifique du Centre de recherche du CISSS de Chaudière-Appalaches qui s'est tenue le 10 avril 2025, ainsi que lors du 61e Colloque ASRDLF tenu le 27 juin 2025.

En tant que première auteure, j'ai conduit la conceptualisation de l'approche, rédigé la majorité du manuscrit et coordonné l'implémentation du pipeline de classification, incluant l'utilisation de Large Language Models (LLaMA 3.1, Qwen 2.5, Mistral, etc.) et la stratégie de prompt engineering (zero-shot, few-shot, chain-of-thought). Le professeur Mehdi Adda a assuré la direction scientifique du projet, validé la méthodologie et supervisé l'évaluation des modèles, tandis que la professeure Lily Lessard a contribué à adapter la solution aux enjeux de santé psychosociale, en plus de participer à la relecture finale de l'article.

Cet article s'inscrit dans la continuité du projet Vigie Psychosociale, dont l'objectif est de mettre en place des outils d'analyse et de détection automatisée des préoccupations psychosociales au Québec. Il propose un pipeline novateur, fondé sur l'hébergement local des LLMs et l'intégration dynamique de nouvelles catégories, afin de répondre à l'évolution constante des situations psychosociales surtout dans des contextes de bouleversements (pandémie, castatrophes, instabilités politiques, etc.). Les résultats mettent particulièrement

en évidence une précision dépassant 95 % selon les tests de concordance effectuée par les intervenants en santé publique, ainsi qu'une capacité à générer des explications contextualisées. L'approche retenue ouvre ainsi la voie à une intelligence artificielle plus fiable, explicable et adaptée à la complexité du domaine psychosocial et prévoit un déploiement au niveau du serveur de l'UQAR pour plus de teste et d'accessibilité.

1.1 MULTI-LABEL CLASSIFICATION OF EVOLVING PSYCHOSOCIAL CONCERNS USING PROMPT-BASED LARGE LANGUAGE MODELS

Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 00 (2018) 000–000

Procedia Computer Science

www.elsevier.com/locate/procedia

The 15th International Conference on Current and Future Trends of Information and Communication Technologies in Healthcare (ICTH 2025)
October 28-30, 2025, Istanbul, Türkiye

Multi-label classification of evolving psychosocial concerns using prompt-based large language models

Adnane Fatima-Azzahrae[a,*], Adda Mehdi[a], Lessard Lily[b], Turcotte Simon[c]

[a]*Département de Mathématiques, Informatique et Génie, Université du Québec à Rimouski (UQAR), Canada (QC)*
[b]*Département des sciences de la santé, Université du Québec à Rimouski (UQAR), Canada (QC)*
[c]*Études urbaines À l'INRS, Chaire CIRUSSS, UQAR, Canada (QC)*

Abstract

Classifying short, evolving textual descriptions of psychosocial concerns presents major challenges for traditional machine learning (ML) and deep learning (DL) models, which often struggle to adapt to emerging or overlapping categories. Multi-label classification further complicates the task, requiring flexible mechanisms capable of assigning multiple relevant labels to a single input. This study introduces a local, prompt-based classification approach that reframes the task as text generation, enabling dynamic and hierarchical labeling without retraining. The system leverages open-source large language models (LLaMA 3.1, LLaMA 3.3, Mistral, and Qwen) to associate each concern with one or more categories while generating contextual explanations to support professional interpretation. LLaMA 3.1 achieved the highest accuracy (97.2%) at the subcategory level, outperforming both classical baselines and other LLMs.

Keywords: Artificial intelligence, NLP, Psychosocial concerns, Classification, Large Language Models, Ollama, Deep learning.

1. Introduction

The public health directorate of Chaudière-Appalaches, in collaboration with Université du Québec à Rimouski (UQAR), collects qualitative data to monitor psychosocial well-being through the Vigie-Psychosociale initiative. Unlike structured indicators, free-text concerns are short, diverse, and context-dependent, requiring scalable systems for hierarchical multi-label classification and the detection of emerging categories without retraining.
Prior research [2], using traditional ML and DL methods showed acceptable performance under fixed-label settings,

* Corresponding author: Adnane Fatima-Azzahrae
E-mail address: fatimaazzahrae.adnane@uqar.ca

but failed to adapt to evolving categories. These limitations underscored the need for a more adaptive and explainable approach to analyzing qualitative data at scale. Building on these observations, we propose to reformulate the classification task using local open-source large language models (LLMs). Rather than assigning fixed labels, our method relies on contextual prompting to dynamically associate each concern with one or more thematic categories. This prompt-based strategy enables the system to operate without retraining, while remaining aligned with privacy constraints, as models are deployed locally. Recent advances in natural language processing (NLP) have shown that LLMs can effectively support classification and summarization tasks, even with limited training data, by leveraging prompt-based or few-shot strategies [7]. In mental health and clinical contexts, LLMs have been applied for screening, risk assessment, and thematic synthesis [19]. Although we could not directly benchmark against proprietary SOTA LLMs such as GPT-4 or Claude due to confidentiality and data protection constraints, published evaluations on comparable tasks indicate that our local approach based on LLaMA 3.1 and Mistral provides a solid compromise between performance and compliance with ethical and legal requirements for sensitive data [24]. Building on our earlier work with classical ML and DL techniques, we compare five locally deployed open-source LLMs : LLaMA 3.1, LLaMA 3.2,LLaMA 3.3 Qwen 2.5, and Mistral 7B, to balance classification accuracy with privacy requirements. Beyond mere categorization, the system delivers structured explanations that enable public health teams to concentrate on higher-level insights and targeted interventions. In initial tests using anonymized data from Vigie-Psychosociale, our best-performing model surpassed existing classical ML baselines by improving the classification accuracy and the capacity to detect emerging psychosocial concerns. Moreover, prompt-based explanations offered clearer, context-rich justifications for each classification, facilitating quicker validation by domain experts and uncovering subthemes that were previously overlooked.

The paper is organized as follows: Section 2 reviews relevant literature on ML, DL, and LLMs, especially in medical contexts, highlighting limitations with sensitive data. Section 3 details the proposed methodology, including model deployment and few-shot prompting. Section 4 covers implementation and system integration. Section 5 presents results, discussion, and model analysis. Section 6 concludes with key contributions and future research directions.

2. Literature review

Traditional classification assigns a single label per instance [13], but in the context of our project, psychosocial monitoring requires multi-label classification due to overlapping themes in short texts. Respondents often describe complex situations needing multiple category labels linked to health or social services [2, 25]. Challenges include imbalanced data, brief texts, and evolving labels as new concerns emerge. Classical methods retrain models periodically, while recent approaches use dynamic strategies to detect and add new categories with minimal retraining. This sets the stage for the following literature review on multi-label methods for short, evolving, and imbalanced text data. Multi-label text classification (MLTC) assigns multiple categories per document, posing challenges beyond single-label tasks [13]. Existing methods fall into two groups: problem-transformation (Binary Relevance, Label Powerset, Classifier Chains) [16, 21]and algorithm-adaptation (ML-kNN, SVM-based) [4, 3]. Data imbalance causes few labels to dominate while others remain rare, affecting label frequencies and co-occurrences [2]. Ensemble variants like RAkEL and ECC redistribute label influence, while MLSMOTE and related methods generate synthetic examples [8].

Most traditional methods assume a fixed label space. MuENL [18] introduced a framework to detect emerging labels in streams, but its effectiveness depends on early label occurrences and it lacks semantic integration with existing categories. Recent advances [26] address these limitations by using counterfactual analysis and semantic alignment to incorporate new categories dynamically while reducing retraining costs. In parallel, ontology-based classification has been explored to address semantic coherence in evolving label spaces. Ontologies such as SNOMED CT or UMLS have been used to enforce hierarchical consistency, improve interpretability in biomedical NLP tasks [27] and confirm that ontology-guided machine learning can enhance robustness and alignment with domain knowledge. However, in the context of transversal data, enforcing a rigid ontology could artificially constrain interpretation and introduce bias, ultimately reducing the contextual richness that LLMs are able to capture. A further related stream of work involves automated triage systems, which employ ML and NLP to prioritize or route patient complaints in emergency and digital health contexts. Reviews of these systems point to notable accuracy gains while at the same time drawing attention to persistent difficulties, including handling brief and imprecise inputs, dealing with categories that are not always stable, and ensuring that automated outputs remain under professional control [28].

LLMs are based on Transformer architectures [7][6] with self-attention to capture long-range dependencies and generate coherent text. Pretrained on large datasets, they perform diverse tasks without costly fine-tuning [12], enabling contextual redirection via instructions and avoiding exhaustive retraining [15]. Prompt-tuning has shown its effectiveness [14], adapting LLMs to medical tasks like clinical extraction and summarization with less annotated data [15]. Few-shot and zero-shot prompting leverage embedded knowledge for complex problems, useful in clinical Q&A and report summarization [9].

However, risks include factual errors, biases, and robustness challenges in sensitive, regulated contexts, as seen in GPT-4 clinical evaluations [10]. Lack of systematic validation across cohorts and standardized resilience protocols [11] hinder safe adoption. Prompt sensitivity and bias remain issues; frameworks like DSPy [17] offer control over information flow. Yet, robust testing and reproducibility in clinical and psychosocial domains face challenges due to variable data and absent standard evaluation [14]. Regulated environments like healthcare increasingly prefer on-premise language models to avoid external API dependencies and comply with data-protection rules. Compact architectures — for example,Mistral [20], Qwen, and LLaMA [22]— can run on institutional hardware, preserving data locality during inference. Fine-tuning these models on internal data is feasible: LLaMA derivatives deployed in hospitals achieve task performance comparable to cloud models while keeping protected health information [22]. Evaluations of Mixtral, Qwen, and BioMistral on 15 extraction categories show local models can match or exceed remote APIs when calibrated to domain terms [24].

3. Methodology

The system follows a modular pipeline composed of four phases: data preparation, classification, label verification, and explanation generation (Fig. 1). This structure ensures flexibility, interpretability, and extensibility across psychosocial domains.

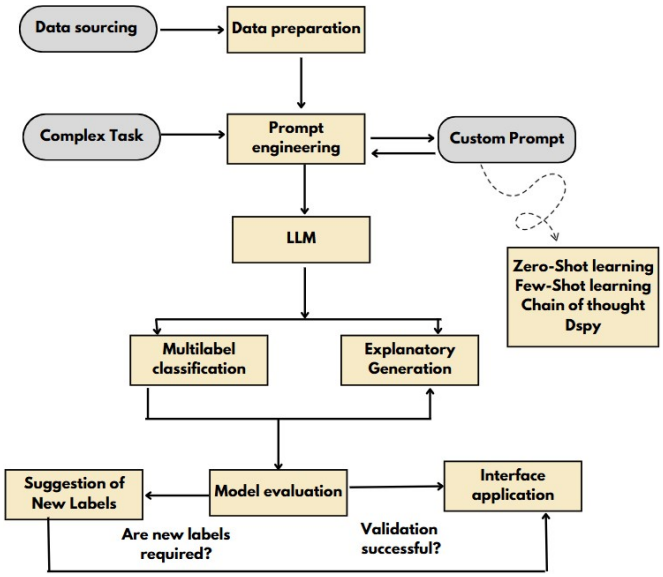


Fig. 1. Modular architecture approach for multilabel classification and generative explanation using LLMs

3.1. Data sourcing and preparation

Data originate from the Vigie-Psychosociale initiative, collected via LimeSurvey and exported in structured Excel format, with entries averaging 5–6 words. The dataset features a hierarchical structure comprising 8 Level-1 Classes and 41 Level-2 Classes, with entries averaging 5 to 6 words in length.

While LLMs have demonstrated the ability to perform certain preprocessing tasks, such as error detection and data imputation [5], some preprocessing steps remain beneficial to enhance their performance and ensure data consistency.

Structuring and harmonizing textual data can help reduce ambiguities that might otherwise affect classification accuracy and interpretability. The key preprocessing steps include: null value removal, elimination of duplicates, and filtering of uninterpretable entries to reduce semantic ambiguity and improve consistency.

3.2. Prompt engineering and optimization

As outlined in Section 2, prompt engineering constitutes the central element of our approach. We steer the LLM with contextual instructions that embed the current label inventory and, when needed, a handful of illustrative examples. This prompt is rebuilt at run-time from the database, guaranteeing that any label addition, removal, or hierarchy change is reflected without retraining[14]. To leverage the intrinsic capabilities of LLMs in this study, we employ three implicit learning strategies. Zero-shot learning is applied to generic categories that are well represented in the LLM's pre-training corpus [1]. When the psychosocial concern uses informal or highly specialised vocabulary, we switch to few-shot learning, embedding a small, diversity-oriented set of illustrative examples to improve generalisation [7]. For the most challenging cases, we invoke chain-of-thought (CoT) prompting, which guides the model through an explicit reasoning sequence: first extracting salient keywords, then mapping them to candidate labels, and finally providing a justification for each choice.

3.2.1. Prompt optimization with DSPy

DSPy integrates three components for prompt optimization [17]. The Signature maps ⟨context, question⟩ to a response, where context is the textual concern and the output is a set of labels with brief explanations. Few-shot examples embedded in the Signature capture domain-specific variability and guide interpretation. The Module applies chain-of-thought reasoning via a locally hosted LLM (Ollama) in three steps: keyword extraction, label matching, and concise justification, improving accuracy and interpretability. The Optimizer varies prompt phrasing, directive detail, example order, and explanation depth, selecting lightweight model checkpoints when needed. Since labels are dynamically retrieved at inference, taxonomy changes trigger Signature regeneration and prompt re-optimization, ensuring alignment without retraining.

3.3. Dynamic multi-label classification

The classification task is framed as a text-to-sequence generation problem in which a dynamic prompt is constructed at runtime based on the current inventory of labels stored in a relational database. This structure enables healthcare professionals to add, remove, or reorganize categories through the user interface without requiring model retraining or backend modifications.

At each execution, the system queries the database to retrieve the complete and up-to-date list of labels (including hierarchy if applicable), then integrates it into the prompt. Any change in the taxonomy immediately affects the model's behavior, ensuring full alignment between classification outputs and the evolving set of categories. Because psychosocial concerns often contain multiple co-occurring issues, the model is designed to support multi-label predictions (Fig. 2). It assigns at least one label per concern, with no upper bound imposed. To maintain semantic precision, the prompt explicitly instructs the model to rely only on explicitly stated content, avoiding speculative inference or latent-topic association. The integration of Few-Shot Learning and Chain-of-Thought (CoT) strategies reinforces the model's ability to handle multi-label outputs.

3.4. Explanation generation: A decision support tool

In the context of classifying psychosocial concerns, it is not sufficient to merely assign relevant labels. It is equally essential to provide clear and comprehensible explanations that allow healthcare professionals to interpret the results effectively. Explanation generation thus supports decision-making by offering a structured overview of the possible causes underlying the identified concerns and uncovering interconnections between them and Leveraging the reasoning capabilities of LLMs (Fig. 3).

To refine the explanation, the system sends a secondary query to the model, retrieving key terms extracted from all concerns previously associated with a given label. By leveraging this enriched information, the model generates structured and contextually relevant responses that incorporate causal hypotheses, explicit category correlations, and

tailored recommendations.

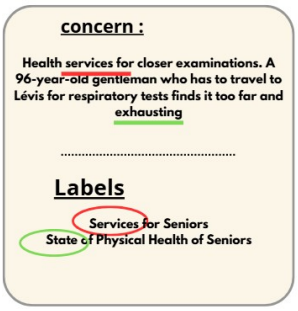


Fig. 2. Multilabel association based on a textual concern in NLP

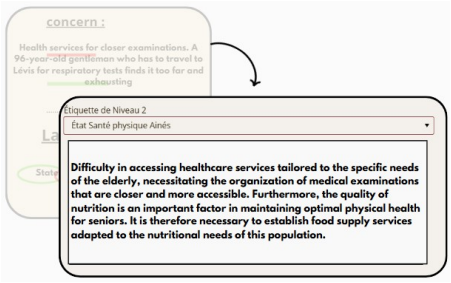


Fig. 3. Causal analysis of psychosocial concerns: Contextual generative explanation

3.5. Prompting for dynamic generation of emerging themes

Beyond selecting existing labels, in this study, it is often necessary to propose new categories that have not been previously defined. To address this need, a dynamic label generation mechanism has been implemented, leveraging the flexibility of prompting [14] and the ability of LLMs to synthesize generic concepts from specific textual inputs.

This process relies on a dedicated prompt designed to encourage the model to generate a label that is more general and abstract than the initial concern. The goal is to avoid reusing exact terms from the source text but rather to distill a generic, concise concept (maximum three words) that can be applied in similar contexts. This mechanism enables the inventory of labels to be progressively enriched, enhancing the system's adaptability to new forms of concerns or emerging contextual developments. The mechanism relies on DSPy, which handles both the generation of labels and their automatic validation, ensuring compliance with predefined criteria (e.g., avoidance of initial terms, appropriate length). In cases of non-compliance, a new iteration is triggered to refine the output. This process enables the creation of an evolving inventory aligned with the changing needs of the psychosocial domain. Dynamic label generation is integrated into the pipeline as a complement to the existing stages. Once a new label is generated and validated by healthcare professionals, it can be added to the relational database. Thus, during future iterations, this label becomes part of the inventory consulted by the dynamic prompt, contributing to the system's continuous adaptation.

3.6. Evaluation strategy for the LLM-based classification and explanation

In this study, evaluation was not treated as a final or isolated step but was integrated into the methodological workflow from the outset. It informed key decisions regarding prompt formulation, classification logic, and explanation generation. Supported by continuous expert feedback, this iterative process ensured that system outputs remained contextually and clinically relevant throughout development.

In the absence of a pre-annotated ground truth and given the nuanced nature of psychosocial concerns, the evaluation extended beyond standard metrics such as precision or recall. It aimed to assess the thematic accuracy, contextual coherence, and interpretability of outputs, particularly in complex, multi-label cases involving overlapping or emerging categories. Expert feedback played a central role in this process, helping to identify inconsistencies, refine reasoning patterns, and optimize prompting strategies. Multiple configurations were tested under controlled conditions to isolate methodological effects and avoid attribution of differences to stochastic model behavior. Special attention was given to the evaluation of generated explanations. Rather than assessing stylistic aspects, experts focused on clarity, logical consistency, and alignment with plausible clinical reasoning. This emphasis on interpretability underscored the system's intended role as a decision-support tool, enhancing its applicability in real-world healthcare contexts.

4. Implementation Details

The approach uses locally deployed LLMs (LLaMA 3.1, Qwen 2.5, Mistral 7B) [23], ensuring data control and security by avoiding external services. This setup offers stable, low-latency operation for fast prompt testing and supports custom metrics like semantic coherence and thematic relevance beyond standard cloud APIs [30]. Since no annotated benchmark exists for the psychosocial relatd data we used, our public health team developed a triage procedure to assess the professional acceptability of label sets. In practice, the Algorithm 1 automatically accepts outputs that fall within an acceptable range of labels and coherence checks. Predictions are flagged if they contain too many labels, if more than two labels appear incoherent with the text, or if the model indicates that no suitable label exists. In such cases, the output is routed to a professional for validation or for the creation of a new category.

Algorithm 1 Professional-in-the-loop triage

Input : Respondent free text t
Output: Decision $\in$ {ACCEPT, REJECT}; optional new label

```
labels ← LLM(prompt, t) if |labels| > 6 then
    return REJECT
incorrect ← ProfessionalCountIncorrect(labels) if incorrect > 2 then
    return REJECT
if labels == {"No suitable label"} then
    L_new ← ProfessionalValidateNew() if L_new accepted then
        add L_new to DB
    return ACCEPT
return ACCEPT
```

The system is designed with two connected components for efficient, scalable deployment. Open source models run on a high-performance GPU server at the University of Quebec at Rimouski, equipped with NVIDIA A100 80GB GPUs. The main application interface, managing user interaction, file handling, prompt dispatch, and visualization, runs on a separate, low-resource virtual machine. The LLM environment is containerized using Singularity, and communicates securely via an SSH tunnel, protecting sensitive ports from public exposure.

5. Results and discussion

5.1. Expert-based evaluation of classification quality

We evaluated the approach using a stratified sample of 2,000 textual concerns across 8 Level-1 and 41 Level-2 categories. Each instance was processed in two stages: first, the classification pipeline powered by LLMs. Then manually reviewed and corrected by health professionals. Feedback was iteratively integrated to align the system with expert expectations. In the absence of a gold-standard corpus, we introduce four expert-centric metrics that reflect clinical validity and operational utility:

- PAR – Professional Acceptability Rate: proportion of model outputs fully endorsed by professionals.
- OGF – Over-Generation Frequency: share of instances where the model produced more than six labels, signalling semantic drift and cognitive overload.
- EER – Excess-Error Rate: proportion of cases with more than two incorrect labels inside an otherwise correct set, a direct proxy for precision loss.
- ELA – Emergent-Label Acceptance: fraction of newly suggested labels that professionals judged valid.

The model that achieved the most consistent performance in classifying granular Level-2 subcategories was LLama3.1:405b as shown in Panel B of Table 1, particularly excelling in nuanced thematic areas. In this same context, we assessed the impact of different prompting configurations (Zero-Shot, Few-Shot) on classification quality.

Table 1. Expert-in-the-loop metrics (Panel A), classification accuracy (Panel B), and latency/emergence metrics (Panel C)

Panel A – Expert-in-the-loop metrics

Metric	LLaMA 3.1	LLaMA 3.3	Mistral 7B	Qwen 2.5
PAR	0.997	0.974	0.974	0.935
OGF	0.001	0.004	0.004	0.030
EER	0.024	0.096	0.094	0.214
ELA	0.972	0.908	0.910	0.791

Panel B – Classification accuracy

Model	Level 1	Level 2
LLaMA 3.1	98%	97.2%
LLaMA 3.3	93.5%	90.5%
Mistral 7B	93.5%	90.5%
Qwen 2.5	90%	79.45%

Panel C – Latency and emergence metrics

Model	Classification. (s)	Verification/Emergence. (s)	Emergence. (%)
LLaMA 3.1	2.8 ± 0.5	4.2 ± 0.3	1.8
LLaMA 3.3	1.1 ± 0.2	2.9 ± 0.2	4.3
Mistral 7B	1.8 ± 0.4	2.2 ± 0.3	3.6
Qwen 2.5	2.1 ± 0.6	3.6 ± 0.4	6.1

At Level-1, both configurations produced comparable results in broad category assignments. However, at Level-2, Few-shot prompting with seven examples improved Level-2 accuracy from 85% to 97%, highlighting the substantial impact of providing relevant examples. Chain-of-Thought prompting improved label consistency, especially in ambiguous cases involving overlapping categories like financial distress as an independent concern versus its role as a contributing factor to family stress. As detailed in Panel A of Table 1, LLaMA 3.1 and Mistral exhibited selective multi-label outputs, whereas Qwen 2.5 and LLaMA 3.3 tended to over-assign categories, occasionally affecting interpretability. Contrary to what might be perceived as a limitation, the approach's ability to assign multiple labels to a single concern is, in fact, an asset. It highlights the interdependencies between different psychosocial dimensions, underscoring the intrinsic complexity of certain situations. Experts also found the generated explanations clear and useful, especially with CoT prompting. However, some complex situations (e.g., multifactorial family situations or trajectories of cumulative stress) revealed limits in semantic depth, prompting calls for more enriched justifications. Moreover, the dynamic label generation mechanism emerges as a major asset and was also well received, although professional validation remained essential to maintain clinical relevance and avoid semantic drift. Finally, the Panel C in Table 1 confirm the practical scalability of the system. To operationalize this balance, the professional decision algorithm applies simple triage rules: predictions with excessive labels, incoherent assignments, or novel categories are flagged for expert review, while the majority of outputs are accepted directly. In practice, fewer than 3% of cases required professional intervention for LLaMA 3.1:405b, as reflected in high PAR and ELA scores (97%). This selective escalation ensures scalability by limiting expert workload to exceptional cases while maintaining accuracy.

6. Conclusion and future perspectives

This study presented a prompt-based classification approach reframing multi-label psychosocial classification as text generation. Using locally hosted LLMs, it dynamically selects existing labels or generates new ones without retraining. Tested on expert-annotated data, this approach showed strong accuracy, adaptability, and interpretability. Combining zero-shot, few-shot, and chain-of-thought prompting, it handles short, evolving, and overlapping concerns effectively. Dynamic label generation and expert validation enhance its public health relevance. Limitations include lack of continuous learning, limited multilingual support, and some generative noise, which must be addressed for broader application. Future work will expand to diverse contexts, develop real-time feedback, and benchmark new LLM models.

References

[1] W. Yin et H. Schütze, "Zero-shot Text Classification: A Survey", ACM Computing Surveys,2020.

[2] Fatima-Azzahrae, A., Amraoui, R., Adda, M., Lessard, L. Automatic classification of psychosocial. Procedia Computer Science (EUSPN 2024), Leuven, Belgique, sept. 2024, vol. 251, pp. 390–397. Elsevier, https://doi.org/10.1016/j.procs.2024.11.125

[3] Zhang, M.-L., & Zhou, Z.-H. (2007). ML-KNN: A lazy learning approach to multi-label learning. Pattern Recognition, 40(7), 2038–2048. https://doi.org/10.1016/j.patcog.2006.12.019

[4] Madjarov, G., Kocev, D., Gjorgjevikj, D., & Džeroski, S. (2012). An extensive experimental comparison of methods for multi-label learning. Pattern Recognition, 45(9), 3084–3104.

[5] Zhang, H., Dong, Y., Xiao, C., & Oyamada, M. (2024). Large Language Models as Data Preprocessors. arXiv:2308.16361

[6] T. Wolf, L. Debut, V. Sanh, J. Chaumond, et al., "Transformers: State-of-the-Art Natural Language Processing", in Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP): System Demonstrations, 2020, pp. 38–45.

[7] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, S. Agarwal, A. Herbert-Voss, G. Krueger, T. Henighan, R. Child, A. Ramesh, D. Ziegler, J. Wu, C. Winter, and D. Amodei, "Language Models are Few-Shot Learners," arXiv:2005.14165, 2020. doi: 10.48550/arXiv.2005.14165.

[8] Charte, F., Rivera, A. J., del Jesus, M. J., & Herrera, F. (2015). MLSMOTE: Approaching imbalanced multilabel learning through synthetic instance generation. Knowledge-Based Systems, 89, 385–397. https://doi.org/10.1016/j.knosys.2015.07.019

[9] K. Singhal, P. X. Liu, et al., "Large Language Models Encode Clinical Knowledge", Nature, vol. 610, pp. 60–67, 2022.

[10] H. Nori, N. King, S. M. McKinney, D. Carignan, and E. Horvitz, "Capabilities of GPT-4 on Medical Challenge Problems," arXiv:2303.13375.

[11] Singh, R., Padmanabhan, B. (2023). "Challenges in Benchmarking and Evaluating Large Language Models for Medical AI." IEEE Access, 11:59875–59890.

[12] Hossain, E., Rana, R., Higgins, N., Soar, J., Barua, P. D., Pisani, A. R., Turner, K. (2023). "Natural Language Processing in Electronic Health Records in Relation to Healthcare Decision-making: A Systematic Review." Journal of Biomedical Informatics, 136, 104234. DOI: 10.1016/j.jbi.2023.104234

[13] Boutell, M. R., Luo, J., Shen, X., & Brown, C. M. (2004). Learning Multi-Label Scene Classification. Pattern Recognition, 37(9), 1757–1771. https://doi.org/10.1016/j.patcog.2004.03.009

[14] P. Liu, W. Yuan, J. Fu, Z. Jiang, et al., "Pre-train, Prompt, and Predict: A Systematic Survey of Prompting Methods in NLP", ACM Computing Surveys, 2023.

[15] Peng, C., Yang, X., Smith, K. E., Yu, Z., Chen, A., Bian, J., Wu, Y. (2023). "Model Tuning or Prompt Tuning? A Study of Large Language Models for Clinical Concept and Relation Extraction." Proceedings of the 2023 Conference on Empirical Methods in

[16] Zhang, M.-L., & Zhou, Z.-H. (2014). A Review on Multi-Label Learning Algorithms. IEEE Transactions on Knowledge and Data Engineering, 26(8), 1819-1837.

[17]]Khattab, O., Singhvi, A., Maheshwari, P., Zhang, Z., Santhanam, K., Vardhamanan, S., Haq, S., Sharma, A., Joshi, T. T., Moazam, H., Miller, H., Zaharia, M., Potts, C. (2023). "DSPy: Compiling Declarative Language Model Calls into Self-Improving Pipelines." Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (EMNLP).

[18] Zhu, Y., Ting, K.-M., & Zhou, Z.-H. (2016). Multi-label learning with emerging new labels. In Proceedings of the 2016 IEEE 16th International Conference on Data Mining (ICDM). IEEE.https://doi.org/10.1109/ICDM.2016.0137

[19] Mazumdar, H., Chakraborty, C., Sathvik, M. S., & Mukhopadhyay, S. (2023). GPTFX: A novel GPT-3 based framework for mental health detection and explanations. IEEE Journal of Biomedical and Health Informatics, PP(99), 1-8. https://doi.org/10.1109/JBHI.2023.3328350

[20] Jiang, A. Q., Sablayrolles, A., Mensch, A., Bamford, C., Chaplot, D. S., de las Casas, D., Bressand, F., Lengyel, G., Lample, G., Saulnier, L., Lavaud, L. R., Lachaux, M.-A., Stock, P., Le Scao, T., Lavril, T., Wang, T., Lacroix, T., & El Sayed, W. (2023). Mistral 7B: A high-quality and efficient open-weight language model. arXiv. https://arxiv.org/abs/2310.06825

[21] Read, J., Pfahringer, B., Holmes, G., & Frank, E. (2009). Classifier Chains for Multi-label Classification. Machine Learning, 85(3), 333-359.

[22] H. Touvron, T. Lavril, G. Izacard, et al., "LLaMA: Open and Efficient Foundation Language Models", in International Conference on Learning Representations (ICLR), 2023.

[23] Rieke, N., Hancox, J., Li, W., et al. (2020). "The future of digital health with federated learning." NPJ Digital Medicine, 3(1), 119. DOI: 10.1038/s41746-020-00323-1

[24] Meddeb, A., Ebert, P., Bressem, K. K., Desser, D., Dell'Orco, A., Bohner, G., Kleine, J. F., Siebert, E., Grauhan, N., Brockmann, M. A., Othman, A., Scheel, M., & Nawabi, J. 2024. Evaluating local open-source large language models for data extraction from unstructured reports on mechanical thrombectomy in patients with ischemic stroke.

[25] Amraoui, R., Adnane, F.-A., Adda, M., & Lessard, L. (2024). NLP and topic modeling with LDA, LSA, and NMF for monitoring psychosocial well-being in monthly surveys. Procedia Computer Science https://doi.org/10.1016/j.procs.2024.11.126

[26] Ren, L., Liu, Y., Ouyang, C., Yu, Y., Zhou, S., He, Y., & Wan, Y. (2025). DyLas: A dynamic label alignment strategy for large-scale multi-label text classification. Information Fusion, 110, 103081. https://doi.org/10.1016/j.inffus.2025.103081

[27] Cai, L., Zhang, C., Xu, D., & Wang, F. (2023). Integrating domain knowledge for biomedical text analysis with deep learning: A comprehensive review. Journal of Biomedical Informatics, 145, 104420. https://doi.org/10.1016/j.jbi.2023.104420.

[28] Da'Costa, A., Teke, J., Origbo, J. E., Osonuga, A., Egbon, E., & Olawade, D. B. (2025). AI-driven triage in emergency departments: A review of benefits, challenges, and future directions. International Journal of Medical Informatics, 195, 105838. https://doi.org/10.1016/j.ijmedinf.2025.105838

[29] Wang, A., Liu, C., Yang, J., Weng, C. Fine-tuning large language models for rare disease concept normalization.

[30] Fu, Y., Xue, L., Huang, Y., Brabete, A.-O., Ustiugov, D., Patel, Y., & Mai, L. (2024). ServerlessLLM: Low-latency serverless inference for large language models. arXiv:2401.14351.

[31] Wang, L., Chang, M., & Feng, J. (2005). Parallel and sequential support vector machines for multi-label classification. Center for Information Sciences, Peking University, Beijing, China.

CONCLUSION GÉNÉRALE

Ce mémoire a exploré, par le biais de quatre articles complémentaires, la classification et l'analyse automatisée de préoccupations collectées par la Vigie psychosociale à travers les questionnaires du Lime Survey. Les méthodes abordées vont des techniques classiques d'apprentissage automatique à la modélisation thématique, pour finalement adopter les modèles de langage de grande taille comme solution finale et rentable. Ensemble, ces travaux mettent en lumière la capacité de l'IA à catégoriser de façon plus précise et flexible des données textuelles en général et de notre projet en particulier, à gérer le déséquilibre des classes et à proposer des explications transparentes pour des utilisateurs non-initiés à l'IA.

Le premier article a montré l'efficacité des algorithmes traditionnels (k-NN, SVM, XGBoost) et des méthodes de deep learning basé sur les Transformers et combinées au différentes approches (fine-tuning, few-shot learning) pour la classification de textes courts, fortement dépendants de stratégies de prétraitement (lemmatisation, TF-IDF, Word Embeddings). Le deuxième article s'est davantage concentré sur la modélisation thématique (LDA, LSA, NMF), démontrant l'apport de ces approches pour extraire, comparer et suivre l'évolution de thèmes psychosociaux. Le troisième article, enfin, s'est penché sur les versions locales des LLMs, pour proposer une classification dynamique, multi-catégorielle et explicable, en reformulant la tâche de classification en un processus de génération textuelle.

Ce travail a permis d'explorer différents modèles, mais il a également mis en lumière plusieurs limites rencontrées en pratique. Les méthodes traditionnelles de classification, bien qu'efficaces sur des données bien structurées, se sont révélées sensibles au déséquilibre des classes, produisant des résultats biaisés lorsque certaines catégories étaient sous-représentées. Elles se sont également montrées peu adaptées à la nature évolutive des étiquettes : l'ajout ou la modification de catégories nécessitait un réentraînement complet du modèle, ce qui ralentissait le processus d'analyse. Par ailleurs, elles ne permettaient pas de

proposer automatiquement de nouvelles étiquettes lorsque des préoccupations inédites apparaissaient dans les données, ce qui limitait leur capacité à suivre les dynamiques émergentes. Enfin, les modèles traditionnels manquaient d'une compréhension contextuelle fine, ce qui les rendait moins performants face aux énoncés très courts ou ambigus, fréquents dans les réponses des participants.

Pour surmonter ces difficultés, le recours aux grands modèles de langage s'est imposé comme une alternative plus flexible. Cependant, ces modèles présentent à leur tour leurs propres contraintes. Bien qu'ils améliorent la classification et permettent la génération d'explications textuelles, celles-ci ne sont pas toujours parfaitement cohérentes ni suffisamment spécifiques, il arrive que les explications demeurent trop générales ou qu'elles n'éclairent pas directement la catégorie assignée. De plus, le recours aux LLMs implique un coût computationnel significatif, particulièrement pour des modèles de grande taille comme LLaMA 3.1 : 405B, ce qui peut limiter leur déploiement à grande échelle. Enfin, la mise en place d'un tel système nécessite un suivi continu pour s'assurer du maintien de la qualité et de la précision, en particulier pour la détection des thématiques émergentes, afin d'éviter toute dérive ou perte de performance dans le temps.

Afin de dépasser ces limites et d'améliorer la robustesse du système, plusieurs pistes de développement sont envisagées. Pour rendre les explications générées par les modèles plus spécifiques et contextualisées, l'intégration d'une approche Retrieval-Augmented Generation (RAG) constitue une solution prometteuse. En connectant les LLMs à une base de connaissances régulièrement mise à jour, il devient possible d'enrichir les réponses par des exemples réels et validés, ce qui réduit les risques de généralisations excessives et d'explications trop vagues. Parallèlement, des pistes d'optimisation computationnelle doivent être explorées pour réduire le coût de l'inférence et améliorer la faisabilité opérationnelle du système. Des techniques telles que la quantification et la distillation de modèles peuvent être mises en œuvre afin de conserver une performance élevée tout en diminuant la consommation de ressources.

Une autre perspective importante consiste à élargir le spectre des modèles testés. L'évaluation de nouveaux LLMs, permettra de comparer leurs performances, leur robustesse et la qualité de leurs explications dans le contexte spécifique des préoccupations psychosociales. Enfin, la mise en place de mécanismes d'auto-évaluation intégrés est envisagée, afin de permettre au système de détecter de manière autonome les prédictions incertaines, les baisses de performance ou les explications trop génériques. Un tel dispositif, combiné à un suivi longitudinal automatisé basé sur des métriques de performance telles que la précision, renforcerait la fiabilité et la stabilité du système tout en réduisant la nécessité d'interventions manuelles fréquentes.

RÉFÉRENCES BIBLIOGRAPHIQUES

Abdurahman, S. A.-M. (2023). Perils and opportunities in using large language models in psychological research. *PNAS Nexus*, 245.

Amraoui, r., Adnane, F., Adda, M., & Lessard, L. (2024). NLP and topic modeling with LDA, LSA, and NMF for monitoring psychosocial well-being in monthly surveys. *Procedia Computer Science*.

Beguš, G. D. (2023). Large Linguistic Models: Investigating LLMs' metalinguistic abilities. *ArXiv*.

Buda, M. M. (2018). A systematic study of the class imbalance problem in convolutional neural networks. *Neural Networks*.

Caron, J. (2019). Santé mentale : concepts, mesures et déterminants. *Santé mentale au Québec, Volume 42, numéro 1*, 125-145.

Duggal, P. S. (2013). Big Data Analysis: Challenges and Solutions. *In Proceedings of the International Conference on Cloud*, 269-270.

Fatima azzahrae , a., Amraoui, R., Adda, m., & Lessard, L. (2024). Automatic classification of psychosocial concerns: From traditional approach to deep learning. *Procedia Computer Science*.

H.Touvron, T. L. (2023). LLaMA: Open and Efficient Foundation Language Models. *International Conference on Learning Representations (ICLR)*.

Huang, G. L. (2024). From Explainable to Interpretable Deep Learning for Natural Language Processing in Healthcare: How Far from Reality? *Computational and Structural Biotechnology Journal*.

Le Glaz, A. H.-D. (2021). Machine Learning and Natural Language Processing in Mental Health: Systematic Review. *Journal of Medical Internet Research*.

Oliveira, A. B. (2013). Psychosocial Impact. *Gellman M.D., Turner J.R. (eds), Encyclopedia of behavioral medicine. Springer, New-York.*, 1583–1584.

Patel V., S. S.-M. (2019). The Lancet Commission on global mental health and sustainable development. *Erratum : The Lancet, 393(10185). https://doi.org/10.1016/S0140-6736(19)30982-6.*, p. e21.

T.Brown, B. M. (2020). Language Models areFew-Shot Learners. *NeurIPS*.

Tunstall, L. R. (2022). SetFit: Efficient Few-Shot Learning Without Prompts. *ArXiv*.

Venkatesh, V. S. (2024). Machine Learning Techniques to Predict Mental Health Diagnoses : A Systematic Literature Review. *Clin Pract Epidemiol Ment Health*.