UNIVERSITÉ DU QUÉBEC À MONTRÉAL

MODÉLISATION DE SUJETS POUR LA DÉTECTION DE RISQUES EN SANTÉ MENTALE

MÉMOIRE

PRÉSENTÉ

COMME EXIGENCE PARTIELLE

DE LA MAÎTRISE EN INFORMATIQUE

PAR

MAXIME D. ARMSTRONG

OCTOBRE 2022

No alarms and no surprises, please.

— Radiohead

REMERCIEMENTS

Mon parcours universitaire n'aurait été le même sans mon entourage, particulièrement depuis mon arrivée à la maîtrise. Merci à mes parents, Jocelyne et Michel, pour leur soutien inconditionnel. Merci à Aurélien, Fred, Jérème, Mario, Pierre-Benoit, Serge, Tim, Tom et Ursule pour leur écoute, leurs distractions et leur présence. Merci à mes coéquipiers chez MobilityData pour leur flexibilité et leurs encouragements. Merci à Diego et Fanny pour leurs conseils, leur esprit d'équipe et leur amitié. Merci à ma directrice de recherche, Marie-Jean Meurs, pour les opportunités et son appui dans chaque projet.

Ce mémoire est dédié à toutes les personnes ayant été affectées, de près ou de loin, par la maladie mentale.

TABLE DES MATIÈRES

LISTE DES TABLEAUX

LISTE DES FIGURES

RÉSUMÉ

Les troubles mentaux ont des effets néfastes à l'échelle planétaire, tant sur les individus que sur la société. Malgré ce problème omniprésent, les personnes affectées peinent à trouver l'aide nécessaire. Les avancées scientifiques dans le domaine de la détection automatique de risques en santé mentale sont prometteuses, particulièrement depuis l'arrivée des médias sociaux, où les utilisateurs se confient régulièrement à leur communauté. Dans ce contexte, ce mémoire s'intéresse à la détection de troubles mentaux à partir de contenus textuels dans les médias sociaux à l'aide de la modélisation de sujets.

La modélisation de sujets est une technique puissante permettant d'expliquer la structure des sujets d'un corpus. Avec cette information, un document est descriptible comme une distribution de sujets, celle-ci pouvant servir comme ensemble d'attributs utiles à diverses tâches de détections de troubles mentaux. Toutefois, cette technique est sensible à la composition du vocabulaire d'un corpus et ignore souvent la sémantique des mots observés, résultant en des performances variables selon les troubles étudiés.

Dans un premier temps, une étude comparative présente les capacités de modélisation d'attributs selon le type de trouble mental observé. Pour y arriver, plusieurs modèles de sujets sont entraînés sur divers corpus pour observer comment le vocabulaire affecte les performances. Ces dernières sont prometteuses, particulièrement pour la détection de l'anorexie.

Par la suite, des expériences sont menées pour mettre en parallèle la modélisation de sujets traditionnels avec une nouvelle approche intégrant la sémantique des mots grâce aux plongements de mots. L'accent est mis sur la détection de la dépression au travers d'une tâche de classification binaire. Les résultats démontrent une amélioration lors de l'usage d'attributs obtenus via la nouvelle approche.

MOTS CLÉS : Modélisation de sujets, plongements de mots, classification binaire, analyse de sentiments, détection du risque, santé mentale, dépistage de la dépression, médias sociaux, apprentissage automatique, traitement automatique du langage naturel.

ARTICLES PRÉSENTÉS DANS CE MÉMOIRE

Topic Models for Assessment of Mental Health Issues

Maxime D. Armstrong, *Diego Maupomé*, *Marie-Jean Meurs*

Proceedings of the 34th Canadian Conference on Artificial Intelligence, Canadian AI 2021

PubPub, Canadian Artificial Intelligence Association (CAIAC), AI2021

Topic Modeling in Embedding Spaces for Depression Assessment

Maxime D. Armstrong, *Diego Maupomé*, *Marie-Jean Meurs*

Proceedings of the 34th Canadian Conference on Artificial Intelligence, Canadian AI 2021

PubPub, Canadian Artificial Intelligence Association (CAIAC), AI2021

AUTRES PUBLICATIONS

Early Mental Health Risk Assessment through Writing Styles, Topics and Neural Models

Diego Maupomé, *Maxime D. Armstrong*, *Raouf Moncef Belbahar*, *Josselin Alezot*, *Rhon Balassiano*, *Marc Queudot*, *Sébastien Mosser*, *Marie-Jean Meurs*

Proceedings of the 11th Conference and Labs of the Evaluation Forum, CLEF 2020

CEUR Workshop Proceedings, CLEF 2020 Working Notes, CLEF 2020

Early Detection of Signs of Pathological Gambling, Self-Harm and Depression through Topic Extraction and Neural Networks

Diego Maupomé, *Maxime D. Armstrong*, *Fanny Rancourt*, *Thomas Soulas*, *Marie-Jean Meurs*

Proceedings of the 12th Conference and Labs of the Evaluation Forum, CLEF 2021

CEUR Workshop Proceedings, CLEF 2021 Working Notes, CLEF 2021

Leveraging Textual Similarity to Predict Beck Depression Inventory Answers

Diego Maupomé, *Maxime D. Armstrong*, *Fanny Rancourt*, *Marie-Jean Meurs*

Proceedings of the 34th Canadian Conference on Artificial Intelligence, Canadian AI 2021

PubPub, Canadian Artificial Intelligence Association (CAIAC), AI2021

Position Encoding Schemes for Linear Aggregation of Word Sequences

Diego Maupomé, *Fanny Rancourt*, *Maxime D. Armstrong*, *Marie-Jean Meurs*

Proceedings of the 34th Canadian Conference on Artificial Intelligence, Canadian AI 2021

PubPub, Canadian Artificial Intelligence Association (CAIAC), AI2021

Automatically Estimating the Severity of Multiple Symptoms Associated with Depression

Diego Maupomé, *Maxime D. Armstrong*, *Raouf Belbahar*, *Josselin Alezot*, *Rhon Balassiano*, *Fanny Rancourt*, *Marc Queudot*, *Sébastien Mosser*, *Marie-Jean Meurs*

Early Detection of Mental Health Disorders by Social Media Monitoring

Studies in Computational Intelligence, volume 1018. Springer, Cham, 2022

INTRODUCTION

En 2017, 792 millions de personnes vivaient avec un trouble de santé mentale, représentant 10,7% de la population mondiale. Selon les données observées, les troubles anxieux et la dépression étaient les plus fréquents, totalisant plus de la moitié des cas (Dattani *et al.*, 2021). Or, les problèmes reliés à la santé mentale ont des conséquences globales et individuelles désastreuses. Ceux-ci sont à l'origine d'énormes pertes économique à l'échelle mondiale, qui devraient atteindre 6 milliards de dollars US par an d'ici 2030. (Bloom *et al.*, 2012). Pour des cas sévères, l'espérance de vie peut diminuer de 10 à 20 ans (Organization *et al.*, 2015). Malgré ces répercussions, près de la moitié des personnes ayant besoin de soins de santé mentale n'obtiennent pas l'aide adéquate au Canada et aux États-Unis (Statistics Canada, 2019; Abuse *et al.*, 2020). Puisque la prévention efficace de troubles mentaux comme la dépression peut limiter les dégâts causés par ceux-ci (Muñoz *et al.*, 2010), leur détection précoce est un enjeu de taille.

Ce problème peut être approché grâce à des techniques d'Intelligence Artificielle (IA). Parmi les différentes approches existantes, il est possible de considérer le contenu textuel produit par des individus pour procéder à la détection du risque en santé mentale. Dans un tel cas, on fera plus précisément appel à des techniques issues du Traitement Automatique du Langage Naturel (TALN). Particulièrement, la modélisation de sujets et les plongements de mots sont des options intéressantes pour approcher ce type de problème. L'utilisation de ces techniques permet de travailler plusieurs tâches relatives à la détection de troubles mentaux, telles que l'**analyse de sentiments** et la prédiction via la classification.

Les travaux mettant de l'avant la modélisation de sujets et les plongements de mots se sont avérés prometteurs quant à l'évaluation et l'analyse du risque de divers troubles mentaux (Resnik *et al.*, 2015; Preoţiuc-Pietro *et al.*, 2015; Coppersmith *et al.*, 2015; Maupomé et Meurs, 2018; Ive *et al.*, 2018). Notamment, ceux-ci démontrent la grande versatilité des attributs obtenus grâce aux techniques utilisées. Par exemple, de tels attributs peuvent être mis à profit pour des tâches de régression et de classification de plusieurs types. Aussi, les résultats et leur interprétabilité présentent un bon potentiel lors de la mise en application de telles techniques du TALN. La compréhension du contexte de prédiction étant primordiale pour un professionnel de la santé mentale, les avancées présentées dans l'état de l'art sont encourageantes pour la détection automatique des troubles mentaux.

Toutefois, les diverses techniques utilisées dans ces travaux affichent des résultats variées selon le trouble mental étudié, suggérant la nécessité d'approfondir les travaux du domaine. À la vue des ces écarts de performances, la question de recherche suivante se pose :

Comment les attributs obtenus grâce à différentes techniques de modélisation permettent de détecter le risque lié à divers troubles mentaux ?

Le présent mémoire cherche donc à répondre à cette question, avec un focus sur des tâches de classification binaire et sur l'utilisation des modèles de sujets *Latent Dirichlet Allocation* (LDA) et *Embedded Topic Model* (ETM).

La structure de ce mémoire est la suivante. Le Chapitre 1 présente l'état de l'art des différentes approches en détection automatique du risque en santé mentale. Le Chapitre 2 présente les diverses techniques du TALN utilisées dans nos recherches. Le Chapitre 3 présente une publication parue à la conférence Canadian AI 2021 sur la détection du risque de la dépression, de l'auto-mutilation et de l'anorexie grâce à la modélisation de sujets. Le Chapitre 4 met en avant une seconde publication parue à la conférence Canadian AI 2021, portant sur l'application de la modélisation de sujets dans les espaces de plongements de mots dans le contexte de la détection du risque de la dépression. Ce mémoire se conclut par d'éventuelles pistes de travaux de recherche en modélisation de sujets pour la détection du risque en santé mentale, ainsi que par une mise en perspective du potentiel de cette approche.

CHAPITRE 1
DÉTECTION DE RISQUES EN SANTÉ MENTALE PAR LE TEXTE

1.1 Données

Les médias sociaux constituent une source de prédilection pour les données textuelles en lien avec la santé mentale. Sur les différentes plateformes, les personnes utilisatrices évoquent leurs pensées et sentiments sous formes de commentaires ou publications. La littérature présente plusieurs travaux mettant à profit de telles données pour procéder au dépistage de troubles de santé mentale.

Le contenu de Reddit [1] est fréquemment utilisé, notamment pour ses divers fils de discussions dédiés à la santé mentale, comme *Anxiety* [2], *StopSelfHarm* [3] et *SuicideWatch* [4]. Chaque fil porte généralement sur un thème bien précis, indiqué par son titre. Par exemple, les personnes utilisatrices de Reddit se rendant sur le fil *Anxiety* participent à des discussions en lien avec les troubles anxieux. Plusieurs types de personnes interagissent dans ces discussions : certaines personnes ont été diagnostiquées avec les troubles en question, d'autres se questionnent sur ceux-ci alors que d'autres souhaitent réconforter des personnes qui en ont besoin.

Gkotsis *et al.* (2016) ont étudié des contenus partagés dans quelques uns des fils de discussion de Reddit, incluant *Anxiety*, *StopSelfHarm* et *SuicideWatch*, afin d'évaluer le caractère unique de chaque fil selon les thèmes abordés par les personnes utilisatrices. Par exemple, le fil *Anxiety* est considéré unique si le thème abordé dans celui-ci ne se retrouve pas dans les autres fils étudiés. Suite à leur évaluation, les chercheurs ont construit un corpus textuel regroupant les contenus de 16 fils de discussions, soient ceux qui abordent un thème s'avérant unique parmi les fils étudiés. Ce corpus, bien que non annoté au niveau des publications individuelles, permet de représenter la langue relative à des troubles mentaux précis, avec des approches de modélisation de sujets et de plongements de mots à apprentissage non supervisé. En effet, les expériences des chercheurs ont révélé que le texte présent dans chaque fil était distinctif en comparaison aux autres fils étudiés. Ainsi, utiliser un tel corpus pour des tâches de modélisations de sujets et de

1. Reddit. `https://www.reddit.com`
2. Anxiety Subreddit. `https://www.reddit.com/r/Anxiety`
3. StopSelfHarm Subreddit. `https://www.reddit.com/r/StopSelfHarm`
4. SuicideWatch Subreddit. `https://www.reddit.com/r/SuicideWatch`

plongements de mots permet de capturer le langage relatif aux différents troubles mentaux observés. Par exemple, en modélisation de sujets, l'utilisation de ce corpus permettrait de modéliser des sujets fortement liés à certains troubles, comme les troubles anxieux et l'automutilation, dû à la présence des fils *Anxiety* et *StopSelfHarm* dans le corpus. Des modèles créés à partir d'un tel corpus ouvrent donc la porte à la détection de risques en santé mentale par le texte, puisqu'ils permettent d'évaluer les productions textuelles d'une personne et d'identifier si cette dernière utilise un langage similaire à celui présent dans le corpus.

Toutefois, des annotations sont nécessaires quand il est question de tâches de classification. Coppersmith *et al*. (2014) proposent corpus annoté, constitué de données publiques obtenues sur Twitter [5]. Celui-ci s'intéresse à 4 troubles de santé mentale, à savoir la dépression, le trouble bipolaire, le trouble de stress post-traumatique et le trouble affectif saisonnier. Pour annoter les données, les créateurs attribuent une étiquette positive aux personnes utilisatrices s'étant déclarées diagnostiquées dans leur publications. Des personnes utilisatrices de contrôle sont choisies au hasard parmi celles n'ayant pas fait une telle déclaration, de façon à équilibrer le corpus. Des corpus portant sur d'autres troubles de santé mentale, comme l'anorexie et l'auto-mutilation, suivant une règle d'annotation similaire, sont fournis dans le cadre de la campagne d'évaluation eRisk [6] (Losada *et al.*, 2018, 2019, 2020), ceux-ci puisant leur contenu dans les fils de discussion de Reddit. Comme d'autres campagnes d'évaluation, eRisk proposent divers problèmes à la communauté scientifique, ces problèmes étant présentés sous la forme de tâches à accomplir accompagnées de règles précises. Dans le cas d'eRisk, les tâches sont reliées à la détection de risques en santé mentale par le texte. Les équipes de recherche désirant participer à la campagne d'évaluation créent des systèmes visant à régler les problèmes proposés et soumettent les résultats obtenus. Les organisateurs de la campagne d'évaluation peuvent ensuite mesurer et comparer les résultats soumis par les différentes équipes, de façon à exposer la performance des différents systèmes. La publication présentée au Chapitre 3 de ce mémoire découle notamment de ma participation avec notre équipe à l'édition 2020 d'eRisk.

Les méthodes d'annotation semi-automatiques sont beaucoup moins coûteuses en ressources et en temps qu'une collecte de type questionnaire obtenue dans un cadre clinique. En revanche, elles sont également sujettes aux erreurs d'annotations. Une fausse déclaration de diagnostic introduit un faux positif dans un corpus et une déclaration de diagnostic absente, un faux négatif.

5. Twitter. `https://twitter.com`

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

À l'opposé, la confirmation clinique du diagnostic d'une personne permet d'éviter les données mal annotées. Par exemple, le corpus présenté par Merchant *et al.* (2019) est composé de statuts Facebook obtenus avec l'autorisation de participant.e.s d'une étude clinique. En parallèle, ces personnes ont également fourni leur dossier médical de façon à identifier leurs différents diagnostics. Bien que la qualité soit au rendez-vous, il est également important de considérer la quantité de données cliniques amassées pour la création d'un corpus, puisque celle-ci peut avoir une incidence sur l'entraînement des modèles.

1.2 Approches

Les systèmes de détection de troubles mentaux basés sur la modélisation de sujets et les plongements de mots sont abordés fréquemment dans la littérature. Resnik *et al.* (2015) comparent divers types de modèle de sujets afin d'analyser le langage lié à la dépression. Alors que tous les modèles découlent de LDA, certaines versions se basent sur des étiquettes ou sur des mots d'ancrage, inclus dans le jeu de données et liés au domaine d'intérêt, afin de guider l'apprentissage des sujets. Avec leurs expériences, les auteurs mettent en évidence la performance des attributs obtenus selon le type de modèle.

Preoţiuc-Pietro *et al.* (2015) mettent en application diverses techniques, telles que LDA et Word2Vec, afin d'obtenir des grappes de termes pour observer les sujets abordés par les utilisateurs de réseaux sociaux. Pour justifier leur approche, les auteurs se basent sur l'intuition que les grappes obtenues sont évocatrices en termes de contexte et de sémantique. Les grappes sont ensuite combinées entre-elles pour procéder à une tâche de classification binaire. Une étude comparative (Coppersmith *et al.*, 2015) révèle les performances prometteuses des deux précédentes approches dans un contexte de classification binaire considérant la dépression et le syndrome post-traumatique.

Plus récemment, Maupomé et Meurs (2018) ont fait appel à un modèle LDA pour une tâche de classification binaire pour la détection précoce de la dépression. Ceux-ci font usage d'attributs obtenus grâce au modèle de sujets pour entraîner un perceptron multicouche. Ce dernier est ensuite utilisé dans un module de prédiction, dans lequel les chercheurs proposent un seuil de prédiction qui décroît de façon itérative. La quantité des productions textuelles évaluées augmente à chaque itération dans la tâche considérée. Lors de la première itération, la quantité des productions textuelles étant petite, le seuil de prédiction doit être élevé, c'est-à-dire strict, pour ne pas émettre trop facilement de prédictions faussement positives. En effet, la présence de plusieurs mots liés à la dépression dans une petite quantité de texte pourrait biaiser les résultats. Étant décroissant à chaque itération, le seuil de prédiction permet de plus en plus facile-

ment l'émission de prédictions positives, soit au fur et à mesure que la quantité des productions textuelles évaluées augmente. L'approche proposée par les chercheurs s'est avérée prometteuse dans ce contexte.

Parallèlement, Ive *et al.* (2018) utilisent un modèle neuronal hiérarchique comprenant une couche de plongements de mots dans une tâche de classification en classes multiples. Considérant le langage associé à 11 troubles mentaux différents, le but du modèle est d'identifier à quel trouble les extraits de textes donnés en entrée sont associés.

Ma participation au travaux de l'équipe de recherche de Marie-Jean Meurs à l'Université du Québec à Montréal a permis de proposer deux autres approches de détection de risques en santé mentale mettant à profit la modélisation de sujet. La première approche (Maupomé *et al.*, 2021a) utilise la similarité textuelle afin de prédire les réponses du *Beck Depression Inventory* (BDI) (Beck *et al.*, 1961). Le BDI est un formulaire d'auto-évaluation en 21 points mesurant la sévérité de la dépression. Obtenir les prédictions de réponses à ce formulaire pourrait donc permettre d'estimer le niveau de dépression d'un individu. Pour émettre une prédiction, l'approche mesure la similarité entre les productions textuelles de personnes utilisatrices de réseaux sociaux et des formulaires BDI préalablement complétés. Les travaux comparent l'utilisation de la modélisation de sujets grâce à LDA à celle d'encodeurs neuronaux pour mesurer la similarité textuelle, toutes deux présentant des résultats prometteurs. La deuxième approche (Maupomé *et al.*, 2021b) découle de nos recherches pour l'édition 2021 de la campagne d'évaluation d'eRisk (Parapar *et al.*, 2021). cette approche utilise également la similarité textuelle, mais pour prédire le risque de dépendance au jeu pathologique. Pour y arriver, l'approche mesure la similarité entre les productions textuelles de personnes utilisatrices de réseaux sociaux et des témoignages de joueurs pathologiques. Les travaux se concentrent sur l'utilisation de la modélisation de sujets grâce à ETM pour mesure la similarité textuelle. Le performances obtenues suggèrent que l'approche est intéressante pour détecter les individus à risque de dépendance de jeu pathologique.

1.3 Applications et considérations

Les différentes approches de détection de risques en santé mentale par le texte ouvre la voie à plusieurs applications concrètes. Notamment, les approches automatiques pourraient être d'une grande utilité pour les professionnel.le.s de la santé mentale. Les personnes utilisatrices de réseaux sociaux sont susceptibles de dévoiler des informations sur leur condition à leurs amis et connaissances sur ces plateformes. Lors de consultations, une personne utilisatrice pourrait autoriser les professionnel.le.s s'occupant de son dossier

à avoir accès à ses productions textuelles présentes sur les réseaux sociaux, de façon à aider l'évaluation de son cas. Or, une énorme quantité d'information se retrouve sur ces réseaux et il serait difficile pour les professionnel.le.s de parcourir manuellement toutes les productions textuelles d'une personne. L'utilisation de systèmes mettant à profit les approches automatiques peut régler ce problème. Grâce à ces systèmes, les professionnel.le.s de la santé mentale pourraient avoir rapidement accès à différentes informations à partir des productions textuelles d'une personne : résumé des textes, analyse des sentiments observés et risques potentiels. Ces informations pourraient alors être utilisées par les professionnel.le.s pour venir appuyer leur choix de diagnostics ou encore pour poser des actions plus rapidement en cas de besoin. Il est à noter que les systèmes automatiques ne viendraient pas remplacer les professionnel.le.s de la santé mentale. Il s'agit simplement d'outils supplémentaires pour les aider dans leur prise de décision.

De plus, de tels systèmes automatiques pourraient être mis en place directement sur les réseaux sociaux. Il serait alors possible pour les modérateur.trice.s d'assurer un suivi des publications textuelles évoquant des risques de santé mentale. Par exemple, il serait possible d'alerter les professionnel.le.s adéquats si un risque de tentative de suicide est détecté.

Toutefois, l'utilisation de ces systèmes automatiques soulèvent également des questions éthiques. La mise en place de ces systèmes sur les réseaux sociaux pourrait facilement devenir pernicieuse. En effet, entre de mauvaises mains, ces systèmes pourraient devenir des outils de surveillance, présentant des enjeux éthiques importants. Il est important que la détection de risques en santé mentale par le texte soit utilisée pour aider les individus et non pour leur nuire. La mauvaise utilisation de tels systèmes pourrait aussi mener à des cas de discrimination. Par exemple, des employeurs pourraient analyser les productions textuelles publiques sur les réseaux de personnes postulant pour un emploi. En cas de risques de santé mentale détectés, ces employeurs pourraient décider de rejeter la candidature des personnes concernées.

Cependant, des moyens existent pour s'assurer que ces systèmes ne servent qu'à améliorer la condition des personnes requérant des soins de santé mentale. Notamment, des actions concrètes peuvent être posées par les instances gouvernementales pour protéger la population. Par exemple, des lois afin d'interdire le potentiel usage néfaste de ces systèmes de détection pourraient voir le jour. Les systèmes de détection automatique de risques de santé mentale par le texte demeurent des outils pour les humains et, comme tout outil, un cadre réglementaire peut être mis en place pour assurer que leur usage sert le bien commun.

CHAPITRE 2
NOTIONS PRÉLIMINAIRES

2.1 Réseaux neuronaux

Les réseaux neuronaux consistent en une famille d'algorithmes fréquemment utilisés en IA. Leur structure, composée de neurones agencés sous plusieurs couches, permet d'estimer des fonctions complexes. Dans le cas le plus simple, celui du perceptron classique (Rosenblatt, 1958), le réseau neuronal est formé de deux couches, soit la couche d'entrée et la couche de sortie. Le réseau est alors dit monocouche, puisqu'il ne contient qu'une seule couche, celle de sortie, en plus de sa couche d'entrée. Dans d'autres cas plus complexes, le réseau neuronal peut également comporter une ou plusieurs couches cachées entre ses couches d'entrée et de sortie. La Figure 2.1 présente l'exemple du **réseau neuronal à deux couches**, soit composé de seulement une couche cachée et une couche de sortie en plus de sa couche d'entrée.

Alors que les couches cachées et de sortie sont composées de neurones, la couche d'entrée sert à passer les données d'entrée au reste du réseau. Chaque neurone comporte des **poids**, jouant un rôle important pour son activation, soit l'état du neurone après la réception de ses données d'entrée, déterminant le signal envoyé en sortie au reste du réseau. Les **poids** d'un neurone pondèrent ses données d'entrée avant de procéder au calcul de son activation. Si le neurone est activé, il envoie un signal, qui sert de donnée d'entrée aux neurones de la couche suivante du réseau. Dans les faits, les **poids** du réseau neuronal constituent ses paramètres, c'est-à-dire qu'ils sont ajustés lors de l'entraînement, permettant l'apprentissage du modèle. L'activation d'un neurone prend également en compte un **biais**, soit une valeur constante permettant d'influencer l'état du neurone. Le **biais** permet notamment de limiter le surentraînement, c'est-à-dire un apprentissage trop ajusté aux données d'entrée résultant en un modèle ayant tendance à estimer une mauvaise sortie pour des données non observées à l'entraînement. La Figure 2.2 illustre les différentes composantes d'un neurone.

Pour déterminer l'activation d'un neurone, l'activation nette net et une fonction d'activation $f(\cdot)$ sont nécessaires. Lorsque la $t-1^{\text{e}}$ couche contient N neurones, l'activation nette du j^{e} neurone de la t^{e} couche est donnée par

$$net_j = \sum_{i=0}^{\mathrm{N}} x_i w_{ji} + w_{j0},$$

Figure 2.1 Réseau neuronal à deux couches.

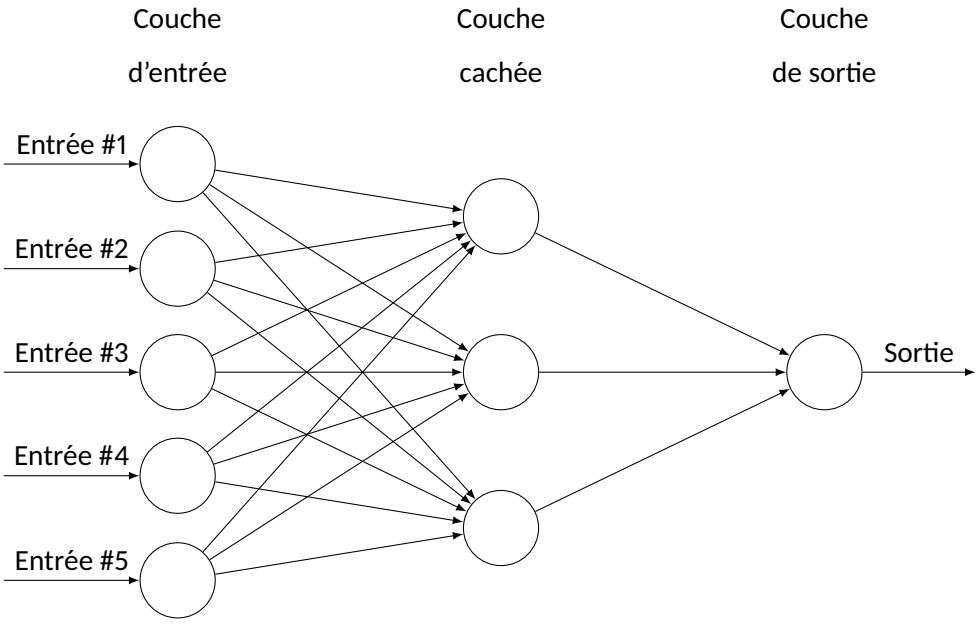


Figure 2.2 Structure d'un neurone

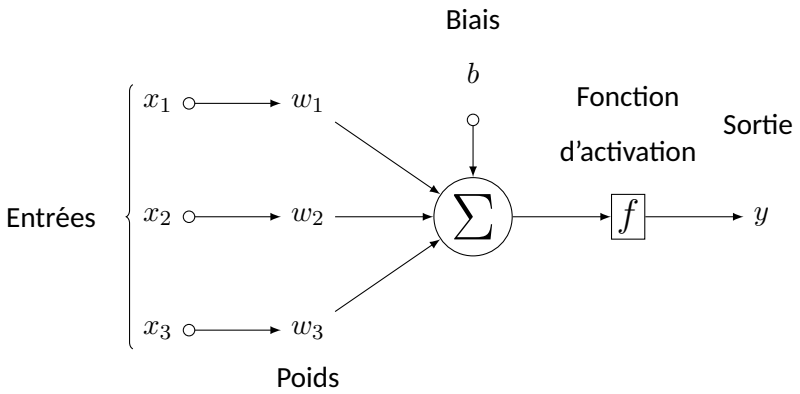
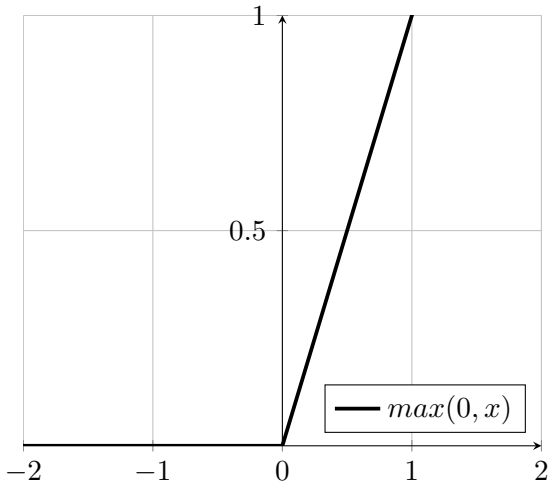


Figure 2.3 La fonction ReLU.



où w_{ji} est le **poids** entre le i^{e} et le j^{e} neurone et w_{j0} est le **biais** du j^{e} neurone. Alors, l'activation du j^{e} neurone est obtenue par $f(net_j)$. En pratique, la fonction d'activation choisie est généralement la fonction logistique ou encore la fonction ReLU, définies respectivement aux Équations 2.1 et 2.2.

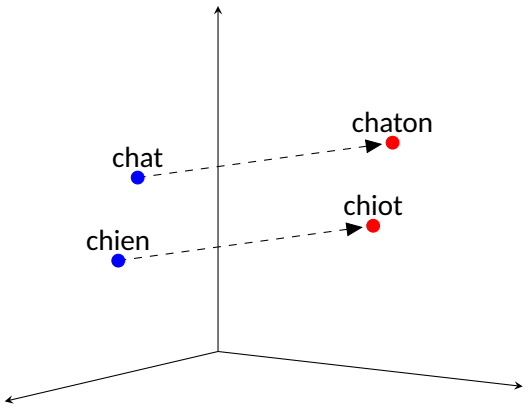
$$\sigma(x) = \frac{1}{1 + e^{-x}} \tag{2.1}$$

$$ReLU(x) = max(0, x) \tag{2.2}$$

Par exemple, si la fonction d'activation sélectionnée est ReLU, le j^{e} neurone est dit activé si $ReLU(net_j) > 0$. Autrement, le signal envoyé aux neurones de la prochaine couche du réseau est nul puisque celui-ci est égal à 0. En observant la Figure 2.3, illustrant la fonction ReLU, il est possible de constater que ce dernier cas survient uniquement lorsque la valeur de net_j est négative. Ceci met en avant l'importance des **poids** et du **biais** d'un neurone, qui ont un impact direct sur le résultat obtenu pour net_j.

Lors de l'entraînement, les paramètres du modèle, soit les **poids**, sont appris par **descente de gradient** dans un processus de **rétropropagation** (Rumelhart *et al.*, 1986). Une fois les sorties obtenues lors d'une propagration avant, les paramètres du modèle sont ajustés en sens inverse du réseau, de façon à minimiser l'erreur d'entraînement du modèle.

Figure 2.4 Relation sémantique entre des mots dans un espace vectoriel.



2.2 Plongements de mots

2.2.1 Fondements

Le concept de proximité sémantique sert de fondation à la technique des plongements de mots. En effet, cette dernière permet d'encoder des mots similaires par des vecteurs rapprochés, c'est-à-dire qui sont séparés par une petite distance, dans un espace vectoriel. La Figure 2.4 illustre la sortie recherchée d'un algorithme de plongements de mots afin de représenter la **relation sémantique**, soit la relation expliquant les liens de signification existants entre les mots. Comme *chat* est à *chaton* ce que *chien* est à *chiot*, les distances entre leurs vecteurs dans l'espace de plongements devraient refléter la **relation sémantique** qui les lie. Cette représentation multidimensionnelle est utile à diverses tâches d'IA, notamment l'**analyse de sentiments**.

Deux familles d'approches dominent le domaine des plongements de mots : les approches à modèle vectoriel et les approches à représentation distribuée. La première famille d'approches est à l'origine de la représentation de termes dans un espace multidimensionnel. Celle-ci se base sur les matrices de cooccurrence pour modéliser la **relation sémantique** entre les mots. Étant donné un corpus, la représentation de la relation entre deux termes peut être donnée par le nombre de cooccurrences par phrase de ces derniers. Le Tableau 2.1 présente la matrice obtenue pour un tel exercice sur un corpus contenant les phrases *Le chat est assis sur le banc* et *Le livre est posé sur le banc*. En pratique, considérer seulement le compte des cooccurences manque de finesse pour capturer la **relation sémantique** entre les termes d'un corpus. Des métriques plus précises sont généralement choisis, comme la fréquence de cooccurrence. Pour un corpus de N mots, la matrice de cooccurence générée est de taille $\mathrm{N} \times \mathrm{N}$. Elle est donc généralement très creuse et difficilement utilisable en pratique. Par exemple, l'application de mesures de distance dans un

Tableau 2.1 Exemple de matrice de cooccurrence de mots.

	chat	assis	banc	livre	posé
chat	1	1	1	0	0
assis	1	1	1	0	0
banc	1	1	1	1	1
livre	0	0	1	1	1
posé	0	0	1	1	1

Tableau 2.2 Exemple de représentation distribuée de mots.

	inanimé	se déplace	en bois	en papier
chat	0	1	0	0
banc	1	0	1	0
livre	1	0	0	1

contexte de très grandes dimensions donne des résultats peu significatifs ; plus l'espace vectoriel est grand, plus les vecteurs semblent à la même distances les uns des autres. Or, comprendre la distance séparant les vecteurs de mots dans l'espace de plongements est nécessaire, puisqu'elle permet d'interpréter la **relation sémantique**. Ce phénomène, aussi nommé ≪le fléau de la dimension≫, peut être atténué par diverses techniques de réduction de dimension, telle que la méthode de la décomposition en valeurs singulières (DVS). Toutefois, ces techniques s'avèrent généralement insuffisantes pour construire des plongements de mots capturant correctement la **relation sémantique** entre les mots d'un corpus.

La deuxième famille d'approches, initialement apparue grâce au travaux de Bengio *et al.* (2000), a permis de remédier à ces problèmes de dimension. Les approches issues de cette famille se basent sur une représentation distribuée des mots présents dans le corpus, où chacune des dimensions de l'espace vectoriel correspond à une caractéristique, de l'anglais *feature*, soit un aspect qualifiant un mot. Par exemple, on peut imaginer qu'une dimension d'une telle représentation capture la caractéristique *inanimé* pour les mots d'un corpus. Ainsi, des mots comme *banc* et *livre* devrait avoir des valeurs similaires pour cette caractéristique, alors que le mot *chat* non. Ces caractéristiques permettent de rapprocher certains mots dans l'espace vectoriel, lorsque leur sens est similaire. La Tableau 2.2 donne un exemple concret d'une représentation distribuée de mots. L'appartenance à une caractéristique y est donnée par une valeur bi-

naire ; 1 si appartenance, 0 sinon. Une ligne de la matrice est alors un vecteur de valeurs binaires de dimension D, soit le nombre de caractéristiques du modèle, permettant de positionner un mot dans l'espace de plongements. Intuitivement, faire appel à une représentation distribuée de mots permet une meilleure généralisation qu'une représentation vectorielle comme la matrice de cooccurrence. En effet, l'apprentissage d'une représentation distribuée de mots considère le contexte dans lequel les mots apparaissent. Par exemple, en présence des phrases *Le chat est assis sur le banc* et *Le chat est couché sur le banc*, l'apprentissage de la représentation distribuée de mots capte le contexte pour que les mots *assis* et *couché* soient rapprochés dans l'espace de plongements. Le modèle généralise donc le niveau d'appartenance des mots à certaines caractéristiques, ce que la matrice de cooccurrence ne parvient pas à faire. Dans la pratique, une représentation distribuée de mots est beaucoup plus utile qu'un modèle vectoriel comme la matrice de cooccurrence, puisque celle-ci est de taille $\mathrm{N} \times \mathrm{D}$, réglant ainsi le problème du fléau de la dimension.

L'architecture de cette famille d'approches repose sur les réseaux neuronaux, tels que présentés à la partie 2.1. Lors de l'entraînement des plongements de mots, les paramètres du modèle sont appris par **descente de gradient** dans un processus de **rétropropagation** de façon à maximiser la vraisemblance du modèle. Le but de ce processus est d'ajuster les paramètres du modèle de façon à ce que les plongements de mots obtenus soient les plus probables possibles en fonction des données observées en entrée. Dans un tel cas, on dit que les paramètres du modèle sont les plus vraisemblables possibles, donc que la vraisemblance du modèle est maximisée. Le processus d'apprentissage de ce type de modèle est toutefois coûteux dû à la grande quantité de paramètres du modèle. Depuis les travaux initiaux, des avancées ont permis la création de modèles beaucoup plus rapides, tel que Word2Vec (Mikolov *et al.*, 2013a).

2.2.2 Approche à représentation distribuée : Word2Vec

Afin d'évaluer une représentation distribuée des mots présents dans le corpus, le modèle Word2Vec considère chacun des mots en fonction de son contexte. Ainsi, la représentation vectorielle apprise pour un mot dépend des mots voisins de ce dernier dans le corpus.

Le modèle Word2Vec regroupe deux algorithmes d'apprentissage, Skip-gram et *Continuous Bag-of-Words* (CBOW) (Mikolov *et al.*, 2013a), approchant chacun le contexte différemment. Le premier, Skip-gram, vise à prédire le contexte à partir d'un mot donné. À l'inverse, le second, CBOW, vise à prédire un mot à partir d'un contexte donné.

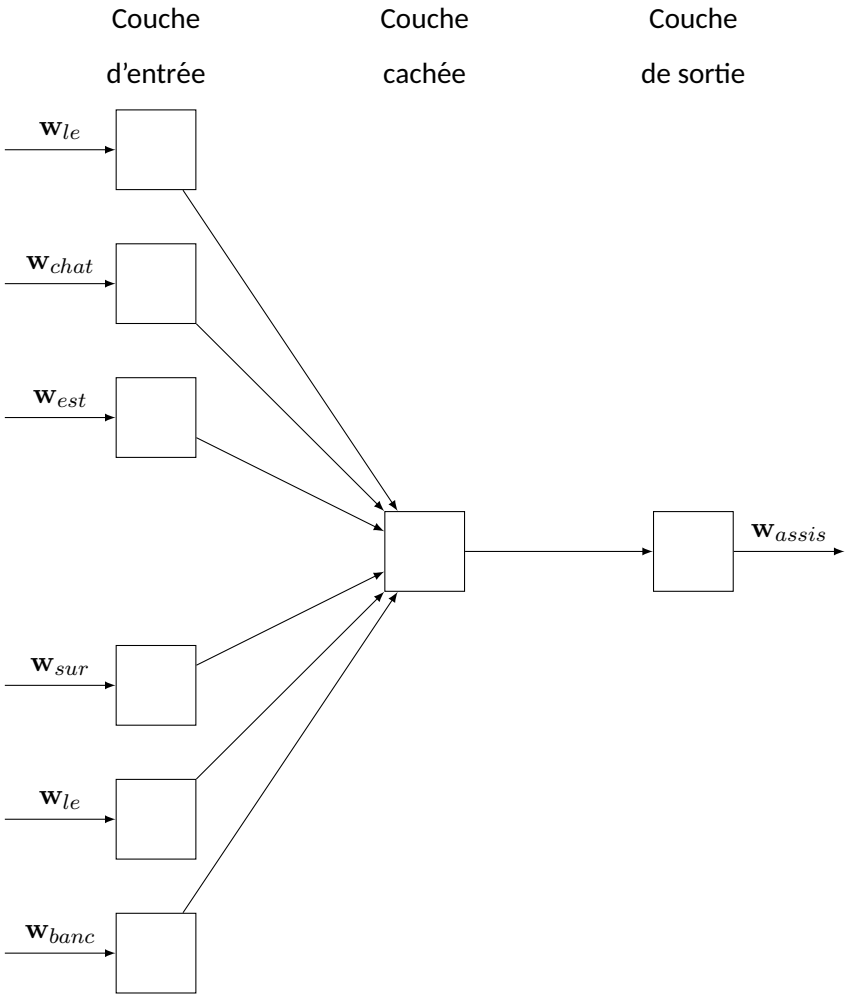
En pratique, chacun des deux algorithmes est basé sur une architecture de **réseau neuronal à deux couches**. Étant donnée une fenêtre de contexte de taille C, le modèle prend en entrée et donne en sorties les représentations vectorielles des différents mots issues du contexte, selon l'algorithme choisi. Par exemple, pour une fenêtre de contexte *le chat est assis sur le banc* où $\mathrm{C} = 3$, le modèle Skip-gram utilisera le vecteur du mot *assis* pour estimer les vecteurs de tous les autres mots du contexte. À l'inverse, CBOW utilisera les vecteurs des mots du contexte pour estimer le vecteur du mot *assis*. Les Figures 2.5 et 2.6 présentent respectivement l'architecture de CBOW et Skip-gram pour l'exemple considéré, où $\mathbf{w}$ est la représentation vectorielle d'un mot donné. À des fin de simplification, les différents éléments de chaque couche ont été illustrés par des carrés. Les détails complets de chacun de ces éléments sont expliqués dans la publication de Mikolov *et al.* (2013a). Comme on peut le voir dans ces figures, l'apprentissage du modèle est intimement lié au contexte dans lequel le mot d'intérêt se retrouve. Lors de l'entraînement, les poids de la couche cachée sont ajustés de façon à apprendre les plongements de mots du modèle, qui sont représentés par une matrice de taille $\mathrm{N} \times \mathrm{D}$, où N est le nombre de mots du vocabulaire et D le nombre de caractéristiques du modèle. En pratique, la valeur de D est donnée en paramètre d'entrée du modèle.

Pour optimiser la **relation sémantique** apprise, l'algorithme Skip-gram (Mikolov *et al.*, 2013b) modélise sa représentation distribuée en maximisant la log probabilité des mots du contexte étant donné le mot considéré, telle que

$$\frac{1}{\mathrm{T}} \sum_{t=1}^{\mathrm{T}} \sum_{-\mathrm{C} \leq j \leq \mathrm{C}, j \neq 0} \log P(w_{t+j} | w_t),$$

où T est le nombre de mots dans l'ensemble d'entraînement et w_t le t^{e} mot. En effet, le but du modèle étant de capturer la **relation sémantique** entre un mot et son contexte, la probabilité d'observer les mots du contexte étant donné le mot considéré devrait être la plus haute possible. Pour reprendre l'exemple précédent, *le chat est assis sur le banc*, une haute probabilité du mot *chat* étant donné le mot *assis* indique que de retrouver le mot *chat* dans le contexte du mot *assis* est très probable. À l'inverse, ce processus de maximisation ne donnerait pas une haute probabilité d'un mot comme *livre* étant donné le mot *assis* si ces deux mots n'apparaissent pas dans le même contexte, reflétant la **relation sémantique** entre ces deux mots. Pour l'implémentation mathématique de ce processus de maximisation, les détails du calcul de $P(w_{t+j}|w_t)$ sont donnés dans la publication de Mikolov *et al.* (2013b). Dans le cas de l'algorithme CBOW, les mêmes étapes sont suivies, mais cette fois-ci pour la probabilité du mot considéré étant donné le contexte, soit $P(w_t|w_{t+j})$.

Figure 2.5 Architecture de l'algorithme *Continuous Bag-of-Words* de Word2Vec pour l'exemple *le chat est assis sur le banc*.

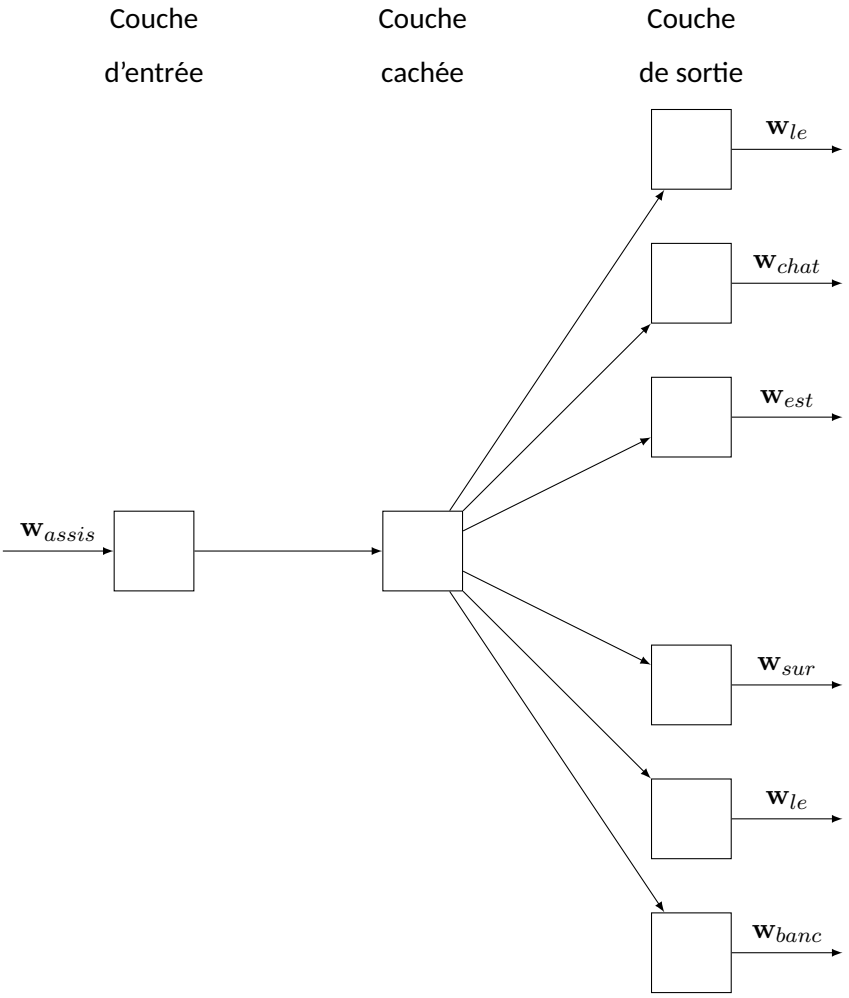


2.3 Modélisation de sujets

2.3.1 Fondements

Un sujet est une matière abordé dans un texte. En général, un document est composé de plusieurs sujets, où certains prédominent sur d'autres. Étant donné un corpus de documents, la modélisation de sujets est une technique expliquant la structure des sujets inclus dans le corpus ainsi que leur distribution pour chacun des documents. Pour parvenir à représenter les sujets d'un corpus, cette technique se base deux intuitions fondamentales. Premièrement, un document constitue un mélange de plusieurs sujets, mais traite principalement d'une petite partie de ceux-ci. Deuxièmement, un sujet est représenté par une collection de mots plus ou moins distinctifs pour celui-ci. Prenons un exemple concret pour illustrer ces deux intuitions :

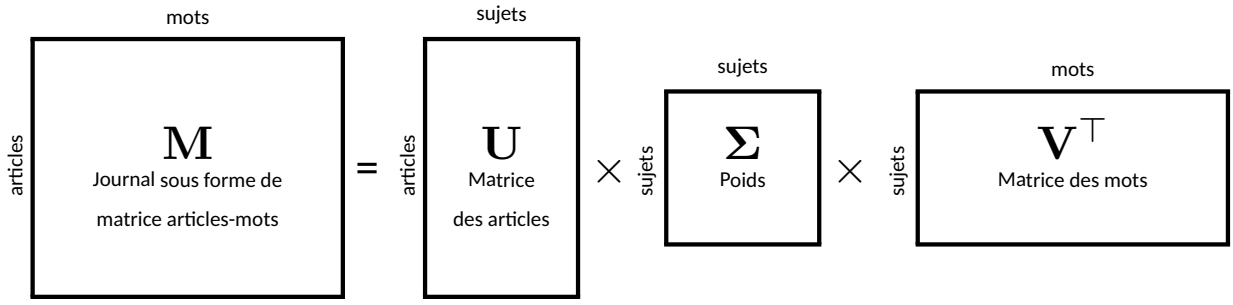
Figure 2.6 Architecture de l'algorithme Skip-gram de Word2Vec pour l'exemple *le chat est assis sur le banc*.



un journal regroupe des articles sous les sections santé, culture et économie ; il s'agit donc d'un corpus de documents. Un article de la section culture traite du financement des artistes de la relève, alors celui-ci est composé d'au moins deux sujets, un plus culturel et un plus économique. Selon le texte de l'article, il est possible qu'un des deux sujets soit présent en plus grande proportion, donc qu'il explique le mieux le contenu observé. D'autre part, chacun de ces sujets a un vocabulaire qui lui est associé ; *spectacle* est associé à la culture alors qu'*inflation* est associé à économie. Toutefois, certains mots plus généraux sont parfois associés à plusieurs sujets ; ils sont donc moins distinctifs. Dans notre exemple, *tarif* peut être lié autant au sujet culturel, *tarif de spectacle*, qu'à l'économique, *tarif bancaire*.

Pionière dans le domaine, la technique de l'analyse sémantique latente (LSA), de l'anglais *Latent Semantic Analysis*, fait appel à la réduction de dimensions pour tenter d'exposer les sujets qui se cachent dans la struc-

Figure 2.7 Décomposition en valeurs singulières d'un journal sous forme de matrice articles-mots.



ture d'un corpus (Deerwester *et al.*, 1990). Cette structure, dite latente, est la composition non-observable du corpus qui explique les liens entre les différents documents. Pour une matrice de documents-termes donnée, représentant le corpus, le modèle LSA met à profit la méthode DVS afin de factoriser la matrice telle que

$$\mathbf{M} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^\top. \tag{2.3}$$

Cette décomposition exhibe les matrices $\mathbf{U}$ et $\mathbf{V}$, qui représentent respectivement la matrice documents-sujets et la matrice termes-sujets, où les colonnes sont triées en ordre décroissant conformément aux **poids** décrits par la matrice singulière $\mathbf{\Sigma}$. Ainsi, les K premières colonnes des matrices $\mathbf{U}$ et $\mathbf{V}$ contiennent les K sujets exprimant le mieux le corpus donné en entrée. En pratique, le nombre de sujets K est donné en paramètre. Si l'on suit notre exemple précédent, il est possible de représenter un journal sous forme de matrice où chaque ligne correspond à un article et chaque colonne à un mot. Cette matrice est décomposable en utilisant la méthode DVS, comme présentée dans la Figure 2.7. Ainsi, les K premières colonnes des matrices $\mathbf{U}$ et $\mathbf{V}$ expliquent les sujets les plus représentatifs de notre journal initial. Conséquemment aux sections du journal, il est possible d'observer des sujets fortement liés à la culture, santé et économie dans les K colonnes.

Toutefois, bien que théoriquement pertinente, LSA s'avère peu pratique, notamment dû à l'immense quantité de données nécessaire pour obtenir des résultats significatifs. Par exemple, un journal avec seulement une centaine d'articles ne serait pas assez volumineux pour obtenir des sujets interprétables.

Plus tard, la technique de l'analyse sémantique latente probabiliste (PLSA), de l'anglais *Probabilistic Latent Semantic Analysis*, a remplacé l'utilisation de DVS par une approche probabiliste afin d'expliquer la structure latente de sujets d'un corpus (Hofmann, 1999). Cette technique fait appel à un **processus génératif**, soit un processus appris dans le but d'approximer la sortie produite par un processus réel. En d'autres

termes, le **processus génératif** comporte des paramètres qui sont ajustés de façon à répliquer au mieux ses données d'entrée. Dans le cas de la modélisation de sujets, un tel processus tente de reproduire les documents originaux selon leur composition de sujets. En ce sens, PLSA modélise $P(d, w)$, soit la **probabilité conjointe** d'observer un document d et un terme w ensemble, de façon à remplir chaque case de la matrice de documents-termes. Pour y parvenir, la technique estime les probabilités par rapport à tous les sujets $z \in \mathcal{Z}$.

$$P(d, w) = P(d) \sum_{z \in \mathcal{Z}} P(z|d) P(w|z) \tag{2.4}$$

La probabilité $P(d, w)$ modélisée par PLSA est exprimée par l'Équation 2.4, où $P(d)$ et $P(z)$ sont respectivement les probabilités d'observer un document d et un sujet z dans un corpus donné. $P(z|d)$ et $P(w|z)$ sont estimées par lors du **processus génératif**. En pratique, l'**algorithme d'espérance-maximisation** est utilisé pour y parvenir. Ce dernier étant hors de la portée de notre étude, plus d'explications sur son fonctionnement sont disponibles dans la publication de Dempster *et al.* (1977).

PLSA débutant par l'estimation de la probabilité $P(d)$, certaines limitations se posent quant à son utilisation. Notamment, il est impossible pour le modèle d'évaluer les sujets d'un document non observé lors de l'entraînement, ce qui constitue un frein considérable à la mise en œuvre de cette technique. Pour reprendre l'exemple du journal, un modèle PLSA entraîné sur les articles d'une certaine édition ne pourrait estimer les sujets des articles issus d'une autre. Pourtant, les deux éditions du journal comportent les mêmes sections et possiblement un vocabulaire similaire. Or, seul les sujets des articles observés lors de la modélisation peuvent être estimés *a posteriori*. La mise en production de modèles ayant souvent pour but d'évaluer de nouvelles données, cette réalité s'avère hautement contraignante dans la pratique.

2.3.2 Approche bayésienne : *Latent Dirichlet Allocation*

Bien que les techniques LSA et PLSA permettent d'évaluer la structure latente de sujets d'un corpus, celles-ci présentent des limitations importantes empêchant leur déploiement à grande échelle, notamment leur incapacité à évaluer la composition de sujets d'un document non observé à l'entraînement. La technique LDA (Blei *et al.*, 2003) propose une version bayésienne de PLSA, faisant appel à des distributions suivant la **loi de Dirichlet**, afin de régler ces problèmes. La **loi de Dirichlet** est une loi de probabilité à variables aléatoires multinomiales, notamment utile lorsqu'un problème est lié à une distribution considérant plusieurs dimensions. En effet, la distribution de cette loi est paramétrée par un vecteur de nombre réels positifs, soient ses paramètres de concentration.

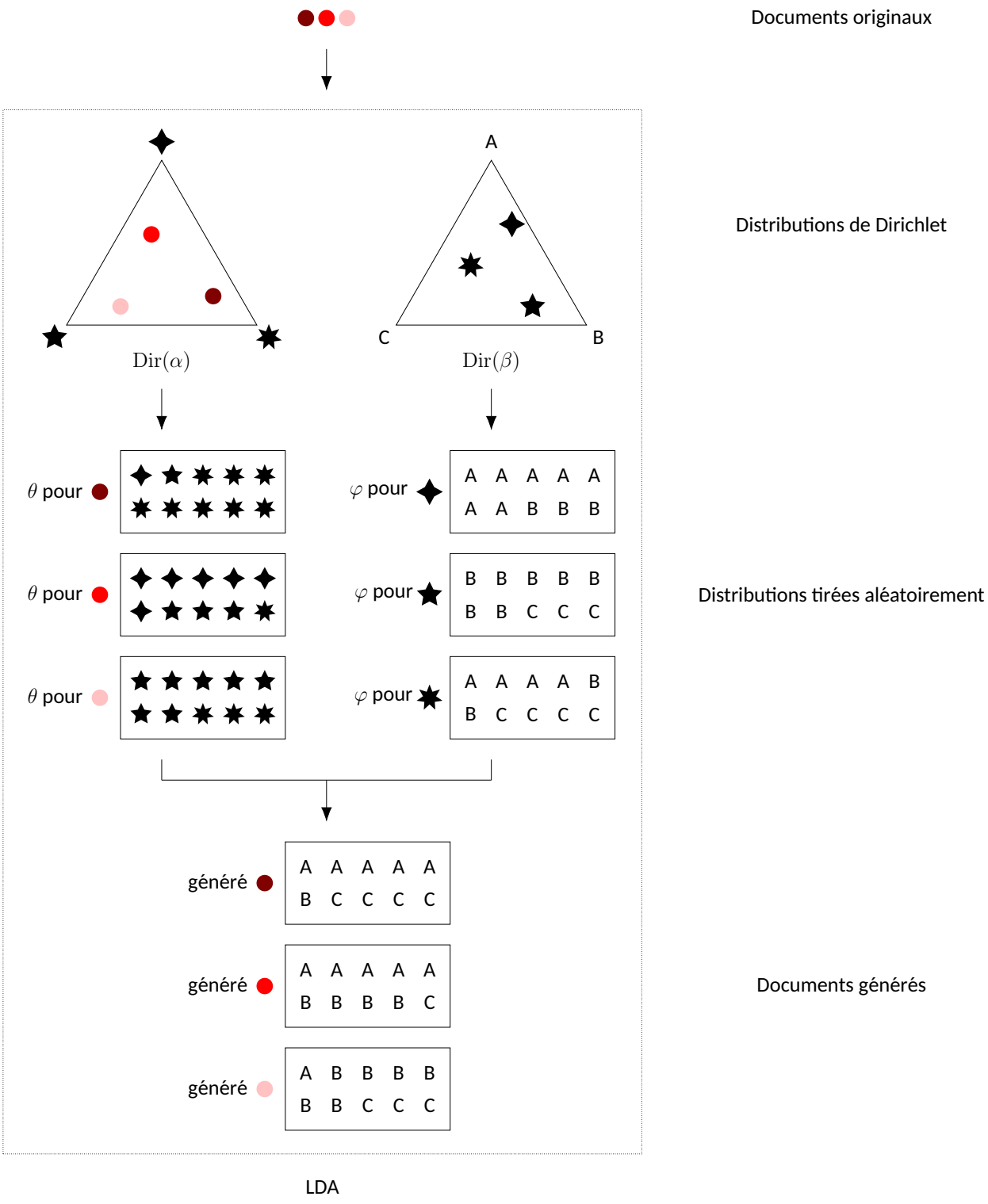
Dans le cas de LDA, deux distributions sont prises en considération de façon à représenter au mieux la structure de sujets observée : $\mathrm{Dir}(\alpha)$, la distribution de Dirichlet des sujets par document, et $\mathrm{Dir}(\beta)$, celle des termes par sujets, où α et β représentent les paramètres de concentration des distributions de Dirichlet dénotées $\mathrm{Dir}(\cdot)$. Ici, $\mathrm{Dir}(\alpha)$ et $\mathrm{Dir}(\beta)$ expliquent respectivement les relations documents-sujets et termes-sujets. Pour reprendre l'exemple du journal, le vecteur α pourrait avoir 3 dimensions, une pour chacune des sections culture, santé et économie, soient les potentiels sujets du corpus. Parallèlement, le vecteur β serait de dimension égale à la taille du vocabulaire, soit le nombre de termes uniques dans le corpus. Dans les faits, la taille du vecteur α peut être donnée en entrée au modèle ; il s'agit du nombre de sujets souhaité en sortie.

Lors de l'entraînement, les paramètres du modèle sont appris de manière à optimiser la ressemblance entre les documents issus du **processus génératif** à partir de $\mathrm{Dir}(\alpha)$ et $\mathrm{Dir}(\beta)$ et ceux donnés en entrée. Étant donnés K, M et N, les nombre respectifs de sujets, documents et termes, sont sélectionnés aléatoirement θ_i, la distribution de sujets du i^{e} document, et φ_k, la distribution de termes du k^{e} sujet. Alors, les documents sont générés en sélectionnant un sujet $z_{i,j}$ à partir de θ_i et un terme $w_{i,j}$ à partir de $\varphi_{z_{i,j}}$ pour toutes positions i, j, où $i \in \{1, \ldots, \mathrm{M}\}$ et $j \in \{1, \ldots, \mathrm{N}_i\}$. Ainsi, le modèle ajuste à la fois les paramètres de $\mathrm{Dir}(\alpha)$ et $\mathrm{Dir}(\beta)$.

Pour illustrer le **processus génératif** de LDA, considérons le cas simple d'un modèle de 3 sujets entraînés à partir d'un corpus de 3 documents, contenant chacun 10 termes, et un vocabulaire de 3 termes uniques. Donc, les vecteurs de paramètres α et β ont chacun 3 dimensions, relativement au nombre de sujets souhaités et à la taille du vocabulaire. La Figure 2.8 présente une itération du **processus génératif** de LDA suivant cet exemple. À des fins de simplification, les sujets sont identifiés par différentes étoiles, puis que les termes du vocabulaires sont A, B et C. Au début du processus, les documents originaux sont donnés en entrée au modèle LDA. Au cours de l'itération, ceux-ci sont positionnés de façon aléatoire dans $\mathrm{Dir}(\alpha)$, afin d'obtenir une distribution de sujets θ par document. Parallèlement, les sujets sont positionnés de façon aléatoire dans $\mathrm{Dir}(\beta)$, afin d'obtenir une distribution de terme φ par sujet. Les différentes distributions θ et φ servent ensuite à générer des documents, tentant de reproduire les documents originaux. Avant la prochaine itération, les paramètres α et β sont ajustés de façon à améliorer le modèle.

L'inférence du modèle LDA est obtenue grâce à une **approximation bayésienne variationnelle**, soit une technique permettant de contourner les problèmes causés par les variables non observées lors de l'in-

Figure 2.8 Exemple d'une itération du processus génératif de LDA.



férence bayésienne. Pour y parvenir, un **algorithme d'espérance-maximisation** est mis en œuvre afin de maximiser la vraisemblance du modèle. Ce processus a pour but de trouver les paramètres α et β permettant de modéliser les sujets les plus probables, soit ceux pouvant reproduire au mieux les documents originaux lors du **processus génératif**. Dans un tel cas, les paramètres α et β sont les plus vraisemblables, donc ceux qui maximisent la vraisemblance du modèle. Pour plus de détails techniques, ces concepts sont décrits plus amplement dans la publication originale de Blei *et al.* (2003).

L'apprentissage de $\mathrm{Dir}(\alpha)$ et $\mathrm{Dir}(\beta)$ permet d'évaluer *a posteriori* la composition de sujets d'un document, que celui-ci ait été observé ou non à l'entraînement. Cet aspect, distinguant LDA de ses prédécesseurs, permet notamment l'utilisation de la modélisation de sujets pour diverses tâches de l'IA, y compris la classification et l'**analyse de sentiments**.

2.3.3 Approche avec plongements de mots : *Embedded Topic Model*

Récemment, la technique ETM a vu le jour, reprenant le concept de **processus génératif** de LDA dans son architecture (Dieng *et al.*, 2020). Toutefois, celle-ci remplace les distributions de Dirichlet par une approche axée sur les plongements de mots et, ainsi, prend en compte la **relation sémantique** existante entre les mots du corpus. Bien que LDA occupe une place de standard dans le domaine de la modélisation de sujets, son **processus génératif** ne parvient pas à capturer la **relation sémantique** puisque celui-ci traite chaque document du corpus comme un sac de mots, soit une représentation considérant seulement le nombre d'occurrences pour chaque mot du document, indépendemment du contexte original. Intuitivement, considérer la **relation sémantique** entre les mots d'un corpus peut mener à la modélisation de sujets plus représentatifs de la réalité, puisque ceux-ci prennent en compte cet aspect important de la langue. Les résultats présentés dans la publication de Dieng *et al.* (2020) le démontrent bien, la technique ETM se distinguant de LDA par de meilleures performances.

Dans son implémentation, la technique ETM considère un espace de dimension de L dans lequel sont plongés les termes et les sujets. Alors, les termes sont représentés par ρ, une matrice de plongements de mots de taille $\mathrm{L} \times \mathrm{N}$, soit un paramètre du modèle. Le k^{e} sujet est exprimé par un paramètre $\beta_k \in \mathbb{R}^L$, un vecteur sous forme de représentation distribuée dans l'espace de plongements, tel que présentée en Section 2.2.1. Un sujet étant représenté dans l'espace de plongements, la relation entre un terme et celui-ci est donné par le produit intérieur de leurs vecteurs, où une grande similarité se traduit par une forte probabilité d'appartenance du terme au sujet. Dans le contexte du produit intérieur, plus haute est la valeur

numérique obtenue, plus grande est la similarité.

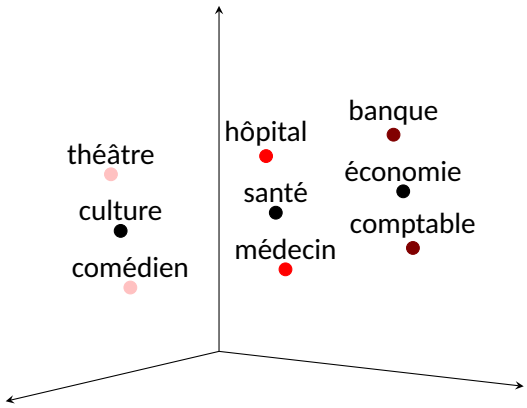
Contrairement à LDA, ETM fait appel à la **loi logit-normale** lors de son **processus génératif**. Sa distribution, dénotée $\mathcal{LN}(\cdot)$, suit la fonction logistique vue à l'Équation 2.1, de façon à ce que son logit, soit la fonction réciproque de la fonction logistique, suive la distribution normale. ETM étant basé sur le **processus génératif** de LDA, l'entraînement du modèle suit les mêmes étapes. La distribution de sujets θ_i du i^{e} document est sélectionné en suivant la **loi logit-normale**, paramétrée telle que $\mathcal{LN}(0, \mathbf{I})$, où $\mathbf{I}$ est la **matrice identité**. La distribution de termes φ_k pour le k^{e} sujet est donnée par le produit intérieur de la matrice ρ et du vecteur β_k. Alors, comme pour LDA, les documents sont générés en sélectionnant un sujet $z_{i,j}$ à partir de θ_i et un terme $w_{i,j}$ à partir de $\varphi_{z_{i,j}}$ pour toutes positions i, j, où $i \in \{1, \ldots, \mathrm{M}\}$ et $j \in \{1, \ldots, \mathrm{N}_i\}$. L'implémentation de la **loi logit-normale** et du calcul des plongements de mots lors du **processus génératif** du modèle ETM étant hors de la portée de ce mémoire, il est possible d'en obtenir les détails dans la publication originale de Dieng *et al.* (2020).

Lors de son **processus génératif**, le modèle ETM cherche à optimiser la position des sujets dans l'espace de plongements de façon à reproduire au mieux les documents originaux. Le but est donc d'ajuster les paramètres β du modèle à chaque itération. Reprenons l'exemple du journal. Si le modèle ETM est paramétré de façon à obtenir 3 sujets, il serait possible que ceux-ci s'apparentent aux sections *culture*, *santé* et *économie* du journal. La Figure 2.9 illustre un simple exemple de plongements de mots et de sujets qui pourraient être obtenu avec le modèle ETM. Les points *culture*, *santé* et *économie* seraient les 3 sujets du modèle pour l'exemple du journal, alors que les autres points seraient des mots se retrouvant dans le vocabulaire des documents originaux. En pratique, les sujets du modèle ne sont pas accompagnés d'une étiquette ; la thématique est interprétée selon la position du sujet dans l'espace de plongements.

Tout comme pour LDA, l'apprentissage du modèle ETM nécessite l'usage de l'**approximation bayésienne variationnelle**. Pour y parvenir, un algorithme d'estimation de Monte Carlo de la **borne inférieure de la preuve** est mis en œuvre afin de maximiser la log-vraisemblance marginale, mais cette fois-ci pour les paramètres β et ρ. Cette implémentation étant hors de la portée de notre étude, les détails des concepts sont disponibles dans la publication originale de Dieng *et al.* (2020).

Lors de son entraînement, le modèle ETM est en mesure d'estimer ses propres plongements de mots, c'est-à-dire de représenter les mots qu'il reçoit en entrée dans un espace vectoriel. Celui-ci peut également

Figure 2.9 Exemple de plongements de mots et de sujets produits par ETM.



utiliser des plongements pré-entraînés via d'autres techniques, notamment ceux obtenus grâce au modèle Word2Vec présenté à la Section 2.2.2, pour ensuite estimer la position de ses sujets dans cet espace vectoriel.

2.3.4 Évaluation

Étudier les performances d'un modèle de sujets sur un corpus non observé à l'entraînement est fondamental en pratique. La perplexité permet cette étude, en mesurant la capacité de généralisation d'un modèle de sujets lors de l'observation d'un corpus de test. En effet, la perplexité est étroitement liée à la vraisemblance d'observer des mots dans de nouveaux documents. Plus cette dernière est élevé, plus la perplexité du modèle est basse et meilleur est le modèle. Étant donné un corpus de test D_{test} comprenant M documents d, la log-vraisemblance est donnée par

$$\ell(D_{test}) = \sum_{i=1}^{\mathrm{M}} \log P(d_i), \tag{2.5}$$

où $P(d_i)$ est la probabilité d'observer le contenu du document d_i avec le modèle de sujets évalué. Suivant cette équation, la perplexité est définie par

$$\mathrm{PP}(D_{test}) = \exp\left\{-\frac{\ell(D_{test})}{\mathrm{N}}\right\},$$

où N est le nombre total de mots dans D_{test}.

Reprenons l'exemple du journal pour donner un cas d'évaluation concret. Un modèle de sujets a été entraîné à partir du journal J_a. Il est donc possible d'évaluer le modèle sur deux journaux concurrents, J_b et J_c, non observées à l'entraînement. Si $\mathrm{PP}(J_b) = 50$ et $\mathrm{PP}(J_c) = 220$, alors le modèle de sujets reconnaît plus

facilement les documents de J_b. En pratique, il est probable que les contenus des journaux J_a et J_b soient similaires. Par exemple, ces journaux pourraient couvrir les mêmes sections.

Pour évaluer l'interprétabilité d'un modèle de sujets, la cohérence de sujet (Mimno *et al.*, 2011) est considérée. Cette dernière additionne les résultats obtenus par une fonction de score prenant en entrée les paires des termes $w_i, \ldots, w_{\mathrm{N}}$ décrivant les sujets du modèles, telle que

$$\mathrm{TC} = \sum_{i<j} score(w_i, w_j).$$

Fréquemment, la fonction de score utilisée est l'**information mutuelle ponctuelle**, qui estime la dépendance statistique de deux variables grâce au logarithme d'un rapport de probabilités. En pratique, les n mots ayant la plus haute probabilité d'appartenance au sujet sont pris en compte. Plus les sujets d'un modèle ont une bonne cohérence de sujet, plus les sujets du modèle sont interprétables. Dans notre exemple du journal, imaginons les sujets S_x et S_y. Si les cohérences calculées sont $\mathrm{TC}_{S_x} = 0.35$ et $\mathrm{TC}_{S_y} = 0.20$, alors S_x est plus cohérent que S_y. Les n mots pour S_x devrait présenter une plus grande signification entre eux, par exemple *musique*, *critique*, *album* et *spectacle*, tandis que ceux de S_y pourraient être *médecin*, *femme*, *aide* et *traitement*, pour lesquels la dépendance semble moins marquée. De même manière, un modèle de sujets ayant une cohérence de sujet moyenne supérieure à un autre modèle est considéré comme plus interprétable.

2.4 Classification

2.4.1 Classification binaire

Le problème de la classification fait référence au domaine de la reconnaissance statistique des formes. Une forme permet de représenter un élément réel grâce à une série de D attributs mesurables le qualifiant. on dit que l'élément réel sous cette représentation est un exemple de la forme, noté $\mathscr{X}$. Reprenons l'idée du journal mentionné à la Section 2.3.1. Un article peut être représenté sous la forme d'une série d'attributs, comme son nombre de mots, d'images et de paragraphes. Alors, chaque article du journal peut être représenté sous cette même forme, ayant chacun des valeurs d'attributs différentes le caractérisant. Tous ces exemples sont alors un ensemble d'exemples, noté $\mathscr{X}$.

Pour un ensemble d'exemples $\mathscr{X}$ observé, il existe Q classes. Une classe est une catégorie à laquelle un exemple peut appartenir, en fonction d'observations faites sur ce dernier. Dans le cas des articles, il serait

Tableau 2.3 Exemple simple d'ensemble d'entraînement dans un contexte de classification binaire.

Attributs			Étiquette
nombre de mots	nombre d'images	nombre de paragraphes	est un article culturel
250	5	3	positif
750	1	6	négatif
400	4	2	positif
1000	0	8	négatif
650	1	6	négatif

possible d'avoir les classes *santé*, *culture* ou *économie*, soit les sections du journal. Les valeurs d'attributs observés pour un exemple permettent donc de le classer, soit de définir son appartenance à une classe. Par exemple, un article avec un grand nombre d'images pourrait appartenir à la classe *culture*, alors qu'un article avec un grand nombre de paragraphes pourrait appartenir à la classe *économie*. Ainsi, lors du processus de classification, une étiquette $\mathcal{Y}_i$, correspondant à une des Q classes, est attribuée un exemple $\mathcal{X}_i$. (Webb et Copsey, 2011, p. 2).

Lorsque $\mathcal{Y}_i$ est connue *a priori* pour tout $\mathcal{X}_i \in \mathscr{X}$, l'ensemble d'exemples est noté $\mathscr{X}_t$, soit un ensemble ou jeu de données d'entraînement. Ici, le problème de classification est dit de nature supervisée. Dans un tel cas, il est possible d'entraîner des classifieurs, ou fonction de classification, en considérant les couples d'exemple $\mathcal{X}_i$ et d'étiquette $\mathcal{Y}_i$ issus de $\mathscr{X}_t$. Ces classifieurs pourront être utilisés pour prédire *a posteriori* la classe d'un exemple non vu à l'entraînement.

Le problème de classification binaire est une spécification du problème de la classification, où $\mathrm{Q} = 2$. Dans un tel cas, on cherche à vérifier l'appartenance des formes à un concept (Cornuléjols et Miclet, 2018, p. 27). Alors, un exemple est dit positif s'il est une instance d'un concept donné, négatif dans le cas contraire.

Pour reprendre le cas des articles de journal, pour une tâche de classification binaire, il serait possible de considérer le concept *est un article culturel*. Dans ce contexte, un article culturel appartiendrait à la classe positive, tandis qu'un article économique appartiendrait à la classe négative. Le Tableau 2.3 présente un simple ensemble d'entraînement pour une tâche visant à classer des articles de journal. Dans le cas présent, les exemples ont $\mathrm{D} = 3$ attributs et sont accompagnés d'une étiquette dont la valeur est *positif* ou *négatif*, selon la classe binaire à laquelle l'exemple appartient.

En pratique, un exemple est manipulé sous forme vectoriel, où chaque dimension du vecteur correspond à un des attributs de l'exemple. Les attributs d'un exemple peuvent prendre des valeurs discrètes, numériques ou continues. Pour cette raison, il est possible d'effectuer une tâche de classification binaire avec des attributs obtenus grâce à la modélisation de sujets, comme la composition de sujets d'un document, représentée par un vecteur de probabilités, donc de valeurs continues.

2.4.2 Évaluation

Les performances d'un classifieur peuvent être évaluées par rapport à sa capacité à émettre correctement des prédictions. Alors qu'un classifieur prédit si un exemple est positif ou négatif, cette prédiction est dite vraie si elle correspond à l'observation externe de l'exemple, c'est-à-dire que la classe réelle de l'exemple correspond à la classe prédite, fausse autrement. Dans un contexte binaire, quatre résultats sont possibles. Les vrais positifs, dénotés VP, sont les exemples réellement positifs évalués positifs par le classifieur. De manière similaire, on obtient les faux positifs (FP), vraix négatifs (VN) et faux négatifs (FN) (Cornuéjols et Miclet, 2010, p. 58).

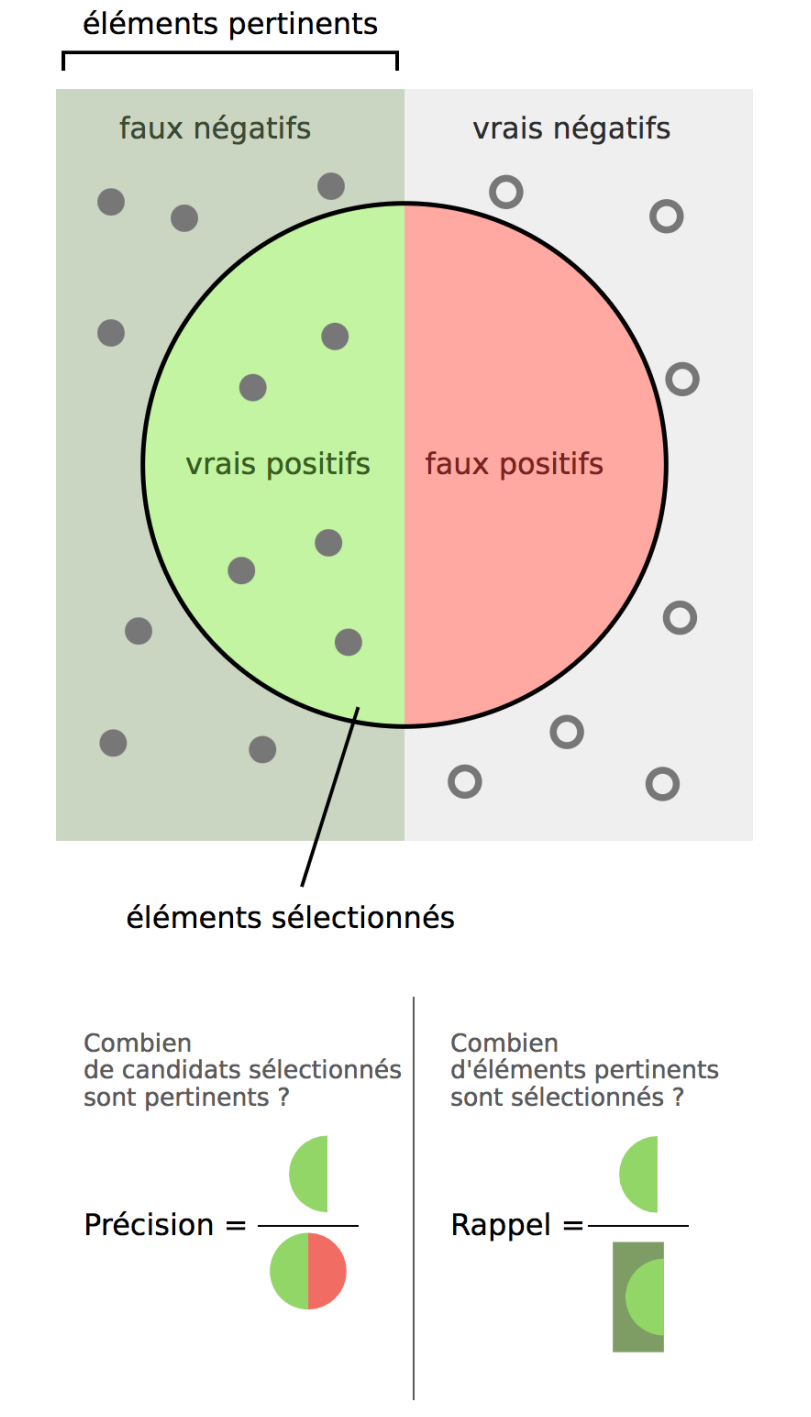
Ces observations empiriques peuvent être utilisées pour calculer les performances du classifieur. Pour évaluer adéquatement les résultats d'une tâche de classification, trois métriques sont généralement prises en compte, la précision, le rappel et la F-mesure, définies par les Équations 2.6, 2.7 et 2.8 respectivement. La Figure 2.10 illustre le calcul de la précision et du rappel, ainsi que leur lien avec les éléments positifs et négatifs. Pour certaines tâches, il est important que tous les exemples positifs soient bien classés, au risque d'introduire des faux positifs dans le classement. Dans un tel cas, la métrique d'intérêt est le rappel. Si au contraire, une tâche requiert que les exemples classés positifs soient uniquement des vrais positifs, quitte à obtenir des faux négatifs, la précision est la métrique à considérer. Pour décrire la performance générale du classifieur, la F-mesure est une métrique utile, soit la moyenne harmonique de la précision et du rappel (Powers, 2020).

$$\text{précision} = \frac{\mathrm{VP}}{\mathrm{VP} + \mathrm{FP}} \tag{2.6}$$

$$\text{rappel} = \frac{\mathrm{VP}}{\mathrm{VP} + \mathrm{FN}} \tag{2.7}$$

$$\text{F-mesure} = 2 \times \frac{\text{précision} \times \text{rappel}}{\text{précision} + \text{rappel}} \tag{2.8}$$

Figure 2.10 Illustration du calcul de la précision et du rappel (Wikimedia Commons, 2018).



Observons un exemple de calcul concret : les prédictions obtenues par un classifieur sont $\mathrm{VP} = 45$, $\mathrm{FP} = 20$, $\mathrm{VN} = 30$ et $\mathrm{FN} = 5$. Dans ce cas, la précision est $45/(45+20) = 0.69$, le rappel $45/(45+5) = 0.90$ et la F-mesure $2 \times (0.69 \times 0.90)/(0.69+0.90) = 0.78$. Ici, le classifieur est meilleur pour classer correctement les exemples positifs, mais parvient moins bien à classer les exemples pertinents à la tâche.

2.4.3 Régression logistique binaire

Pour accomplir une tâche de classification binaire, il est possible de faire appel à la régression logistique binaire. Modèle statistique répandu, la régression logistique peut être appliquée dans un contexte d'apprentissage automatique afin de créer un classifieur (Rakotomalala, 2014). Étant donné une série de valeurs $(x_1, \ldots, x_{\mathrm{D}})$, noté X, ce modèle cherche à prédire la valeur Y. Dans le cas binaire, cette dernière prend une valeur $y \in \{1, 0\}$, où 1 dénote le cas positif, négatif autrement. Concrètement, pour la classification, les valeurs de X sont données par les attributs d'un exemple $\mathcal{X}_i$. Ainsi, prédire la valeur Y revient à prédire l'étiquette $\mathcal{Y}_i$ de l'exemple $\mathcal{X}_i$.

La régression logistique base son hypothèse fondamentale à partir de la probabilité conditionnelle $P(Y = y_k|X)$. Dans le contexte binaire, la règle de décision du modèle se base sur le rapport de probabilité des deux classes. Toutefois, puisque $P(0|X)$ est égal à $1 - P(1|X)$, la règle peut être exprimée uniquement grâce à $P(1|X)$, tel que

$$\text{si } \frac{P(1|X)}{1 - P(1|X)} > 1, \text{ alors } Y = 1.$$

Pour résoudre le problème donné par ce rapport de probabilité, la technique de la régression logistique fait appel au logit, défini sur $]0, 1[$ par

$$\mathrm{logit}(x) = \ln \frac{x}{1-x}. \tag{2.9}$$

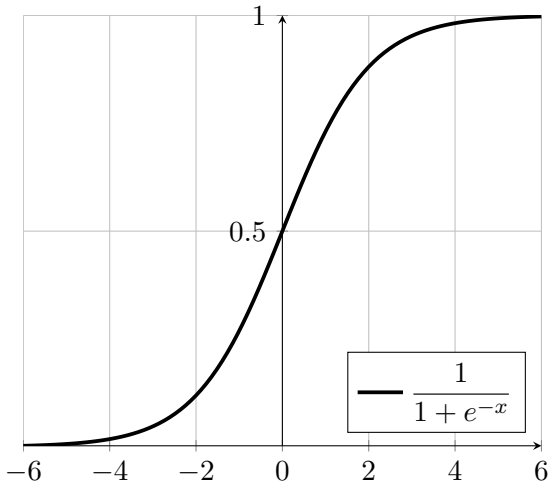
Dans le cas particulier de la régression logistique, l'intention est de calculer le logit de $P(1|X)$, soit

$$\ln \frac{P(1|X)}{1 - P(1|X)}. \tag{2.10}$$

Le logit défini en Équation 2.10 peut également être exprimé grâce aux valeurs de X et aux paramètres $\alpha = (\alpha_0, \alpha_1, \ldots, \alpha_{\mathrm{D}})$ du modèle, tel que

$$\alpha_0 + \alpha_1 x_1 + \cdots + \alpha_{\mathrm{D}} x_{\mathrm{D}}. \tag{2.11}$$

Figure 2.11 La fonction logistique.



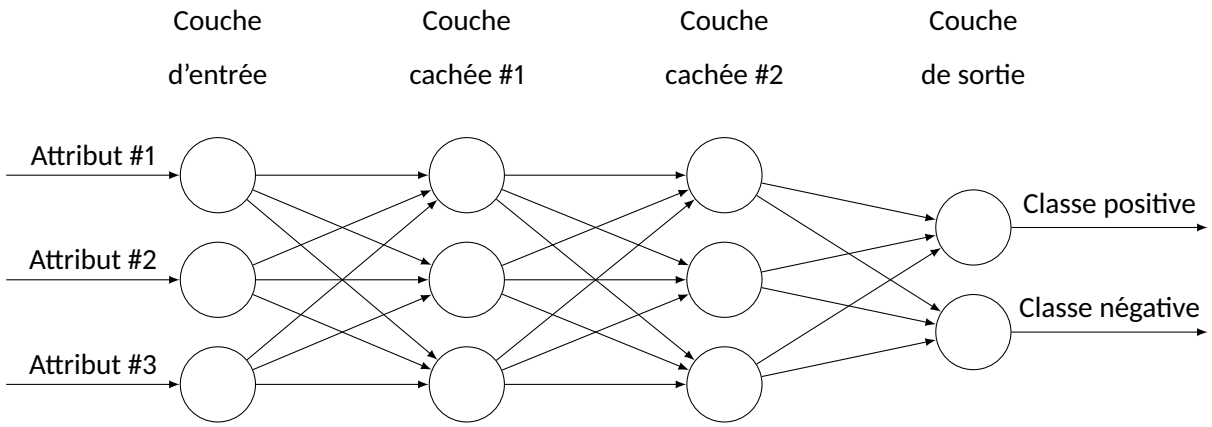
En considérant le logit défini en Équation 2.10, l'Équation 2.9 peut être transformée, révélant que $P(1|X)$ est exprimé par la fonction logistique, définie à l'Équation 2.1, où, dans ce cas particulier, x prend la valeur logit de $P(1|X)$. Étant hors de la portée de ce mémoire, les détails mathématiques permettant de passer de l'Équation 2.10 aux Équations 2.11 et 2.1 sont disponibles dans la publication de Rakotomalala (2014). Les Équations 2.10, 2.11 et 2.1 permettent alors d'exprimer la $P(1|X)$ tel que

$$P(1|X) = \frac{1}{1 + e^{-(\alpha_0 + \alpha_1 x_1 + \cdots + \alpha_{\mathrm{D}} x_{\mathrm{D}})}}.$$

De cette façon, évaluer la règle de décision du modèle nécessite seulement l'utilisation de la fonction logistique, des valeurs de X et des paramètres α du modèle. En pratique, les classifieurs basés sur la régression logistique binaires sont souvent choisis pour cette caractéristique, les rendant simples à implémenter et rapide à entraîner. Estimer la règle de décision du modèle de cette manière est possible puisque, comme illustré à la Figure 2.11, la fonction logistique a un codomaine défini sur $]0, 1[$. Ainsi, lorsque $P(1|X) = 0.99$, $Y = 1$ puisque $0.99/(1 - 0.99) = 99$, donc plus grand que 1. À l'inverse, lorsque $P(1|X) = 0.01$, $Y = 0$ puisque $0.01/(1 - 0.01) = 0.01$, donc plus petit que 1.

Lors d'un processus d'entraînement supervisé, les paramètres α du modèle sont appris grâce à un ensemble d'exemples donné $\mathscr{X}_t$. Cet apprentissage est obtenu par l'estimation du maximum de vraisemblance d'observer les exemples de $\mathscr{X}_t$, pour lequel les détails techniques sont donnés dans la publication de Rakotomalala (2014). Les paramètres α permettant de maximisant la probabilité d'observer les exemples de $\mathscr{X}_t$ sont ceux retenus par le modèle entraîné. En pratique, les paramètres sont optimisés par **descente de gradient**. La méthode la plus couramment appliquée est l'algorithme de Newton-Raphson (Tenenhaus, 2007,

Figure 2.12 Perceptron multicouche à deux couches cachées dans un contexte de classification binaire.



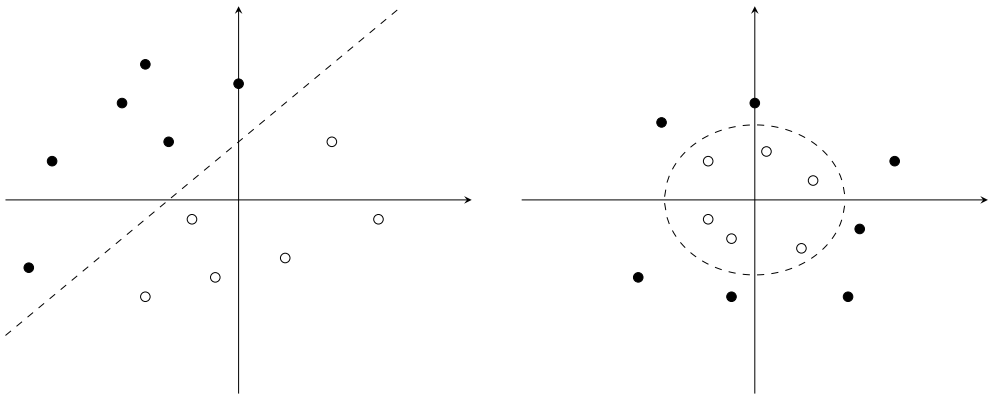
Chapitre 11), où les paramètres optimisés $\hat{\alpha}$ sont obtenus de façon itérative.

2.4.4 Perceptron multicouche

Issu des algorithmes de la famille des réseaux neuronaux discutés à la Section 2.1, le perceptron multicouche est une **méthode paramétrique** permettant de modéliser des classifieurs complexes. Différemment du perceptron classique, celui-ci se compose d'au minimum trois couches : une couche d'entrée, une ou plusieurs couches cachées et une couche de sortie (Duda *et al.*, 2000, Chapitre 6). Les couches d'entrée et cachées peuvent prendre différentes tailles. Généralement, leur taille est ajustée pour refléter la dimension des vecteurs d'exemples données en entrée. Quant à elle, la taille de la couche de sortie est ajustée au nombre de classes observées dans l'ensemble d'entraînement, de façon à pouvoir émettre la prédiction du modèle. Lors de la classification, le neurone ayant la plus haute valeur d'activation de la couche de sortie détermine l'étiquette attribuée à l'exemple considéré, soit la classe associée à ce même neurone. Ainsi, pour une tâche de classification binaire avec un ensemble d'entraînement $\mathscr{X}_t$ comme celui observé à la Table 2.3, pour lequel les exemples ont $\mathrm{D} = 3$ attributs, un perceptron multicouche aurait une couche de sortie de 2 neurones, ainsi que des couches d'entrée et cachées de 3 neurones chacune. La Figure 2.12 illustre un tel perceptron multicouche avec 2 couches cachées.

Le perceptron multicouche est préféré au perceptron classique pour sa capacité à apprendre des fonctions non-linéaires. En effet, le perceptron classique n'étant pas doté de couches cachées, ses données d'entrée sont transformées une seule fois par les **poids** du modèle, soit en passant de la couche d'entrée à la couche

Figure 2.13 Problèmes linéaire et non-linéaire.



(a) Problème linéaire (b) Problème non-linéaire

de sortie. Cette transformation est un produit scalaire, donc linéaire, ce qui ne permet pas au modèle de classer des données suivant un patron non-linéaire. Les couches cachées du perceptron multicouche permettent d'outrepasser cette limitation et d'apprendre des fonctions non-linéaires. Ceci s'avère très utile lors de la classification, selon les données considérées. Par exemple, bien que les perceptrons classique et multicouche peuvent classer les données de la Sous-figure 2.14(a), seul le perceptron multicouche peut y parvenir pour les données de la Sous-figure 2.14(b).

2.4.5 Méthode des k plus proches voisins

La méthode des k plus proches voisins est une **méthode non-paramétrique** souvent appréciée pour sa simplicité. En effet, ce modèle n'apprend pas de paramètres lors d'un entraînement. Ici, c'est une fonction de distance, préalablement choisie, qui permet d'évaluer la prédiction. Cette méthode nécessite seulement trois éléments pour émettre une prédiction : la valeur de k, c'est-à-dire le nombre de voisins à considérer lors de la prédiction, une fonction de distance et les données d'entraînement (Webb et Copsey, 2011, p. 152–154).

Pour parvenir à émettre sa prédiction, le modèle calcule la distance entre les représentations vectorielles d'un exemple à classer $\mathcal{X}$ et de tous les exemples issus d'un ensemble d'entraînement $\mathscr{X}_t$ donné. Le modèle est en mesure de choisir k voisins, soit les k exemples les plus proches de $\mathcal{X}$. La prédiction est effectuée par vote majoritaire : l'étiquette de la classe la plus observée parmi les k voisins est attribuée à $\mathcal{X}$. Par exemple, dans le cas binaire, si $k = 3$ et que deux des trois plus proches voisins de $\mathcal{X}$ sont des exemples de la classe

positive, alors $\mathcal{X}$ est prédit comme étant un exemple de la classe positive.

Lorsque les attributs des exemples sont des valeurs numériques continues de même échelle, la distance euclidienne est généralement utilisée comme distance. Elle est définie par :

$$d(\mathbf{u}, \mathbf{v}) = \sqrt{\sum_{i=1}^{\mathrm{I}} (u_i - v_i)^2}, \tag{2.12}$$

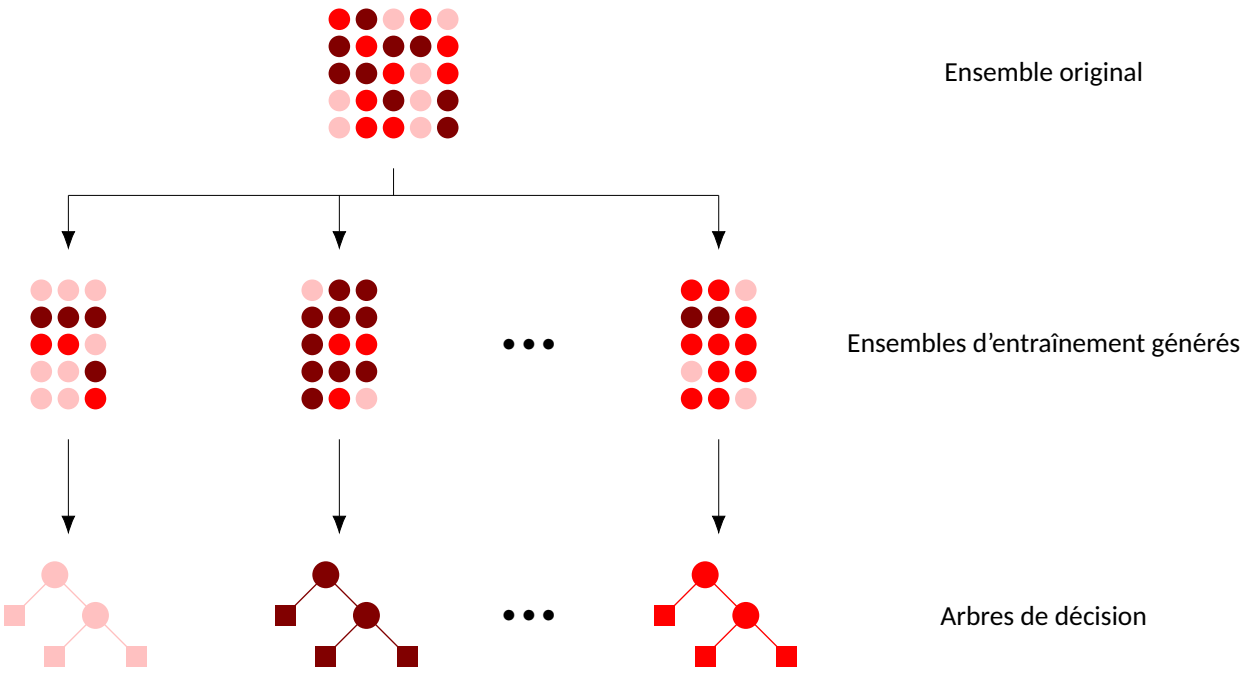
où $\mathbf{u}$ et $\mathbf{v}$ sont des vecteurs de dimension I. En pratique, la valeur de k qui optimise les performances de classification peut être obtenue par une étude empirique en utilisant la **validation croisée**, soit une méthode d'évaluation de la fiabilité basée sur l'échantillonnage. La **validation croisée** divise un ensemble d'entraînement $\mathscr{X}_t$ initial n échantillons afin de tester n fois le modèle. Un à la suite de l'autre, un des n échantillons est utilisé comme ensemble de validation et les $n-1$ échantillons restant sont utilisés comme ensemble d'entraînement. Lorsque les n tests sont complétés, la moyenne des performances obtenues indique la performance générale du modèle. De cette façon, il est possible d'ajuster les paramètres donnés en entrée à un modèle, comme la valeur de k.

2.4.6 Forêt aléatoire

La forêt aléatoire est un algorithme de classification fréquemment utilisé pour bénéficier des avantages des arbres de décision, tout en réduisant leur taux d'erreur. En effet, cette approche tire profit de la création d'un grand ensemble d'arbres de décision partiellement non corrélés. Pour y parvenir, le modèle combine la force du *bagging* à la sélection aléatoire d'attributs pour procéder à la création des différents arbres (Breiman, 2001).

La technique du *bagging*, de l'anglais ***b**ootstrap **agg**regating*, consiste à échantillonner aléatoirement avec remplacement l'ensemble de données original afin de créer un ensemble de données d'entraînement distinct pour chaque arbre de la forêt (Breiman, 1996). En d'autres termes, une donnée de l'ensemble original peut être sélectionnée plusieurs fois lors de l'échantillonnage, permettant de générer un ensemble d'entraînement différent de l'ensemble original. Le processus d'échantillonage est répété pour chaque arbre de la forêt aléatoire, afin que chacun ait un ensemble d'entraînement propre à lui. Ainsi, chaque arbre est construit à partir d'un ensemble de données différent. La Figure 2.14 illustre le processus de création des arbres de décision d'une forêt aléatoire. Chaque ensemble d'entraînement généré en échantillonnant l'ensemble original est composé de données qui influencent la structure de l'arbre de décision créé. Dans les

Figure 2.14 Processus de création d'une forêt aléatoire.



faits, la prédiction émise par un arbre de décision dépend de sa structure, plus précisément de ses noeuds.

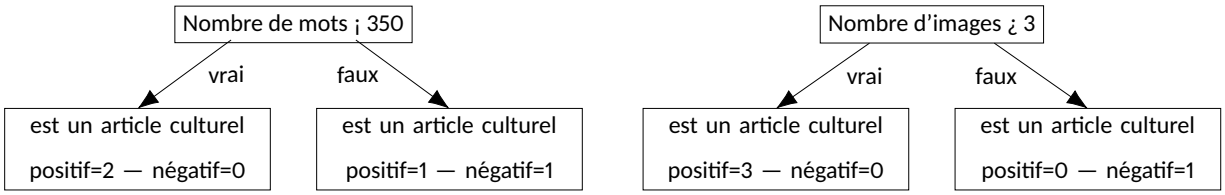
Chaque noeud d'un arbre est créé à partir d'un sous-ensemble d'attributs sélectionnés aléatoirement. Ensuite, chaque attribut du sous-ensemble est évalué de façon à choisir l'attribut divisant le mieux les données à classer. Dans la pratique, l'**indice de Gini** peut être utilisée pour faire cette évaluation (Cornuéjols et Miclet, 2010, p. 413–414), soit une mesure statistique évaluant le niveau d'inégalité de la répartition d'une variable au sein d'une population.

Considérons un exemple concret. Si l'ensemble de données original utilisé pour le processus d'entraînement de la forêt aléatoire est celui de la Table 2.3, alors il est possible d'obtenir l'ensemble d'entraînement généré observé en Table 2.4, conservant les attributs originaux *nombre de mots*, *nombre d'images* et *nombre de paragraphes*. Lors de la création d'un noeud, l'algorithme sélectionne aléatoirement un sous-ensemble de ces attributs, par exemple *nombre de mots* et *nombre d'images*. L'algorithme pourrait alors considérer les noeuds potentiels présentés en Figure 2.15. Si tel était le cas, le noeud de la Sous-figure 2.16(b) serait choisi puisqu'il divise parfaitement les données selon leur classe, tandis que le noeud de la Sous-figure 2.16(a) n'y

Tableau 2.4 Exemple d'ensemble d'entraînement généré par échantillonage aléatoire avec replacement dans le processus de création d'une forêt aléatoire.

Attributs			Étiquette
nombre de mots	nombre d'images	nombre de paragraphes	est un article culturel
250	5	3	positif
250	5	3	positif
750	1	6	négatif
400	4	2	positif

Figure 2.15 Noeuds potentiels pour différents attributs.



(a) Nombre de mots

(b) Nombre d'images

parvient pas, son cas *faux* contenant un exemple pour chacune des classes.

Ce processus de création permet de différencier les arbres de la forêt et, donc, diminuer leur corrélation. La structure de chaque arbre étant différente et la forêt contenant un très grand nombre d'arbres, le risque que la prédiction soit affectée par des erreurs individuelles est diminué. Lors de la prédiction, le modèle fait évaluer l'instance à classer par tous les arbres qui le composent. La règle de décision du modèle devient alors un vote majoritaire, c'est-à-dire que la classe prédite par le plus d'arbres est attribuée.

CHAPITRE 3
MODÈLES DE SUJETS POUR L'ÉVALUATION DE TROUBLES MENTAUX

3.1 Contexte et références

Comme vu en 1.2, la modélisation de sujets s'est avérée efficace pour diverses tâches de classification et d'analyse de sentiments. Les performances observées varient en fonction des données, mais surtout des troubles mentaux étudiées. Dans ce contexte, cette publication s'intéresse aux résultats de classification binaire obtenus pour trois troubles distincts : la dépression, l'automutilation et l'anorexie. Les données proviennent de trois tâches d'eRisk (Losada *et al.*, 2018, 2019, 2020). Celles-ci sont constituées de données textuelles produites par des personnes utilisatrices de Reddit. Chaque personne est associée à une étiquette binaire. Grâce à la modélisation de sujets, des attributs sont créés pour chacun des personnes à partir de leurs écrits. Plusieurs modèles LDA sont appris pour étudier les performances des attributs obtenus selon le vocabulaire présent dans le corpus d'entraînement. Les résultats indiquent un bon potentiel, particulièrement pour la détection de l'anorexie.

La publication a été présentée par l'auteur de ce mémoire en mai 2021 à la conférence *Canadian AI 2021, the 34th Canadian Conference on Artificial Intelligence*, à Vancouver, Canada, et est publiée dans les actes de la conférence (PubPub, Canadian Artificial Intelligence Association (CAIAC), AI2021)

L'auteur de ce mémoire a formulé l'hypothèse de recherche ainsi que conçu et implémenté les architectures présentées. Il a également choisi et mené les expériences conformément à l'état de l'art. Finalement, il a effectué la majeure partie de la rédaction de l'article.

3.2 Publication

Topic Models for Assessment of Mental Health Issues

Maxime D. Armstrong, Diego Maupomé, Marie-Jean Meurs

Université du Québec à Montréal

Abstract

In this paper, we explore topic modeling for the assessment of risk for depression, anorexia and self-harm. Using social media textual content from different datasets, we focus on Latent Dirichlet Allocation models, trained on both specific and combined corpora made from these datasets to perform risk detection. We investigate mental health vocabulary and shared topic modeling performance improvements on user classification.

Keywords : Topic modeling · Risk assessment · Social media · Sentiment analysis · Natural language processing

3.2.1 Introduction

On a global scale, mental health issues account for a significant portion of the total burden of disease, up to 13% according to recent studies (World Health Organization, 2013). For instance, in North America in 2018, 18% of Canadians seeking health care did so in relation to mental health, making it the most common reason to request for care ; the proportion being higher within those aged 15 to 34, reaching almost 60% (Statistics Canada, 2020). Among mental health issues, depression is one of the most common. In Europe, depression is a highly prevalent condition, with an estimated 33.4 million people affected (World Health Organization, 2016). In the U.S., depression reaches an annual prevalence rate of 7.8% among adults, which represent a total of 19.4 million people affected every year (Substance Abuse and Mental Health Services Administration, 2020). According to a recent study (Qin *et al.*, 2018), the prevalence rate of depression among the adult population in China is estimated as high as 38%.

Mental health issues also cover eating disorders, such as anorexia, which reach the highest overall mortality rate of any mental illness, with estimates between 10-15% (Arcelus *et al.*, 2011) ; the mortality rate associated with anorexia peaks for females aged 15-24 years old, being 12 times greater than that all other causes of death combined (Smink *et al.*, 2012). Self-harm is also an important issue, particularly among youth. In Canada, which is among the few countries where very detailed statistics are available, in 2013–2014, about 2,500 hospitalizations among youth age 10 to 17 were due to intentional self-harm, representing 1 in 4 injury hospitalizations for youth in this age group (Canadian Institute for Health Information, 2014). Since early intervention on mental health issues gives better treatment results (Arango *et al.*, 2018; Canadian Mental Health Association, 2020a), this work focuses on early assessment of risk for depression, anorexia and self-

harm with topic modeling and automatic detection from text extracted from social media. The long-term goal is to help mental health practitioners acquire better tools to assess the mental status of patients using AI-based systems. This could save precious time in the diagnostic and treatment processes.

Topic modeling has shown promising results for mental health issue detection. Resnik *et al*. (2015) demonstrated that topic models such as Latent Dirichlet Allocation (LDA) (David M. Blei, 2003) can uncover meaningful and promising latent structure within depression-related language collected from Twitter. Coppersmith *et al*. (2015) confirmed the potential of using social media content such as Twitter posts for depression and Post-Traumatic Stress Disorder (PTSD) binary classification. In addition, Zhao *et al*. (2011) showed the potential of using Twitter content for topic extraction, and Jelodar *et al*. (2019) reiterate on topic modeling and LDA capacities to uncover hidden structures related to user behavior in social media. Although previous work demonstrates the effectiveness of topic modeling for depressive-related conditions, there is a lack of horizontal application of this approach to various mental health issues, so as to compare them and assess its potential on a case-by-case basis. Furthermore, combining corpus concerning different mental health issues to perform such approach does not appear to have been tested yet, to the best of our knowledge.

The main contributions of this work are put forward on two different levels. Primarily, we perform binary classification with different specific topic models to observe their potential, and evaluate whether topic models related to mental health assessment showing a distinctive and strong vocabulary achieve better results. Moreover, we explore how shared topic models might improve risk detection of mental health issues sharing similar lexical field. Specifically, we study whether training a single topic model over corpora pertaining to different mental health issues benefits the assessment of such issues.

The rest of the paper is organized as follows. Section 3.2.2 gives a brief overview of the use of topic models for mental health assessment, introduces the data used and our approach. Section 3.2.3 presents our experimental settings and the results obtained, while Section 3.2.4 analyzes them. Finally, Section 3.2.5 proposes future work to be done.

3.2.2 Topic Detection on User-generated Textual Content

Over recent years, social media textual content has been shown to be an interesting basis of discovery of the mental status of its authors (Ive *et al.*, 2018; Merchant *et al.*, 2019). Social media users might share their feelings and circumstances with one another. As such, the topics discussed on Internet fora can help

Tableau 3.1 Summary statistics of the datasets from the eRisk Shared Tasks.

	Depression (2018 T1)		Anorexia (2019 T1)		Self-harm (2020 T1)	
	Risk	Control	Risk	Control	Risk	Control
Nb. users	214	1493	134	1153	145	618
Nb. writings	90 222	986 360	42 493	781 768	18 987	255 964
writings / user	411.6	660.7	317.1	678.0	130.9	414.2
avg. writing length (words)	27.5	22.8	37.3	21.5	31.5	21.5

in detecting at-risk persons (Maupomé et Meurs, 2018; Resnik *et al.*, 2015). By extracting topics from data concerning different mental health issues, this work aims to study how robust LDA is at separating the topics concerning each one. To that end, the experiments evaluate the impact of both shared and specific topic modeling. While specific topic modeling consists in extracting topics from a dataset associated to a sole mental health issue, shared topic modeling is performed on combined corpora, *i.e.* the combination of specific datasets into a single corpus. The data used are borrowed from the recurring eRisk shared task, broadly aimed at assessing the risk for different mental health issues from social media text.

Data. The data consist of three distinct datasets pertaining to depression (Losada *et al.*, 2018), anorexia (Losada *et al.*, 2019) and self-harm (Losada *et al.*, 2020). Each dataset is composed of the written production of English-speaking Reddit users. Each dataset contains writings from two classes of users. Risk users (*i.e.* positive users) have admitted to having been diagnosed with a mental health issue ; control users (*i.e.* negative users) have not. Table 3.1 presents the number of users and writings for each class in each dataset. For positive users, the text content extracted precedes and excludes the mention of diagnosis. All three datasets were originally split into a training set and a test set. However, these partitions show large discrepancies in class proportions and subject verbosity. For this reason, we perform instead five-fold cross-validation on the complete datasets for each task.

Model. The topics discussed in the data are extracted using the well-known LDA algorithm (David M. Blei, 2003). The per-user distribution of topics is then used to predict the risk for the relevant mental health issue. LDA posits documents as having a topic distribution sampled from a Dirichlet prior, each topic having its own distribution over words, also sampled from a Dirichlet distribution. The observation of each word in a document is therefore modeled as the result of sampling a topic from the distribution of K topics assigned to

said document and sampling the word from this topic distribution. As K fluctuates, LDA evaluates different distributions of topics to best fit all the observed documents. By inferring an LDA model, and thus the unobserved topics that resulted in the set of observed documents, LDA allows to discover the underlying structure within a collection of documents. Having fit an LDA model across the selected documents, one can then represent a user's writings as a proportion of topics. These proportions allow to map the user's writings to a prediction of the user's risk.

3.2.3 Experiments and Results

Preprocessing. A training corpus C is built from one or more of the eRisk datasets previously described. In this case, four corpora were created for further model training. The first three corpora consist of the three eRisk datasets as is, denoted as C_D for depression, C_A for anorexia and C_S for self-harm. To test shared topic modeling, the depression and self-harm datasets were combined into a sole corpus, C_{DS}, since both datasets share a similar lexical field, while the anorexia one does not. The documents from the original datasets were simply put together to create a larger corpus.

The LDA models were trained on documents obtained by preprocessing the corpora. Each user of a dataset is considered as a document in a corpus *i.e.* the user's writings are concatenated to generate a sole document. Stop words and short words (3 characters or fewer) are removed in every document. Following (Maupomé *et al.*, 2020), the remaining words are stemmed and extremes are filtered before training, *i.e.* words appearing in less than 20 documents or more than 50% of the documents are removed.

Training. Four different LDA models were tested, one for each of the training corpora. According to the nomenclature of the corpora, they are respectively denoted LDA_D, LDA_A, LDA_S and LDA_{DS}. To improve the external validity of the findings, five-fold cross-validation is performed. Each model and ensuing classifiers are trained on 80% of users and evaluated on the remaining 20%, reported results being the mean across the folds. Once an LDA model is trained, users are mapped to a vector of topics. Finally, four binary classifiers are trained on these vectors—using Logistic Regression (LR), k-Nearest Neighbors (KNN), Multi-Layer Perceptron (MLP) and Random Forests (RF). Note that for LDA_{DS}, two sets of classifiers are created, one for depression and one for self-harm. Thus, only vectors resulting from users' productions belonging to a specific dataset are exploited to train the classifiers and test the prediction of a mental health issue. This ensures that the original labels are respected, without adding "noisy" users from another dataset during the binary classification. Different training-validation splits are used to tune the hyper-parameters, namely

Tableau 3.2 Precision (P), Recall (R) and F-measure (F_m) for depression, self-harm and anorexia for LDA models.

		Classifier											
		LR			KNN			MLP			RF		
	Model	P	R	F_m	P	R	F_m	P	R	F_m	P	R	F_m
Depression	$LDA_D@25$	0.44	0.77	0.56	0.66	0.43	0.52	0.63	0.47	0.53	**0.77**	0.29	0.41
	$LDA_{DS}@25$	0.51	**0.78**	**0.61**	0.65	0.45	0.52	0.64	0.52	0.57	0.74	0.33	0.45
Self-harm	$LDA_S@19$	0.60	**0.76**	0.67	0.71	0.61	0.65	0.80	0.59	**0.68**	**0.86**	0.40	0.54
	$LDA_{DS}@25$	0.20	0.53	0.29	0.26	0.08	0.12	0.29	0.02	0.04	0.10	0.01	0.01
Anorexia	$LDA_A@18$	0.71	**0.87**	0.78	**0.89**	0.70	0.78	0.88	0.72	**0.79**	0.88	0.58	0.70

Tableau 3.3 Vocabulary of the most evocative mental health topic for depression, anorexia and self-harm.

	Top 10 stems (descending order of probability)
$LDA_{DS}@25$	depress-, women, relationship, doctor, pain, medic-, anxieti-, mental, okay, husband
$LDA_A@18$	weight, calori-, disord-, healthi-, doctor, gain, bing, relationship, diet, depress-
$LDA_S@19$	parent, famili-, depress, pain, women, relationship, self, mental, felt, situat-

the number of topics extracted. The number of topics varies from 10 to 30, as such range showed promising results for the assessment of mental health issues (Maupomé et Meurs, 2018; Maupomé *et al.*, 2020).

Evaluation. F-measure (F_m) is used as the main metric to evaluate the results. F_m is the harmonic mean of precision (P) and recall (R) of the positive class, as it is the most difficult one to assess for the task. This allows to balance a retrieval of positive observations that is both exhaustive and discerning. For each corpus, the number of topics achieving the best F_m is reported.

Results. The results for the best model for each corpus are presented in Table 3.2 where $LDA_C@\mathrm{K}$ denotes the LDA model with K topics trained on corpus C. Overall, the best results were obtained for the classification of anorexia, followed by self-harm then depression. For depression, $LDA_{DS}@25$ reached a maximal F_m of 0.61 with an LR classifier. For anorexia, $LDA_A@18$ hit the highest F_m of 0.79 with an MLP classifier. For self-harm, $LDA_S@19$ got to a F_m of 0.68 with a MLP classifier. In addition, for the three best models, the topic with the most evocative mental health vocabulary was extracted from the generated topics. The first 10 stems—showing the highest probability within the topic—are displayed in Table 3.3, in descending

order of probability.

3.2.4 Discussion

The results obtained uphold our initial hypotheses. Some mental health issues appear to show strong and distinctive vocabulary, such as the examples of depression and anorexia presented in Table 3.3, which allows good performance in classification with topic modeling. Using larger datasets for anorexia and depression resulted in human readable topics, contrary to self-harm. The top 10 stems for the most evocative topic for $LDA_A@18$ express significant vocabulary, and appear to be highly related to anorexia, while those for $LDA_S@19$ seem less distinctive to self-harm, including stems like *parent*, *famili-* and *relationship*. Similarly, the $LDA_{DS}@25$ top stems appear also less distinctive than the $LDA_A@18$ ones despite the corpora size, including stems like *women*, *relationship* and *okay*.

In addition, the model trained with combined datasets, which presented similar lexical content, appears to be more effective in some cases. The experiments showed that LDA_{DS} gave better results for depression classification than LDA_D, suggesting that extending corpus with similar mental health-related lexical content could help improve classification. However, tests also demonstrate that LDA_{DS} underperformed for self-harm classification. This may be due to the imbalance between the depression and self-harm datasets, diluting self-harm-specific vocabulary and resulting in weaker classification performance.

It should be noted that the self-harm dataset spans a vocabulary much smaller than the other datasets, even after stemming. At 96k word types before preprocessing, the vocabulary is less than half the size of the vocabularies of the depression set (274k word types) and the anorexia set (238k word types). This also stands for word occurrences, since the self-harm dataset neighbours 6M words, while their counts is much higher for the depression set (25M) and the anorexia set (18M). Furthermore, the vocabulary usage seems less spread out in the self-harm dataset. Combining the depression and self-harm datasets results in a vocabulary with a distribution much more similar to that of the depression dataset. This is illustrated in Figure 3.1. The systems shown better results in comparison to those obtained during the eRisk competition as reported in Table 3.4. It is to be noted though that the splits used for this task were slightly different from those in the previous eRisk shared tasks. Both LDA_D and LDA_{DS} models score better results for depression classification than the team proposing a topic model for eRisk2018 T1 (Losada *et al.*, 2018), which obtained a F_m of 0.42 at the time. With a F_m of 0.79, the proposed LDA_A model surpasses every model proposed for eRisk2019 T1 (Losada *et al.*, 2019), for which the best F_m was 0.71.

Figure 3.1 Distribution of frequency to rank in the vocabularies of the Depression and Self-harm datasets as well as the combined dataset. Frequency is in total number of occurrences. Both axes are in logarithmic scale.

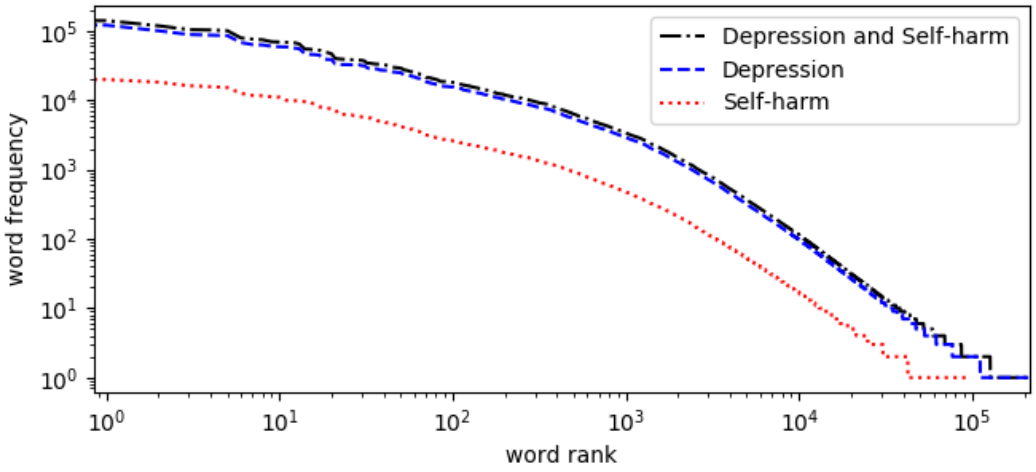


Tableau 3.4 Comparative F-measure (F_m) results between eRisk shared tasks and our best models. TM and BM refer respectively to topic model and best model.

	Model	F_m
Depression	eRisk 2018 TM	0.42
	$LDA_D@25$	0.56
	$LDA_{DS}@25$	0.61
Anorexia	eRisk 2019 BM	0.71
	$LDA_A@19$	0.79

3.2.5 Conclusion and Future Work

Exploring both shared and specific mental health-related topic models, various LDA models were tested to assess automatic detection of depression, anorexia and self-harm. The most evocative topic was extracted for each of the best models to investigate vocabulary distinctiveness, and analyze the classification performance. The reported results confirm the potential of topic modeling for early assessment of mental health issues with distinctive vocabulary. Future work could include training LDA models on combined corpora made from balanced datasets. Also, diverse topic models could be tested with the same corpora, such as Supervised LDA and Supervised Anchor, which showed promising results for depression-related language (Resnik *et al.*, 2015). In addition, an online LDA with infinite vocabulary approach as suggested by (Zhai et Boyd-Graber, 2013) could be interesting for topic models trained on users written online contributions in social media.

Reproducibility. The source code of the proposed systems is licensed under the GNU GPLv3. The datasets are provided on demand by the eRisk organizers.

Acknowledgments. This research was enabled in part by support provided by Calcul Québec and Compute Canada. MJM acknowledges the support of the Natural Sciences and Engineering Research Council of Canada [NSERC Grant number 06487-2017] and the Government of Canada's New Frontiers in Research Fund (NFRF), [NFRFE-2018-00484].

CHAPITRE 4

MODÉLISATION DE SUJETS DANS LES ESPACES DE PLONGEMENT POUR L'ÉVALUATION DE LA DÉPRESSION

4.1 Contexte et références

La similarité sémantique étant prise en compte par les plongements de mots, il est envisageable de les mettre à profit pour détecter des troubles de santé mentale comme la dépression. La sous-section 1.2 présente des approches prometteuses faisant usage de la modélisation de sujets, particulièrement avec LDA, et des plongements de mots. Toutefois, les systèmes jumelants ces deux techniques sont rarement vus dans l'état de l'art. Dans ce contexte, cette publication s'intéresse aux performances du nouveau modèle ETM au sein d'une tâche de classification binaire. Le modèle LDA est utilisé comme base de référence afin de comparer les perfomances obtenues. La tâche se concentre sur la détection de la dépression à partir des données d'eRisk (Losada *et al.*, 2018). Les résultats obtenus suggèrent une amélioration des performances de classification lors de l'utilisation des attributs obtenus grâce au modèle ETM.

La publication a été présentée par l'auteur de ce mémoire en mai 2021 à la conférence *Canadian AI 2021, the 34th Canadian Conference on Artificial Intelligence*, à Vancouver, Canada, et est publiée dans les actes de la conférence (PubPub, Canadian Artificial Intelligence Association (CAIAC), AI2021)

L'auteur de ce mémoire a formulé l'hypothèse de recherche ainsi que conçu et implémenté les architectures présentées. Il a également choisi et mené les expériences conformément à l'état de l'art. Finalement, il a effectué la majeure partie de la rédaction de l'article.

4.2 Publication

Topic Modeling in Embedding Spaces for Depression Assessment

Maxime D. Armstrong, Diego Maupomé, Marie-Jean Meurs

Université du Québec à Montréal

Abstract

This paper presents an investigation of topic modeling in embedding spaces performances in the context of depression assessment. Using the textual content of social media users from the eRisk 2018 dataset, a classification task is performed employing features generated from the Embedded Topic Model. To set contrast with traditional topic modeling, a full comparison with the Latent Dirichlet Allocation model is shown. An extensive range of topics and different preprocessing strategies are studied to demonstrate the efficiency of the models. Our results show a noteworthy improvement in the explored task from the application of the novel topic modeling approach.

Keywords : Topic modeling · Word embeddings · Depression assessment

4.2.1 Introduction

Depression is one of the most pervasive mental disorders, with over 264 million people affected worldwide. (James *et al.*, 2018). Looking more closely at the Canadian situation, 8% of adults will experience major depression in their lifetime. (Canadian Mental Health Association, 2020b). In 2018, nearly one out five health care requests was related to mental health, making it the most common reason for the population to seek assistance (Statistics Canada, 2020). Nevertheless, the stigma encircling mental disorders often leads to treatment avoidance and delays to care, resulting in deficient support and absence of proper expert diagnosis (Henderson *et al.*, 2013). As better treatments result from early intervention (Arango *et al.*, 2018; Canadian Mental Health Association, 2020a), our efforts focus on automatic depression assessment using textual content from social media. Systems using such AI-based methods could embodied better tools to evaluate the mental status of a patient, which could be helpful and time-saving for mental health practitioners. Since understanding an individual's mental condition is essential to the selection of a treatment, tools implementing topic modeling methods would be an asset for the practitioners' interpretation of the results.

The analysis of textual content from social media has proven to be a promising avenue for the automatic assessment of mental health issues, demonstrating efficiency for prediction and classification tasks surrounding depression, suicidality and anxiety disorders (Ive *et al.*, 2018; Merchant *et al.*, 2019; Coppersmith *et al.*, 2015; Shen et Rudzicz, 2017). Topic modeling also displayed considerable potential for depression-related tasks, including classification. Previous work (Resnik *et al.*, 2015; Maupomé et Meurs, 2018) emphasizes the

usage of LDA (Blei *et al.*, 2003) to perform topic extraction in this context, allowing to produce discriminant features and interpretable distributions of words from the induced topics. Word embeddings also demonstrated noteworthy performances on sentiment analysis tasks, such as classifying emotions from texts (Giatsoglou *et al.*, 2017) and evaluating the feelings associated to Twitter posts (Tang *et al.*, 2014). Since both topic modeling and word embeddings methods reached noticeable performances, systems combining them, such as the ETM (Dieng *et al.*, 2020), might improve the results obtained in depression assessment tasks.

Topic modeling in embedding spaces (Dieng *et al.*, 2020) is hence studied in this work, comparing the traditional and the novel topic modeling approaches in the context of depression risk assessment. A classification task is conducted to evaluate the quantitative performances of the models. The ability to extract discriminant topics related to depression is then investigated. The next Section presents the proposed models and the dataset considered for our experiments. Section 4.2.3 describes the performed experiments. Section 4.2.4 analyzes the obtained results while Section 4.2.5 concludes and proposes some directions for future investigations.

4.2.2 Models and Dataset

Models. To demonstrate comparison between both topic modeling approaches, the well-known LDA (Blei *et al.*, 2003), previously used in (Maupomé et Meurs, 2018), is selected as our baseline. LDA uses Dirichlet distributions within a probabilistic generative process to posit K topics from a corpus, each one being its own distribution over words. Every document within the corpus is assumed to be a mixture of the extracted topics, the proportion of which depends on the words observed in the said document.

The ETM (Dieng *et al.*, 2020) employs a similar generative process, however using word embeddings to do so. Contrary to the traditional modeling approach, the topics evaluated by the ETM are expressed as embeddings, *i.e.* vectors, belonging to the embedding space, where the vocabulary is represented. Through its generative process, the ETM computes the distribution of topics over words, where the probability of a word pertained to a topic is estimated across the difference between both of their embeddings.

In both cases, the inferred models evaluate the underlying structure within a corpus, allowing to represent each of the documents as a proportion of topics. Thus, some topics could be discriminant features in tasks such as classification, facilitating the interpretation of results, since their distribution of words over the

Tableau 4.1 Summary statistics of the dataset from the eRisk2018 shared task.

	RISK	CONTROL
# users	214	1493
# writings	90 222	986 360
writings / user	411.6	660.7
avg. words / writing	27.5	22.8

vocabulary is understandable to humans. Here, a document being the collection of a user's writings, as detailed hereafter, these proportions are then exploited to predict the user's depression risk.

Dataset. The dataset considered for our experiments is selected from the eRisk2018 shared task (Losada *et al.*, 2018), which comprises the written production of English-speaking Reddit users. It is divided into two classes of users : RISK (positive) and CONTROL (negative). The users included in the RISK class have admitted to having been diagnosed with depression ; CONTROL users have not. The users' writings, which are posts or comments on the original website, are extracted from various threads of discussion. It is to be noted that only users are labeled, not their individual writings.

The original dataset was split into a training set and a test set, which showed several imbalances in the class distribution. Thus, both subsets have been combined to form a sole dataset, for which the statistics are displayed in Table 4.1. For the purpose of our task, a custom balanced split is created, dividing the unified dataset into a training set (80%), a validation set (10%) and a test set (10%). In doing so, 1365 documents are retained for training, while the validation and test sets kept respectively 171 documents.

4.2.3 Experiments and Results

Model configurations. The ETM variant using *Pre-fitted Word Embeddings* (PWE) is selected, as its interpretability and predictive power obtained the best performances in (Dieng *et al.*, 2020). Since our corpus neighbours 25M words, from which the rare occurences are removed for training, the CBOW algorithm of Word2Vec (Mikolov *et al.*, 2013b) is chosen to generate the PWE as it tends to smooth better bigger datasets with frequent words (Mikolov *et al.*, 2013a). The ETM retains the default parameters put forward in (Dieng *et al.*, 2020), setting the learning rate with the Adam optimizer (Kingma et Ba, 2014) and selecting a ReLU activation function. In parallel, the LDA is configured to perform 50 passes through the corpus during training, with its log perplexity evaluated every time.

Preprocessing & Training. Two preprocessing strategies were applied on the dataset. First is a regular one, *i.e.* a word undergoes no alteration after tokenization, as adopted in (Dieng *et al.*, 2020). Following (Maupomé *et al.*, 2020), a stemming strategy is also selected, where a word is stemmed after tokenization. Models trained with the latest one are denoted MODEL-stem, MODEL-reg otherwise. It is to be noted that each user of the dataset is viewed as a sole document for our experiments, where a user's writtings are concatenated to create the latter.

To train a PWE, a preprocessed corpus is generated after the application of a strategy, in which the documents are divided in smaller chunks of 200 words for computational efficiency. Two independent PWE are fitted, one per strategy, ensuring the presence of every token in the embedding spaces during the training of the ETM models. For the training of the topic models, the stop words and short words (3 characters or fewer) are removed in every document prior to tokenization. Also, extremes are filtered, removing words appearing in less than 20 documents or more than 50% of the documents.

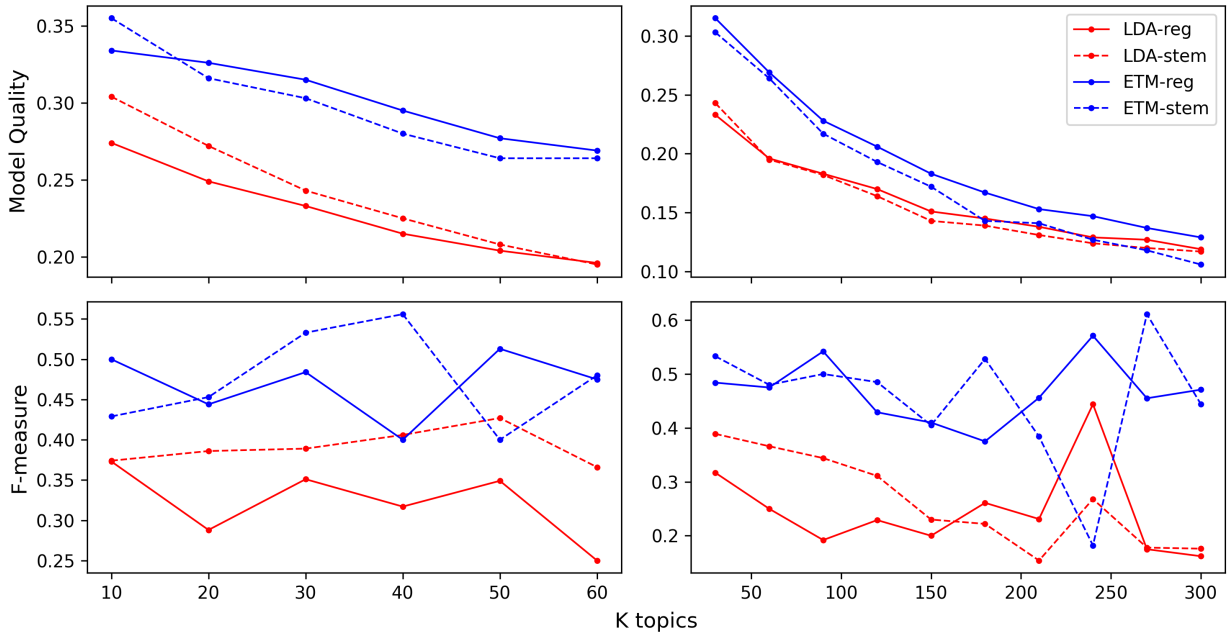
Several numbers of topics have demonstrated efficiency for depression assessment tasks, most of the time ranging between 10 and 50 topics (Resnik *et al.*, 2015; Maupomé et Meurs, 2018; Maupomé *et al.*, 2020). Alongside, the ETM has shown promising results in topic extraction and interpretability using larger numbers, such as 300 topics. Following these ideas, our models are trained on numbers of topics (K) ranging from 10 to 60 in steps of 10, and from 60 to 300 in steps of 30. After completing the model training, each document is used to generate a vector of topics, representing the topic probabilities related to every user. Following (Maupomé et Meurs, 2018), a multilayer perceptron classifier formed of 3 hidden layers is trained on these vectors. The size of each hidden layer is set to 300 neurons for every model.

Evaluation & Metrics. Beyond goodness of fit, an important aspect of topic models is the interpretability of the extracted topics. A few quantitative measures of the semantic soundness of topic models have been proposed. Topics being probability distributions over the entire vocabulary, these measures often focus on the most likely words of each topic. **Topic Coherence** (TC) (Mimno *et al.*, 2011), also known as UMass Coherence, measures the conditional log probability of the M most likely words in a topic. In a complementary manner, **Topic Diversity** (TD) (Dieng *et al.*, 2020) measures the overlap among the M most likely words across topics. Specifically, TD is computed as the proportion of the unique occurences found unique among such words of every topic. In order to balance these two measures, **Model Quality** (MQ) is defined as the product of TC and TD. Undermentioned TC and TD are denoted with a $@M$ mention, indicating the M most

likely words used to calculate these. Following (Losada *et al.*, 2018), the standard metrics of Precision (P), Recall (R) and F-measure (F_m) ($\beta = 1$) are used for classification.

Results. The results obtained show ETM outperforming LDA both in terms of quality and predictive power. Moreover, ETM better sustain both of these aspects over smaller and larger values of K. Figure 4.1 displays a close view of the results gained from the models trained to extract smaller counts of topics, as well as the longer trend from 30 to 300 topics.

Figure 4.1 Evolution of Model Quality (TC@10, TD@25) & F-measure according to K topics.



The ETM-stem model holds the best position for all but one of the proposed metrics, reaching a MQ of .355 and a F_m of .611. The leading P of 0.944 is detained by the LDA-stem model. However, the ETM-stem model is also relatively competent at this level, its highest P being .889. The foremost performances of this model were achieved with four different K topics, which are 10, 50, 150 and 270. Table 4.2 shows a comparison between the models at each of these values of K.

While the split used in our experiments is different from the original eRisk one, the ETM performances are noteworthy compared to the results obtained with LDA in (Maupomé et Meurs, 2018). Table 4.3 displays the contrast between our LDA-stem and ETM-stem models, holding the best metrics within our experiments, and the previously proposed system.

Tableau 4.2 Results for ETM and LDA models for K = 10, 50, 150 & 270.

Model	K	TC@10	TD@25	MQ	P	R	F_m
LDA-reg	10	.306	.896	.274	.268	.611	.373
	50	.273	.746	.204	.221	.833	.349
	150	.266	.566	.151	.116	.722	.200
	270	.252	.505	.127	.101	.667	.175
LDA-stem	10	.334	.908	.304	.233	**.944**	.374
	50	.291	.714	.208	.281	.889	.427
	150	.268	.533	.143	.137	.722	.230
	270	.266	.450	.120	.103	.667	.178
ETM-reg	10	.343	**.972**	.334	.500	.500	.500
	50	.291	.714	.208	.281	.889	.427
	150	.319	.575	.183	.267	.889	.410
	270	.317	.434	.137	.385	.556	.455
ETM-stem	10	**.365**	**.972**	**.355**	.600	.333	.429
	50	.339	.781	.264	**.714**	.278	.400
	150	.324	.530	.172	.262	.889	.405
	270	.320	.369	.118	.611	.611	**.611**

Tableau 4.3 Comparative results between the LDA model from Maupomé et Meurs (2018), LDA-stem and ETM-stem. For our models, only those with the best F_m are displayed.

Model	K	P	R	F_m
UQAMA (Maupomé et Meurs, 2018)	30	.32	.62	.42
LDA-stem	50	.281	.889	.427
ETM-stem	270	.611	.611	.611

4.2.4 Discussion

In view of the results, the models ability to extract topics as astute depression-related features is investigated. Mann-Whitney U test is retained to evaluate the most discriminant topics for each model, for its capacity to assess samples in a binary non-parametric context (Nachar *et al.*, 2008). To perform the test, the per-topic log probability distributions of the positive and negative users of the training set are selected. Table 4.4 presents the most discriminant topics gained from the experiment. The tokens displayed sug-

Tableau 4.4 Top tokens from the most discriminant topics containing the token *depression* (-reg) or *depress-* (-stem) among its 10 first ones. R stands for the topic rank. K = 270.

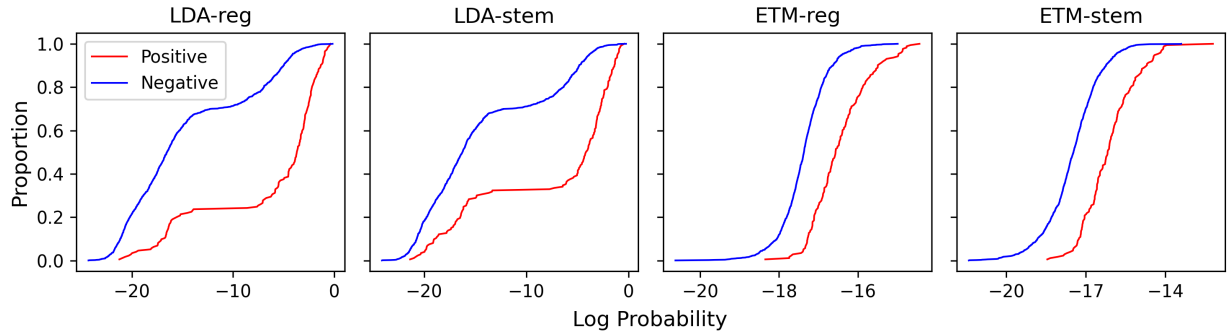
Model	R	Tokens
LDA-reg	1	depression, pain, anxiety, mental, relationship, doctor, health, therapy, haha, depressed
LDA-stem	1	depress-, emot-, anxieti-, mental-, attract-, medic-, sexual-, disord-, suffer-, neg-
ETM-reg	4	mental, brain, depression, comfortable, physical, pain, fitness, disabled, dont, injury
ETM-stem	1	weight-, depress-, reward-, downvot-, crit-, meta-, miser-, quit-, somewhat-, neg-

gest that the models are indeed able to extract an underlying structure related to depression, which could lead to the creation of shrewd features for classification. Further examination is carried out by analyzing the aforesaid per-topic log probability distributions for the selected discriminant topics, as shown in Figure 4.2. The curves observed for the topics extracted by the ETM show a distinct and concise evolution of the distribution, with similarities to a normal one. A closer look of the ETM-stem graph reveals a significant separation around the log probability -17, where about 80% of the negative users are under or reaching this value, versus 20% for the positive users. On the opposite, the topics obtained by the LDA models fail to attribute clear and differentiating log probabilities to either of the positive or negative users. The LDA-stem curves plateau around log probabilities -15 to -7, indicating that the users find themselves largely before or after these points, without a notable division. These findings highlight the ETM aptness to produce astute features, which could explain its predictive power for classification. Furthermore, this suggests that the vocabulary dividing positive users from negative users includes words that are further removed from a lexicon typically associated with depression, as suggested in the Table 4.4. Such information could prove to be a relevant addition when interpreting the obtained results, particularly when an individual's mental condition is assessed by a mental health professional.

4.2.5 Conclusion

The ETM and LDA models were tested on a classification task related to depression assessment. The most performing model, ETM-stem, present significant results both in quality and prediction. The features produced from the topics extracted by the ETM models exhibit a notable dividing power, being able to better differentiate between positive and negative users than the LDA models. Overall, the stemming strategy holds the best position with the ETM-stem and LDA-stem holding the six leading metrics scores. Future work may include feature selection, based on their precision and recall capabilities. Also, in-depth work on

Figure 4.2 Empirical distribution function of the most discriminating depression-related topic extracted per model. The y-axis shows the proportion of users for which the log probability for the selected topic is less than or equal to the value in x. K = 270.



vocabulary analysis to better distinguish between the two types of users could be conducted.

Reproducibility. The source code of the proposed systems is licensed under the GNU GPLv3. The datasets are provided on demand by the eRisk organizers.

Acknowledgements. The authors would like to thank Fanny Rancourt for useful discussions and suggestions. This research was enabled in part by support provided by Calcul Québec and Compute Canada. MJM acknowledges the support of the Natural Sciences and Engineering Research Council of Canada [NSERC Grant number 06487-2017] and the Government of Canada's New Frontiers in Research Fund (NFRF), [NFRFE-2018-00484].

CONCLUSION

Tout au long de ces travaux, différentes méthodes de modélisation de sujets ont été évaluées. L'exercice a démontré que les attributs ainsi obtenus sont pertinents pour procéder à la détection automatique de possibles troubles mentaux. Notamment, l'étude comparative du Chapitre 3 a mis en lumière le comportement des attributs extraits en fonction du vocabulaire observé. Bien que les résultats obtenus soient très différents pour chacun des troubles mentaux étudiés, ceux-ci indiquent un bon potentiel en présence d'un vocabulaire distinctif. Particulièrement, les performances obtenues pour la détection de l'anorexie indique que la modélisation de sujets est une avenue encourageante, celles-ci dépassant les performances de toutes les approches proposées dans le cadre l'édition 2019 de la campagne d'évaluation eRisk. Parallèlement, les résultats observés suite à l'entraînement de modèles de sujets sur des corpus combinés sont prometteurs. Par exemple, regrouper le contenu textuel lié à la dépression et à l'auto-mutilation sous un seul corpus d'entraînement semble mener des modèles de sujets donnant des attributs de meilleure qualité pour la détection de la dépression. Bien que les performances de la détection de l'automutilation soient modestes, les sujets obtenus sont évocateurs et, donc, pertinents lors de l'interprétation des résultats par les professionnel.le.s de la santé mentale. Les caractéristiques linguistiques affichées dans ces sujets pourraient mener à une meilleure compréhension de la situation d'un individu, malgré une éventuelle prédiction erronée lors de la classification. La compréhension de la situation globale d'un individu peut également soutenir le travail des professionnel.le.s de la santé mentale, notamment lors de l'établissement d'un diagnostic. Une prochaine étape de travail serait donc d'approfondir l'étude des caractéristiques linguistiques distinctives associées aux troubles mentaux. Une meilleure compréhension de celles-ci pourrait être bénéfique pour l'apprentissage de modèles de sujets en entraînement supervisé, comme ceux suggérés par Resnik *et al.* (2015), pouvant ainsi mener à des sujets plus interprétables et performants dans des tâches de classification.

De plus, les expériences décrites au Chapitre 4 indiquent que la prise en compte de la sémantique dans le processus de modélisation de sujets est pertinente pour la détection de la dépression. En capturant le contexte, les attributs obtenus sont potentiellement plus discriminants, ce qui constitue un avantage notoire dans le cadre d'une tâche de classification. En même temps, les sujets obtenus grâce à ETM sont évocateurs et différents de ceux obtenus avec LDA. Utiliser les deux approches en parallèle dans une application concrète, par exemple la production de rapports d'analyse destinés à aider les professionnel.le.s de la santé mentale, pourrait être bénéfique. Comme les sujets obtenus pour chaque approche comportent

des caractéristiques linguistiques différentes, leur interprétabilité pourrait être complémentaire, fournissant ainsi un portrait de la situation plus large aux professionnel.le.s concerné.e.s. Les prochains travaux pourraient comprendre la comparaison d'autres techniques de modélisation de sujets intégrant les plongements de mots, tel que lda2vec (Moody, 2016), afin de détecter le risque de troubles mentaux. Aussi, étendre l'étude à plusieurs autre troubles mentaux serait pertinent. Par exemple, les résultats atteints pour la détection de l'anorexie au Chapitre 3 étant très prometteurs, il serait intéressant d'approfondir l'analyse grâce à une technique tel que ETM.

Les observations faites ouvrent la porte à plusieurs autres expériences. Particulièrement, utiliser ces approches avec des données cliniquement validées permettrait d'évaluer le potentiel de mise en application dans la vie réelle. Ces réponses pourraient mener à la conception de meilleurs outils pour soutenir le travail des professionnel.le.s de la santé mentale. Aussi, étudier les nouveaux modèles de sujets intégrant les plongements de mots dans un tel contexte pourrait être grandement bénéfique. Puisqu'ils considèrent la sémantique des mots, ces modèles pourraient permettre une meilleure interprétabilité des résultats, facilitant le travail des professionnel.le.s. Ainsi, une meilleure application et compréhension de la modélisation de sujets dans un contexte clinique favoriserait une meilleure détection de signes de troubles mentaux et aiderait les praticien.ne.s lors de la prise de décision.

GLOSSAIRE

algorithme d'espérance-maximisation algorithme itératif permettant de trouver les paramètres du maximum de vraisemblance d'un modèle probabiliste lorsque ce dernier dépend de variables latentes non observables (Dempster *et al.*, 1977). 18, 21

analyse de sentiments domaine du Traitement Automatique du Langage Naturel s'intéressant à la reconnaissance de sentiments au sein de données textuelles. 1, 11, 21

approximation bayésienne variationnelle méthode permettant de calculer une approximation de la probabilité a posteriori de variables non observées (Webb et Copsey, 2011, p. 126). 19, 22

biais quantité ajoutée au produit des entrées et des poids lors de l'activation d'un neurone afin d'en décaler le résultat. 8, 10

borne inférieure de la preuve quantité à optimiser afin de minimiser la divergence de deux distributions donnés dans le contexte des méthodes bayésiennes variationnelles (Webb et Copsey, 2011, p. 126–128). 22

descente de gradient technique d'optimisation itérative visant à trouver le minimum d'une fonction (Cauchy *et al.*, 1847). 10, 13, 29

indice de Gini mesure statistique évaluant le niveau d'inégalité de la répartition d'une variable au sein d'une population (Gini, 1912). 33

information mutuelle ponctuelle mesure de dépendance statistique entre deux variables aléatoires (Church et Hanks, 1990). 24

loi logit-normale loi de probabilité pour laquelle la fonction logistique suit la loi normale. 22

loi de Dirichlet loi de probabilité à variables aléatoires multinomiale pour laquelle la distribution est paramétrée par un vecteur de nombres réels positifs. 18

matrice identité matrice carrée pour laquelle les éléments de la diagonale principale sont égaux à 1 et tout autre élément est égal à 0. 22

méthode non-paramétrique méthode ne supposant pas la forme de la fonction de classification et étant capable d'estimer cette dernière peu importe sa forme. 31

méthode paramétrique méthode supposant la forme de la fonction de classification et par laquelle les paramètres de cette dernière sont estimés dans le processus d'apprentissage. 30

poids paramètre associé à un neurone d'un réseau neuronal ayant pour but de pondérer les données d'entrée. 8, 10, 17, 30

probabilité conjointe mesure statistique calculant la probabilité que deux événements se produisent ensemble et au même moment. 18

processus génératif processus par lequel des données sont générés en échantillonnant la distribution des entrées ou des sorties. 17, 18, 19, 21, 22

relation sémantique relation expliquant le sens des mots et les liens existant entre eux. 11, 12, 14, 21

réseau neuronal à deux couches réseau neuronal comportant seulement une couche cachée et une couche de sortie en plus de sa couche d'entrée. 8, 14

rétropropagation algorithme facilitant la descente de gradient en calculant les gradients des paramètres en sens inverse de l'ordre d'exécution (Rumelhart *et al.*, 1986). 10, 13

validation croisée méthode d'évaluation de la fiabilité basée sur l'échantillonnage qui utilise différentes parties des données pour tester et entraîner un modèle sur différentes itérations (Webb et Copsey, 2011, p. 582–583). 32

ACRONYMES

BDI *Beck Depression Inventory*. 6

CBOW *Continuous Bag-of-Words*. 13, 14, 47

DVS Décomposition en valeurs singulières. 12, 17

ETM *Embedded Topic Model*. 2, 6, 21, 22, 44, 46, 47, 49, 51, 53, 54

IA Intelligence Artificielle. 1, 8, 11, 21

LDA *Latent Dirichlet Allocation*. 2, 5, 6, 18, 19, 21, 22, 35, 44, 46, 47, 49, 51, 53

LSA Analyse sémantique latente. 16, 17, 18

PLSA Analyse sémantique latente probabiliste. 17, 18

PWE *Pre-fitted Word Embeddings*. 47, 48

TALN Traitement Automatique du Langage Naturel. 1, 2

NOTATION

Nombres et variables

A une variable quelconque.

A une constante.

$\mathscr{A}$ un ensemble.

$\mathbf{A}$ un vecteur.

$\mathbf{A}$ une matrice.

α un paramètre ou variable latente.

Classification

$\mathcal{X}$ un exemple ou instance.

$\mathcal{Y}$ une étiquette.

Distributions

$\mathrm{Dir}(\cdot)$ une distribution de Dirichlet.

$\mathcal{LN}(\cdot)$ une distribution logit-normale.

RÉFÉRENCES

Abuse, S. *et al.* (2020). Key substance use and mental health indicators in the United States : results from the 2019 National Survey on Drug Use and Health.

Arango, C., Díaz-Caneja, C. M., McGorry, P. D., Rapoport, J., Sommer, I. E., Vorstman, J. A., McDaid, D., Marín, O., Serrano-Drozdowskyj, E., Freedman, R. *et al.* (2018). Preventive Strategies for Mental Health. *The Lancet Psychiatry*.

Arcelus, J., Mitchell, A. J., Wales, J. et Nielsen, S. (2011). Mortality Rates in Patients with Anorexia Nervosa and other Eating Disorders : A Meta-Analysis of 36 Studies. Dans *Archives of General Psychiatry*, *68*.

Beck, A. T., Ward, C., Mendelson, M., Mock, J. et Erbaugh, J. (1961). Beck depression inventory (BDI). *Arch Gen Psychiatry*, *4*(6), 561–571.

Bengio, Y., Ducharme, R. et Vincent, P. (2000). A neural probabilistic language model. *Advances in Neural Information Processing Systems*, *13*.

Blei, D. M., Ng, A. Y. et Jordan, M. I. (2003). Latent Dirichlet Allocation. *Journal of machine Learning research*.

Bloom, D. E., Cafiero, E., Jané-Llopis, E., Abrahams-Gessel, S., Bloom, L. R., Fathima, S., Feigl, A. B., Gaziano, T., Hamandi, A., Mowafi, M. *et al.* (2012). *The global economic burden of noncommunicable diseases*. Rapport technique, Program on the Global Demography of Aging.

Breiman, L. (1996). Bagging predictors. *Machine learning*, *24*(2), 123–140.

Breiman, L. (2001). Random forests. *Machine learning*, *45*(1), 5–32.

Canadian Institute for Health Information (2014). Intentional Self-Harm among Youth in Canada.

Canadian Mental Health Association (2020a). Early Intervention. `https://cmha.ca/public-policy/subject/early-intervention`.

Canadian Mental Health Association (2020b). Fast Facts about Mental Illness.

Cauchy, A. *et al.* (1847). Méthode générale pour la résolution des systemes d'équations simultanées. *Comp. Rend. Sci. Paris*, *25*(1847), 536–538.

Church, K. et Hanks, P. (1990). Word association norms, mutual information, and lexicography. *Computational linguistics*, *16*(1), 22–29.

Coppersmith, G., Dredze, M. et Harman, C. (2014). Quantifying mental health signals in Twitter. Dans *Proceedings of the workshop on computational linguistics and clinical psychology : From linguistic signal to clinical reality*, 51–60.

Coppersmith, G., Dredze, M., Harman, C., Hollingshead, K. et Mitchell, M. (2015). CLPsych 2015 Shared Task : Depression and PTSD on Twitter. Dans *Workshop on Computational Linguistics and Clinical Psychology*.

Cornuéjols, A. et Miclet, L. (2010). *Apprentissage artificiel, concepts et algorithmes* (2 éd.). Eyrolles.

Cornuléjols, A. et Miclet, L. (2018). *Apprentissage artificiel : concepts et algorithmes* (3 éd.). Eyrolles.

Dattani, S., Ritchie, H. et Roser, M. (2021). Mental Health. *Our World in Data*. https ://ourworldindata.org/mental-health.

David M. Blei, Andrew Y. Ng, M. I. J. (2003). Latent Dirichlet Allocation. Dans *Journal of Machine Learning Research 3*, 993–1022.

Deerwester, S., Dumais, S. T., Furnas, G. W., Landauer, T. K. et Harshman, R. (1990). Indexing by Latent Semantic Analysis. *Journal of the American society for information science*, *41*(6), 391–407.

Dempster, A. P., Laird, N. M. et Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society : Series B (Methodological)*, *39*(1), 1–22.

Dieng, A. B., Ruiz, F. J. et Blei, D. M. (2020). Topic Modeling in Embedding Spaces. *Transactions of the Association for Computational Linguistics*.

Duda, R. O., Hart, P. E. et Stork, D. G. (2000). *Pattern classification* (2 éd.). Wiley.

Giatsoglou, M., Vozalis, M. G., Diamantaras, K., Vakali, A., Sarigiannidis, G. et Chatzisavvas, K. C. (2017). Sentiment Analysis Leveraging Emotions and Word Embeddings. *Expert Systems with Applications*.

Gini, C. (1912). *Variabilità e mutabilità : contributo allo studio delle distribuzioni e delle relazioni statistiche*. Tipogr. di P. Cuppini. Bologna.

Gkotsis, G., Oellrich, A., Hubbard, T., Dobson, R., Liakata, M., Velupillai, S. et Dutta, R. (2016). The language of mental health problems in social media. Dans *Proceedings of the third workshop on computational linguistics and clinical psychology*, 63–73.

Henderson, C., Evans-Lacko, S. et Thornicroft, G. (2013). Mental Illness Stigma, Help Seeking, and Public Health Programs. *American journal of public health*.

Hofmann, T. (1999). Probabilistic Latent Semantic Analysis. Dans *Proceedings of the Fifteenth Conference on Uncertainty in Artificial Intelligence*, UAI'99, p. 289–296. Morgan Kaufmann Publishers Inc.

Ive, J., Gkotsis, G., Dutta, R., Stewart, R. et Velupillai, S. (2018). Hierarchical Neural Model with Attention Mechanisms for the Classification of Social Media Text Related to Mental Health. Dans *Workshop on Computational Linguistics and Clinical Psychology*.

James, S. L., Abate, D., Abate, K. H., Abay, S. M., Abbafati, C., Abbasi, N., Abbastabar, H., Abd-Allah, F., Abdela, J., Abdelalim, A. *et al.* (2018). Global, Regional, and National Incidence, Prevalence, and Years Lived with Disability for 354 Diseases and Injuries for 195 Countries and Territories, 1990–2017 : A Systematic Analysis for the Global Burden of Disease Study 2017. *The Lancet*.

Jelodar, H., Wang, Y., Yuan, C., Feng, X., Jiang, X., Li, Y. et Zhao, L. (2019). Latent Dirichlet Allocation (LDA) and Topic Modeling : Models, Applications, a Survey. *Multimedia Tools and Applications*.

Kingma, D. P. et Ba, J. (2014). Adam : A Method for Stochastic Optimization. *arXiv preprint arXiv :1412.6980*.

Losada, D. E., Crestani, F. et Parapar, J. (2018). Overview of eRisk 2018 : Early Risk Prediction on the Internet. Dans *CLEF (Working Notes)*.

Losada, D. E., Crestani, F. et Parapar, J. (2019). Overview of eRisk 2019 Early Risk Prediction on the Internet. Dans *International Conference of the Cross-Language Evaluation Forum for European Languages*. Springer.

Losada, D. E., Crestani, F. et Parapar, J. (2020). Overview of eRisk 2020 : Early Risk Prediction on the Internet. Dans A. Arampatzis, E. Kanoulas, T. Tsikrika, S. Vrochidis, H. Joho, C. Lioma, C. Eickhoff, A. Névéol, L. Cappellato, N. Ferro (eds) (dir.). *Experimental IR Meets Multilinguality, Multimodality, and Interaction - International Conference of the CLEF Association (CLEF 2020)*. Springer International Publishing.

Maupomé, D., Armstrong, M. D., Belbahar, R., Alezot, J., Balassiano, R., Queudot, M., Mosser, S. et Meurs, M.-J. (2020). Early Mental Health Risk Assessment through Writing Styles, Topics and Neural Models. Dans *CLEF (Working Notes)*.

Maupomé, D., Armstrong, M. D., Rancourt, F. et Meurs, M.-J. (2021a). Leveraging Textual Similarity to Predict Beck Depression Inventory Answers. Dans *Proceedings of the Canadian Conference on Artificial Intelligence*.

Maupomé, D., Armstrong, M. D., Rancourt, F., Soulas, T. et Meurs, M.-J. (2021b). Early detection of signs of pathological gambling, self-harm and depression through topic extraction and neural networks. *Proceedings of the Working Notes of CLEF*.

Maupomé, D. et Meurs, M.-J. (2018). Using Topic Extraction on Social Media Content for the Early Detection of Depression. *CLEF (Working Notes)*, *2125*.

Merchant, R. M., Asch, D. A., Crutchley, P., Ungar, L. H., Guntuku, S. C., Eichstaedt, J. C., Hill, S., Padrez, K., Smith, R. J. et Schwartz, H. A. (2019). Evaluating the Predictability of Medical Conditions from Social Media Posts. *PLOS ONE*.

Mikolov, T., Chen, K., Corrado, G. et Dean, J. (2013a). Efficient Estimation of Word Representations in Vector Space. *arXiv preprint arXiv :1301.3781*.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. et Dean, J. (2013b). Distributed Representations of Words and Phrases and their Compositionality. *arXiv preprint arXiv :1310.4546*.

Mimno, D., Wallach, H., Talley, E., Leenders, M. et McCallum, A. (2011). Optimizing Semantic Coherence in Topic Models. Dans *International Conference on Empirical Methods in Natural Language Processing*.

Moody, C. E. (2016). Mixing dirichlet topic models and word embeddings to make lda2vec. *arXiv preprint arXiv :1605.02019*.

Muñoz, R. F., Cuijpers, P., Smit, F., Barrera, A. Z. et Leykin, Y. (2010). Prevention of major depression. *Annual review of clinical psychology*, *6*, 181–212.

Nachar, N. *et al.* (2008). The Mann-Whitney U : A Test for Assessing whether Two Independent Samples Come from the Same Distribution. *Tutorials in quantitative Methods for Psychology*.

Organization, W. H. *et al.* (2015). Excess mortality in persons with severe mental disorders. *Geneva, Switzerland : WHO*.

Parapar, J., Martín-Rodilla, P., Losada, D. E. et Crestani, F. (2021). Overview of eRisk at CLEF 2021 : Early Risk Prediction on the Internet (Extended Overview). *CLEF (Working Notes)*.

Powers, D. M. (2020). Evaluation : from precision, recall and F-measure to ROC, informedness, markedness and correlation. *arXiv preprint arXiv :2010.16061*.

Preoţiuc-Pietro, D., Sap, M., Schwartz, H. A. et Ungar, L. (2015). Mental illness detection at the World Well-Being Project for the CLPsych 2015 shared task. Dans *Proceedings of the 2nd workshop on computational linguistics and clinical psychology : from linguistic signal to clinical reality*, 40–45.

Qin, X., Wang, S. et Hsieh, C.-R. (2018). The prevalence of depression and depressive symptoms among adults in china : estimation based on a national household survey. *China Economic Review*.

Rakotomalala, R. (2014). Pratique de la Régression Logistique. Régression Logistique Binaire et Polytomique. Version 2.0. *Lyon, Université Lumière Lyon-2, multigr*.

Resnik, P., Armstrong, W., Claudino, L., Nguyen, T., Nguyen, V.-A. et Boyd-Graber, J. (2015). Beyond LDA : Exploring Supervised Topic Modeling for Depression-Related Language in Twitter. Dans *Workshop on Computational Linguistics and Clinical Psychology*.

Rosenblatt, F. (1958). The perceptron : a probabilistic model for information storage and organization in the brain. *Psychological review*, *65*(6), 386.

Rumelhart, D. E., Hinton, G. E. et Williams, R. J. (1986). Learning Representations by Back-Propagating Errors. *nature*, *323*(6088), 533–536.

Shen, J. H. et Rudzicz, F. (2017). Detecting Anxiety through Reddit. Dans *Workshop on Computational Linguistics and Clinical Psychology*.

Smink, F. E., van Hoeken, D. et Hoek, H. W. (2012). Epidemiology of Eating Disorders : Incidence, Prevalence and Mortality Rates. Dans *Current Psychiatry Reports 14*.

Statistics Canada (2019). Health Fact Sheets : Mental health care needs, 2018.

Statistics Canada (2020). Care Counts : Receiving Care for a Mental Illness, 2018.

Substance Abuse and Mental Health Services Administration (2020). Key Substance Use and Mental Health Indicators in the United States : Results from the 2019 National Survey on Drug Use and Health.

Tang, D., Wei, F., Yang, N., Zhou, M., Liu, T. et Qin, B. (2014). Learning Sentiment-Specific Word Embedding for Twitter Sentiment Classification. Dans *Annual Meeting of the Association for Computational Linguistics*.

Tenenhaus, M. (2007). *Statistique : méthodes pour décrire, expliquer et prévoir*, volume 680. Dunod.

Webb, A. R. et Copsey, K. D. (2011). *Statistical Pattern Recognition* (3 éd.). Wiley.

Wikimedia Commons (2018). Précision et Rappel. `Precisionrappel.svg`. Récupéré le 2022-05-07 de `https://commons.wikimedia.org/wiki/File:Precisionrappel.svg`

World Health Organization. (2013). *Mental Health Action Plan 2013-2020*. World Health Organization.

World Health Organization. (2016). *Preventing Depression in the WHO European Region*. World Health Organization.

Zhai, K. et Boyd-Graber, J. (2013). Online Latent Dirichlet Allocation with Infinite Vocabulary. Dans *International Conference on Machine Learning*.

Zhao, W. X., Jiang, J., Weng, J., He, J., Lim, E.-P., Yan, H. et Li, X. (2011). Comparing Twitter and Traditional Media using Topic Models. Dans *Advances in Information Retrieval*.