UNIVERSITÉ DU QUÉBEC EN OUTAOUAIS

TECHNIQUES D'APPRENTISSAGE MACHINE POUR L'ESTIMATION DU RISQUE SUICIDAIRE SUR LES RÉSEAUX SOCIAUX

MÉMOIRE
PRÉSENTÉ
COMME EXIGENCE PARTIELLE
DE LA MAÎTRISE EN INFORMATIQUE

PAR

SAMUEL BERNIER

DECEMBRE 2023

UNIVERSITÉ DU QUÉBEC EN OUTAOUAIS

Département d'informatique et d'ingénierie

Ce Mémoire intitulé :

TECHNIQUES D'APPRENTISSAGE MACHINE POUR L'ESTIMATION DU RISQUE SUICIDAIRE SUR LES RÉSEAUX SOCIAUX

présenté par
Samuel Bernier

pour l'obtention du grade de maître ès science (M.Sc.)

a été évalué par un jury composé des personnes suivantes :

Dr. Rokia Missaoui ..... Directrice de recherche
Dr. Mohand-Saïd Allili ..... Président du jury
Dr. Mohamed Hamza Ibrahim ..... Membre du jury

Mémoire accepté le : le 4 décembre 2023

To those I couldn't save.

May this work help me save one.

Remerciements

J'aimerais tout d'abord souligner le soutien et les précieux conseils tout au long de ce mémoire de ma directrice de recherche, Dr. Rokia Missaoui, ainsi que les conseils et précieux commentaires de Dr. Mohand-Saïd Allili et Dr. Mohamed Hamza Ibrahim, cruciaux pour le succès de ce mémoire.

Je tiens également à remercier Alexandre Boudry pour son expertise sur les risques suicidaires et Suicide Action Montréal pour les conseils sur l'application théorique de la solution. Je remercie également les membres du jury pour le temps et l'attention accordée à mon travail.

Je tiens d'autant plus à remercier ma famille et mes amis. Leur soutien m'a permis de surmonter les nombreux défis de ce projet.

Table des matières

Liste des figures

Liste des tableaux

Liste des abréviations, sigles et acronymes

API *Application Programming Interface*

ASHA *Adversarial Suicide assessment Hierarchical Attention*

BERT *Bidirectional Encoder Representations from Transformers*

CNN *Convolutional neural networks*

GPT *Generative pre-trained transformer*

ICL *in-context learning - ICL*

LLM *Large language model - LLM*

LSTM *Long short-term memory*

MLM *Masked Language Model*

MLP *Multilayer Perceptron*

NLP *Natural Language Processing*

NSP *Next Sentence Prediction*

PRAW *Python Reddit API Wrapper*

RAFT *Real-World Annotated Few-shot Tasks*

RNN *Recurrent Neural Networks*

SBERT *Sentence Bidirectional Encoder Representations from Transformers*

SetFit *Sentence Transformer Fine-tuning*

SVM *Support vector machines*

TF-IDF *term frequency-inverse document frequency*

Résumé

Au cours des dernières années, l'influence croissante des plateformes de médias sociaux a révolutionné la communication, reliant des millions d'individus dans le monde entier. Parmi ces plateformes, *Reddit*, en tant que forum social de premier plan, est devenu un foyer pour les utilisateurs pour exprimer librement leurs pensées et leurs émotions. Cependant, ce niveau de connectivité sans précédent met également en lumière des défis préoccupants en matière de santé mentale, y compris des signes potentiels de risque de suicide intégrés dans les publications des utilisateurs. De l'autre côté de la médaille, le traitement automatique du langage naturel (TALN) est un sujet au cœur des discussions d'aujourd'hui et a été la source de nombreuses études au cours des dernières décennies. Avec son expansion remarquable, ce domaine de recherche offre des avancées technologiques notables telles que BERT [15], GPT [36] ou T0 [40]. Celles-ci permettent l'élaboration de solutions à d'importants problèmes sociétaux.

Avec l'augmentation alarmante des problèmes de santé mentale parmi les communautés en ligne, il est primordial d'avoir des mécanismes d'identification précoce des individus à risque de suicide. L'objectif de ce mémoire est justement de tirer parti du pouvoir du TALN pour fournir un soutien rapide et précis aux individus en détresse. Le but est de relever le défi crucial de la détection des signes suicidaires et de présenter une solution innovante en utilisant le modèle *SetFit*, une technique de raffinement de Sentence-BERT [38], dans le contexte des médias sociaux comme *Reddit*. Le modèle *SetFit* a la capacité d'analyser des textes non structurés tout en affichant une précision de classification impressionnante, même avec des données

d'entraînement limitées, ce qui en fait un outil puissant pour l'évaluation du risque de suicide.

Après s'être assuré de la pertinence et de l'authenticité des messages, nous avons créé sur la plateforme *Reddit* un ensemble de données de messages comprenant divers modèles linguistiques et expressions émotionnelles pour des cas de suicides potentiels. Cet ensemble de données sert de pierre angulaire à la validation initiale de l'efficacité du modèle prometteur *SetFit* par le biais d'expériences et d'analyses préliminaires.

Abstract

In recent years, the growing influence of social media platforms has revolutionized communication, connecting millions of individuals worldwide. Among these platforms, Reddit, as a prominent social forum, has become a hotbed for users to express their thoughts and emotions freely. However, this unprecedented level of connectivity also brings to light concerning mental health challenges, including potential suicide risk signs embedded within user posts. On the flip side of the coin, Natural Language Processing (NLP) is a subject at the heart of today's discussion and, as such, has been the source of numerous studies over the past decades. With its great expansion, this field of study offers remarkable technological advancements, such as BERT [15], GPT [36], or T0 [40]. These advancements allow the creation of solutions to significant societal problems.

With the alarming rise in mental health problems among online communities, it is vital to have mechanisms for early identification of individuals at risk of suicide. The aim of this M. Sc thesis is precisely to harness the power of NLP to provide rapid and accurate support to individuals in distress. The aim is to tackle the crucial challenge of detecting suicidal signs and present an innovative solution using the *SetFit* model, a Sentence-BERT refinement technique [38], in the context of social media *Reddit*. The *SetFit* model has the ability to analyze unstructured text while displaying impressive classification accuracy, even with limited training data, making it a powerful tool for suicide risk assessment.

After ensuring the relevance and authenticity of posts, a curated dataset of posts encompassing diverse linguistic patterns and emotive expressions for potential suicidal cases based on the Reddit platform was created. This dataset serves as the

cornerstone for initially validating the promising SetFit model's efficacy through preliminary experiments and analysis.

Chapitre 1

Introduction

1.1 Mise en contexte

Les problèmes de santé mentale au sein de la société sont des enjeux de grande envergure affectant un nombre considérable de personnes. En effet, selon l'Organisation mondiale de la Santé (OMS), plus de 800 000 individus partout dans le monde se suicident chaque année [34, 33]. Bien que ce sujet demeure tabou dans la société, une personne, toutes les trois secondes, tente de se suicider. Ce nombre très impressionnant montre la difficulté de la société à faire face au problème de santé mentale à grande échelle. Non seulement plusieurs personnes vivent avec la pression d'avoir des idées suicidaires, mais l'acte, lorsqu'il est commis, affecte d'autant plus les proches des victimes. Selon l'OMS, chaque suicide entraîne une répercussion d'envergure sur au moins six autres personnes [34, 33]. Ces données statistiques établissent bien l'importance de cette crise mondiale qui mérite, de ce fait, une attention particulière. Les experts de la santé travaillent d'arrache-pied, à l'aide de campagnes de sensibilisation, afin d'atteindre le plus de personnes victimes d'idées suicidaires, sans toutefois être en mesure de réduire ce nombre [17].

À l'ère des technologies avancées de l'information, il faut considérer l'utilisation de récentes technologies afin d'affronter ce fléau. À l'aide de nouvelles techniques d'apprentissage automatique, il est possible d'apporter une aide au système de la santé,

afin de cibler le plus de suicidaires potentiels et de prévenir leurs tentatives. L'un des points essentiels à la prévention du suicide est d'identifier les signaux suicidaires avant qu'un individu passe à l'acte [29].

Toutefois, la détection de tels signaux s'avère être une tâche très complexe. Heureusement, l'accès aux médias sociaux publics permet l'identification des signes précoces, dans le but de les analyser. Ceux-ci offrent un aperçu des pensées individuelles d'un utilisateur. Certains médias sociaux, tels que Twitter ou Reddit, sont conçus dans l'objectif de permettre aux utilisateurs d'exprimer leurs pensées et leurs émotions. Ces réseaux sociaux offrent une opportunité aux victimes ayant des pensées suicidaires de chercher de l'aide indirectement puisqu'ils les expriment. Dans un monde utopique, les professionnels de la santé analyseraient le profil d'utilisateur de médias sociaux par million afin de détecter les signes d'idées suicidaires chez un individu. Néanmoins, il manque de professionnels de la santé disponibles pour compléter ce rôle d'analyse. L'utilisation d'algorithmes d'apprentissage automatique aurait donc le potentiel de réaliser cette tâche, agissant comme premier filtre pour établir un portrait d'ensemble aux professionnels de la santé. Un tel filtre permettrait d'apporter l'aide nécessaire aux personnes les plus à risque de passer à l'acte et qui seraient, en temps normal, oubliées par notre société.

Les statistiques de cette problématique montrent l'importance d'utiliser une approche professionnelle face aux personnes plus à risque. En revanche, il peut sembler difficile de détecter les signes précurseurs au passage à l'acte. L'étude réalisée par Monique Séguin et Christian Lafleur [23] souligne un mythe très commun au suicide, qui est qu'une personne suicidaire ne va pas discuter de ces problèmes avant de passer à l'acte. Or, « la grande majorité des personnes qui se sont suicidées ont laissé des indices de leurs intentions. Les messages directs ou indirects que les personnes que nous côtoyons envoient doivent être pris au sérieux. Toutefois, nous devons reconnaître que chez une petite proportion, le geste peut être impulsif ».

Les médias sociaux offrent un avantage extraordinaire quant à la détection d'idée suicidaire chez un individu. Certaines de ces plateformes sont conçues dans le but d'offrir un moyen d'exprimer ses émotions et ses pensées. Les plateformes telles que

Twitter et Reddit sont très populaires sur Internet et offrent justement un tel moyen d'expression.

1.2 Formulation du problème

L'utilisation des techniques d'apprentissage automatique afin de prévenir le suicide est toujours un sujet d'actualité. En effet, la question est abordée par plusieurs recherches qui seront détaillées dans le chapitre « État de l'art ». Le domaine d'apprentissage automatique est propice à la résolution de ce problème et plus précisément le traitement du langage naturel (*Natural Language Processing* - NLP) [59]. Ce vaste domaine comprend des dizaines de techniques et d'algorithmes pouvant effectuer une telle tâche, chacun ayant des aspects positifs et des aspects négatifs qui s'y rattachent. Il s'agit, entre autres, de l'utilisation de la technique « fréquence des termes - fréquence inverse des documents » (TF-IDF - *term frequency-inverse document frequency*) [62], de machines à vecteurs de support (SVM - *Support vector machines*) [61], ou bien des réseaux de neurones (*Neural Networks*) [52]. Toutefois, ces techniques ont tous un problème commun quant à l'approche de cette problématique. Celles-ci classifient ce problème en utilisant une solution binaire, tandis que le problème nécessite une approche à plusieurs variables. Le suicide est un problème très complexe nécessitant la considération d'une dizaine de facteurs afin d'obtenir une estimation utilisable. Ces facteurs peuvent être des facteurs individuels et personnels, familiaux ou même des éléments de la vie, ayant chacun un niveau d'urgence, selon Monique Séguin et Christian Lafleur [23]. Par conséquent, l'utilisation d'algorithmes plus complexes tels que BERT se rapproche davantage d'une solution adéquate [41].

Il est utile de formuler tout d'abord la tâche qui consiste en effet à identifier les utilisateurs potentiellement suicidaires sur Reddit à l'aide de la grille d'évaluation de la dangerosité du passage à l'acte offerte par le gouvernement du Québec [24]. Il s'agit d'un problème de classification multiclasses. En d'autres termes, supposons que $\mathcal{D} = \{t_i, y_i\}_{i=1}^{n}$ représente notre ensemble de données étiquetées (collectées comme décrit dans la sous-section du chapitre « Méthode proposée »), où t_i représente le texte du i-ème message d'auto-déclaration dans les fils de discussions suicidaires de

Reddit. $Y_i \in \{1, 2, 3, ...n\}$ fait référence à l'étiquette du i-ème message des classes « planification du suicide », « tentative de suicide récente », « désespoir », « abus de substance », « impulsivité », « solitude et isolement» et « Soins personnels ».

1.3 Objectif

L'objectif de ce mémoire est d'analyser des documents textuels à l'aide de nouvelles techniques du traitement du langage naturel. Par conséquent, le sous-objectif de ce mémoire de maîtrise est d'offrir un outil pouvant servir de filtre aux professionnels de la santé afin de cibler les personnes les plus à risque selon leurs publications sur internet. Cette solution offre un système décisionnel d'alerte précoce qui reçoit en entrée les nouveaux messages publiés par l'intéressé et détermine ensuite si chacun d'entre eux dénote un signe de suicide potentiel.

Il est important de noter qu'il s'agit d'un mémoire en informatique et non d'un mémoire en santé mentale. L'objectif n'est donc pas de remplacer les professionnels de la santé ni de redéfinir le fonctionnement d'une intervention lors d'une situation de crise suicidaire. Il s'agit plutôt d'offrir un outil aux personnes qualifiées en intervention de situation de crise suicidaire afin d'aider à la résolution de ce problème à grande échelle. Cet outil aurait le potentiel d'être déployé sur les médias sociaux afin de cibler les personnes les plus à risques et ainsi les référer à des professionnels qualifiés afin d'apporter une aide plus rapidement. De cette façon, les efforts du système de santé seraient concentrés vers un groupe de personnes plus à risque parmi des millions de personnes actives sur les médias sociaux. L'exécution manuelle de cette tâche par un professionnel de la santé, d'analyser chaque profil, apporterait assurément un plus grand taux de succès en matière de prévention du suicide, mais cela exigerait un nombre inconcevable de ressources humaines. L'utilisation du traitement du langage naturel pourrait grandement aider à la réduction du passage à l'acte.

Afin d'atteindre notre objectif, nous utilisons la méthodologie suivante :

- Sélection d'une grille du risque du passage à l'acte, incluant les signes à détecter.

- Création d'un ensemble de données pour l'entraînement de l'algorithme suivant la grille d'estimation du passage à l'acte à l'aide d'un expert du domaine.
- Entraînement d'un algorithme de classification à plusieurs classes des techniques du traitement du langage naturel.
- Analyse de profils sur les médias sociaux estimant le risque du passage à l'acte en passant par la création d'un ensemble de données anonymes incluant les profils analysés, ainsi que les signes détectés et le texte relié à ceux-ci. Cet ensemble de données est compilé de façon qu'un professionnel puisse faire sa propre évaluation de l'algorithme.

Une contribution est envisagée par la collecte d'un ensemble de données d'entraînement créé en collaboration avec un expert du domaine et par la création d'un modèle flexible d'estimation de la dangerosité du passage à l'acte pouvant être adapté à plusieurs médias sociaux. Le modèle servira de preuve de concept permettant l'utilisation d'un ensemble de données d'entraînement limité afin de créer un modèle de classification autant efficace que les techniques populaires.

Chapitre 2

Préliminaires

Le sujet de la santé mentale, et plus précisément du suicide, est un enjeu de société qui perdure dans le temps. Par conséquent, il existe plusieurs recherches sur le sujet présentant des pistes de solution distinctes utilisant des algorithmes différents. Ce chapitre est donc séparé en deux sections. La première section fait un rappel des notions de base déjà établies par la communauté scientifique ayant un impact sur notre problématique. Celle-ci inclut des pistes de solutions et des algorithmes abordés par diverses recherches dans le domaine. La deuxième section porte sur l'état de l'art concernant la santé mentale et le suicide. Cette dernière partie est décrite de manière sommaire, puisqu'il ne s'agit pas d'un mémoire sur la santé mentale, l'essentiel est tout de même discuté afin de bien comprendre le problème en question.

2.1 Notion de base

La section suivante se concentre sur les algorithmes utilisés dans le contexte de la problématique, incluant leurs forces et leurs faiblesses, jusqu'à la définition de l'algorithme choisi.

2.1.1 Réseau de neurones

Les réseaux de neurones sont présentement la technique d'apprentissage la plus populaire dans plusieurs domaines de l'apprentissage automatique, particulièrement pour le traitement du langage naturel. Il existe plusieurs types de réseaux de neurones, tels que les réseaux de neurones récurrents (RNN - *Recurrent Neural Networks*), les réseaux de neurones convolutionnels (CNN - *Convolutional neural networks*), les transformateurs (*Transformers*), etc. À la base, ceux-ci offrent une approche imitant le fonctionnement du cerveau humain. Chaque réseau de neurones est, comme le nom l'indique, une représentation graphique des liens entre de plusieurs neurones. Ceux-ci sont placés en plusieurs couches selon le type de réseau de neurones désiré. Les réseaux sont habituellement composés d'une couche d'entrée des données, d'une couche cachée et d'une couche des données de sortie. Ces couches sont composées de neurones qui s'activent avec des poids. Lorsque le réseau est entraîné, le poids entre chaque neurone s'ajuste afin d'obtenir le résultat désiré.

La précodure sert principalement de fondation à des algorithmes très puissants utilisés à ce jour pour le traitement du langage naturel, tels que les LSTM [57], BERT [15], GPT [36] ou T0 [40]. Or, ce dernier est utilisé directement sur notre problématique [14], qui présente des résultats prometteurs avec un taux de précision de 73.22%. En revanche, il existe des algorithmes basés sur ces fondements qui offrent une meilleure performance et une meilleure précision, tels que GPT, T0, BERT ou ASHA [41] utilisant des transformateurs.

2.1.2 Apprentissage contrastif

L'apprentissage contrastif (*Contrastive learning*) est une approche d'apprentissage automatique qui vise à créer des liens dans un ensemble de données en comparant des instances similaires et opposées sans étiquette. Le classement est produit par l'apprentissage des données en encourageant la similarité entre des instances syntaxiquement liées, tout en éloignant les instances dissemblables dans l'espace d'intégration. Pour ce faire, on utilise une formule d'apprentissage minimisant la distance entre les paires positives (similaires) et maximisant la distance entre les paires néga-

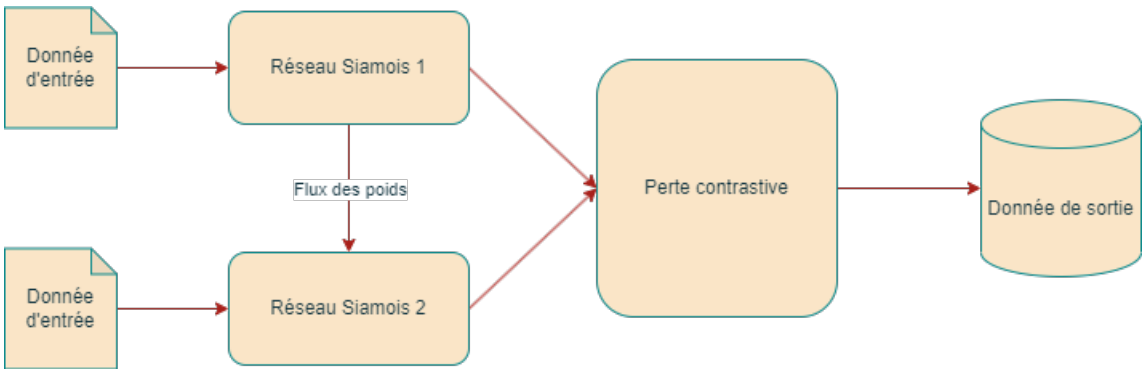
tives (dissemblables). Plusieurs algorithmes utilisant cet apprentissage contrastif ont été proposés pour des tâches de traitement du langage naturel tel que SBERT [38], SimCSE [18] et InfoNCE [50].

Ce type d'apprentissage a été appliqué avec succès à divers domaines, notamment la vision par ordinateur, la reconnaissance vocale et le NLP, prouvant son potentiel d'amélioration des performances des modèles d'apprentissage automatique, en particulier dans les scénarios où les données étiquetées sont rares.

Réseau neuronal siamois

Le réseau neuronal siamois (*Siamese neural network*) est une architecture utilisant l'apprentissage contrastif. Ce dernier est conçu pour apprendre les métriques de similarité ou de distance entre des paires de données d'entrée. Les réseaux siamois se composent d'au moins deux sous-réseaux identiques, également appelés « réseaux frères », qui partagent les mêmes poids et la même architecture [11]. Chaque réseau traite les données d'entrée de manière indépendante, en produisant une représentation vectorielle de taille fixe. Les vecteurs sont ensuite combinés dans une formule de perte contrastive, telle que la perte de triplet, pour calculer un pointage de similarité ou de différence. Ce pointage peut être utilisé pour déterminer la relation entre les instances d'entrée. Voici une représentation d'un réseau neuronal siamois dans le contexte d'analyse d'images :

FIGURE 2.1 – Réseau neuronal siamois



La perte de triplet

La perte de triplet (*Triplet loss*) est une fonction de coût contrastif (*contrastive loss*) pouvant être utilisée dans les réseaux neuronaux siamois. Dans le contexte de perte de triplet, un triplet se compose de trois instances : une ancre (A), une positive (P) et une négative (N). L'ancre et les instances positives sont similaires, tandis que l'ancre et les instances négatives sont dissemblables. L'objectif de la perte de triplet est d'apprendre des représentations telles que la distance entre l'ancre et l'instance positive soit plus petite que la distance entre l'ancre et l'instance négative. Elle se définit par $L(A, P, N) = \max(0, D(A, P) - D(A, N) + \text{marge})$, où $D(A, P)$ indique la distance entre les représentations de l'ancre et des instances positives, $D(A, N)$ spécifie la distance entre les représentations de l'ancre et des instances négatives, et la « marge » est une constante prédéfinie qui impose une séparation minimale entre les paires positives et négatives [63]. Lors de l'entraînement, l'objectif est de minimiser la fonction de perte utilisant des techniques telles que la descente du gradient.

2.2 Domaine d'application

Bien que ce mémoire soit une étude sur le traitement du langage naturel et sur la santé mentale, il demeure primordial d'assurer la compréhension de certaines notions de base. Il est important de clarifier les concepts élémentaires, les méthodes utilisées à ce jour, ainsi que les limites. Cela permet d'assurer une meilleure compréhension de la solution à la problématique présentée. L'exploration du domaine d'application est guidée par des professionnels de la santé ayant de réelles expériences en termes d'estimation du risque suicidaire.

2.2.1 Santé mentale

L'estimation de la dangerosité du passage à l'acte est une problématique très complexe. En effet, la réduction de cette question à un simple « vrai » ou « faux » comme certains articles présentent, n'est pas convenable à ce type de problème. C'est pourquoi les professionnels de la santé universellement utilisent des grilles d'es-

timation permettant de se donner des pistes d'analyses. La Commission de la santé mentale du Canada ainsi que l'Institut canadien pour la sécurité des patients offrent un outil d'évaluation du risque de suicide [32]. Ce document propose une vingtaine de grilles d'estimation avec, encore une fois, chacun des avantages et des défauts respectifs. Tel que mentionné précédemment, puisqu'il s'agit d'un mémoire en informatique et non d'un mémoire en santé mentale, il va donc se limiter à l'explication d'une seule grille. Néanmoins, chacune de ces grilles est évaluée lors de la rédaction de ce mémoire.

La santé mentale et les pensées suicidaires sont une science toujours en évolution à ce jour, étant l'objet de nouvelles recherches et de méthodes d'approche. Par conséquent, les méthodes utilisées lors de ce mémoire sont en date de l'année 2023 et peuvent changer au cours du temps. Lors de la rédaction de ce mémoire, l'Ordre des travailleurs sociaux du Québec utilise une grille d'estimation, publiée par Santé et Services sociaux du Québec, dans un document intitulé « Prévention du suicide » [24]. Le document propose une grille d'estimation du niveau de probabilité d'une tentative de suicide qui est employée par les travailleurs sociaux du Québec, utilisant une approche orientée vers la solution. Celle-ci comprend plusieurs aspects très importants :

1. Elle prend en compte l'urgence suicidaire, les facteurs principalement associés au suicide et les facteurs de protection. Elle accorde un poids, représentant le niveau d'importance, à chacun des facteurs, simplifiant la tâche des travailleurs sociaux.
2. Elle cible les facteurs les plus près du passage à l'acte tel que :
   - **La planification du suicide**. L'évaluation comprend le « quand », le « où » et le « comment ». Puis, elle inclut également les moyens et les préparatifs.
   - **Les tentatives de suicide antérieures**. Une tentative de suicide récente, de la part d'un individu, est un facteur aggravant.
   - **La capacité à espérer un changement**. Un individu vivant un grand désespoir possède un facteur de risque aggravé par ce ressentiment.
   - **L'usage de substances**.

- **La capacité à se contrôler**. Une forte impulsivité est un facteur aggravant.
- **La présence des proches**. La solitude et l'isolement sont des facteurs aggravants.
- **La capacité à prendre soin de soi**. Un individu qui « se laisse aller» est un facteur aggravant. [24]

3. Elle cible certains symptômes des problèmes de santé mentale souvent associés au suicide tels que la dépression par exemple.

De plus, la grille propose une classification de chaque signe en quatre catégories distinctives, en allant du vert, au jaune, à l'orange et au rouge. Les couleurs représentent le niveau de dangerosité du signe identifié, vert étant le plus faible et rouge étant le plus grave. Par exemple, une personne n'ayant aucune planification pour passer à l'acte sera classée dans le « vert » pour ce signe précis. Une personne qui est très impulsive aura un indicateur rouge dans la catégorie « capacité à se contrôler ». Toutefois, l'utilisation de cette grille à ce niveau requiert une formation auprès d'un organisme du domaine, tel que Suicide Action Montréal (SAM) par exemple.

La grille permet d'autant plus d'aider les intervenants à trouver les leviers d'intervention et les actions à entreprendre en fonction de l'estimation finale, mais ces deux aspects sont hors du champ d'application de cette recherche. Il est important de rappeler que l'objectif est d'agir comme filtre afin de référer rapidement les personnes en situation de détresses aux ressources appropriées. Le but n'est pas de remplacer les intervenants du domaine. L'estimation de la dangerosité du passage à l'acte est surtout une évaluation complémentée par l'expertise des professionnels de la santé.

Dans l'objectif d'utiliser cette grille et pour bien comprendre les explications supplémentaires qu'elle fournit, une formation avec l'organisme Suicide Action Montréal est nécessaire, d'où la raison pour laquelle cette recherche est assistée d'un travailleur social reconnu. Les spécificités de l'approche proposée, utilisant cette grille, seront détaillées au quatrième chapitre.

2.2.2 Médias sociaux

Les médias sociaux offrent à chaque individu un endroit pour s'exprimer librement et facilement. Certains d'entre eux sont spécialement conçus dans l'optique d'offrir une plateforme aux utilisateurs afin d'exprimer leurs pensées. En effet, les médias sociaux furent l'objet de nombreuses études sur l'analyse de l'opinion du public pour cette raison précise [46, 31]. Il existe deux médias sociaux en particulier qui ressortent du lot en raison de leur accessibilité au travers d'un API public. Il s'agit Twitter et Reddit.

Reddit

Reddit est une plateforme de publications de nouvelles, de contenu web et de discussions permettant aux utilisateurs de soumettre du contenu gratuitement, tel que des messages textuels, des liens URL, des images et des vidéos. Ce contenu, généré par les utilisateurs, est regroupé par thème par des sous-forums appelés « subreddit ». Ceux-ci comprennent un grand éventail de sujets incluant l'actualité, la science, les loisirs, les divertissements, etc. Les utilisateurs peuvent noter les publications à la hausse ou à la baisse, ce qui influence la visibilité du contenu sur la plateforme. Ils peuvent aussi participer à des discussions en publiant des commentaires sur les soumissions.

Reddit est une source importante pour les chercheurs en traitement du langage naturel en raison de l'important volume et de l'importante diversité de contenu textuel. Avec des millions d'utilisateurs actifs, Reddit génère une grande quantité de données quotidiennement classées par thème (ou « *subreddit* »). Ces données sont riches en langage formel et informel, ce qui en fait une excellente source pour l'entraînement et l'évaluation des modèles d'apprentissage automatique.

La structure unique de Reddit, avec ses fils de commentaires imbriqués et ses interactions avec les utilisateurs, offre plusieurs défis et opportunités pour la recherche en NLP. La plateforme contient des conversations à divers niveaux, où le contexte joue un rôle crucial dans la compréhension de la signification des commentaires in-

dividuels. En outre, la présence d'abréviations, d'émoticônes et d'échanges de codes ajoutent à la complexité du texte.

L'alternative populaire par les chercheurs en traitement du langage naturel est Twitter. Celle-ci a été l'objet de plusieurs recherches du domaine, notamment l'analyse des sentiments, la détection des thèmes, la détection des sarcasmes, le résumé, etc. [41, 25] L'ensemble de données vaste et diversifié de ces deux plateformes offre une source de données riches pour les chercheurs du domaine. Ceux-ci utilisent ces plateformes dans la création de nouvelles technologies et de nouveaux modèles d'apprentissage automatique.

Les avantages de Reddit sont notés par :

- **Des communautés riches et spécialisées** : Reddit héberge un large éventail de communautés, connues sous le nom de "subreddits", dédiées à des sujets, des intérêts et des conditions de santé spécifiques. Cela permet des discussions davantage ciblées. Ces communautés spécialisées peuvent être utiles lorsque vous cherchez à identifier des patients potentiels souffrant de pathologies spécifiques ou à obtenir des informations sur des questions médicales particulières.
- **Des discussions approfondies** : Comparé à la limite de caractères de Twitter, Reddit permet des discussions plus approfondies avec des informations détaillées. Les utilisateurs peuvent fournir des descriptions précises de leurs symptômes, de leurs expériences et de leurs préoccupations, ce qui peut s'avérer précieux pour l'analyse NLP. Cette profondeur dans le contenu permet une analyse et une compréhension plus solide des besoins des utilisateurs en matière de santé.
- **L'anonymat et la protection de la vie privée** : Reddit offre un degré d'anonymat élevé. Les utilisateurs ont tendance à créer un nouveau compte anonyme lors de la publication sur certains sous-forums. Les noms d'utilisateurs peuvent être complètement aléatoires et n'offrent aucun indice sur l'identité de l'auteur. Ce type d'anonymat encourage les utilisateurs à publier leurs pensées parfois cachées et permet une plus grande ouverture de leurs émotions. Il n'est pas rare de découvrir que le compte de l'auteur d'une pu-

blication est créé seulement pour une seule publication ou pour participer à un seul sous-forum.

- **La longévité des discussions** : Les discussions sur Reddit ont tendance à avoir une durée de vie plus longue que sur Twitter, où les tweets s'effacent rapidement au cours du temps. Les messages et les commentaires sur Reddit restent souvent accessibles et consultables pendant une longue période, ce qui facilite l'analyse des données historiques et le suivi de la progression des discussions au fil du temps.
- **Un contenu ciblé et organisé** : Reddit propose des fonctionnalités telles que le *post flair*, le *tagging* et la modération qui permettent de maintenir un contenu organisé et structuré au sein des s*ubreddits*. Cela peut être avantageux pour les méthodes NLP car cela facilite la catégorisation et le filtrage des discussions pertinentes.

2.3 Traitement du langage naturel

Le traitement du langage naturel (*Natural language processing - NLP*), est un domaine d'apprentissage automatique. Son objectif principal est de permettre aux ordinateurs de transformer du contenu textuel non structuré en un contenu textuel structuré. Si l'on prend par exemple un courriel. Le contenu principal d'un courriel est un texte clair représentant de l'information non structurée. Or, les informations telles que la date d'envoi, l'expéditeur et le récipient sont tous des informations structurées. Cela s'explique par le fait qu'elles peuvent être facilement utilisées et interprétées par un ordinateur pour certaines tâches telles que classer les courriels en ordre de réception. Si le désir est de classer les courriels par thématique, l'utilisation du NLP pourra s'avérer utile afin d'analyser le texte et d'en déterminer le thème.

Le NLP englobe un large éventail de tâches, telles que l'analyse des sentiments, la traduction automatique, le résumé de textes et les systèmes de réponses aux questions. Ces tâches requièrent l'application de plusieurs techniques de transformation, incluant la création de jetons (ou « tokenisation »), l'étiquetage du texte ou la reconnaissance des entités. Au cours des dernières années, le domaine a connu plusieurs

avancées technologiques importantes en raison des réseaux de neurones et des transformateurs. Les modèles les plus importants aujourd'hui incluent BERT [15], GPT [36] et T5 [37].

Malgré le progrès impressionnant dans ce domaine, il existe toujours des défis liés à la capture de la subtilité du langage humain. Il existe plusieurs techniques populaires visant à résoudre ce problème.

2.3.1 Classificateur multiclasses

Dans le contexte du traitement du langage naturel, il existe de nombreux types de classificateurs. Afin d'utiliser la grille mentionnée à la section « Domaine d'application », il est essentiel de procéder à l'utilisation d'un classificateur multiclasses, plutôt qu'un classificateur binaire. Contrairement au classificateur binaire qui fait la distinction entre deux classes, comme entre un sentiment soit positif ou soit négatif par rapport à un article de journal, les classificateurs multiclasses peuvent traiter plus de deux classes, par exemple, le thème d'un article de journal étant « finance », « politique » ou « sport », ce qui les rendent adaptés à un plus large éventail de problèmes. La grille d'estimation choisie pour la problématique de ce mémoire compte sept catégories de signes différents, obligeant ainsi à faire l'utilisation d'un classificateur multiclasses. Nonobstant cela, plusieurs recherches, s'attardant à la même problématique, utilisent des classificateurs binaires puisque ceux-ci permettent la classification d'un texte de « risque suicidaire » avec un indicateur « vrai » ou « faux » [13]. La littérature offre plusieurs algorithmes d'apprentissage automatique utilisé dans un tel contexte. Parmi ceux-ci, les plus populaires incluent la régression logistique, les machines à vecteurs de support (SVM), les arbres de décision, les réseaux de neurones, etc.

Régression logistique

La régression logistique est une technique populaire en apprentissage automatique, servant à faire de la classification de données basée sur certains modèles.

Celle-ci s'avère populaire en analyse de données due à sa facilité d'implémentation et au besoin minime d'interaction humaine lors de la conception [56].

Bien que cette technique ne présente pas directement une étude sur la détection du suicide, elle sert de fondement à la création de réseaux de neurones de tout genre. Il est donc primordial de revisiter ces notions de base afin d'assurer la compréhension des prochaines techniques d'apprentissage automatique.

Tout d'abord, cette méthode combine chaque mot de l'ensemble de données source et lui associe une valeur numérique afin de pouvoir les retrouver facilement. Il s'agit d'une technique d'indexation intitulée « codage de paires d'octets » (*Byte pair encoding*) [53]. Par la suite, lors de l'analyse de la base de données, chaque mot va être associé à un poids se situant entre 0 et 1, selon le contexte de l'analyse. Par exemple, dans un contexte de risque du suicide, les mots reliés à des pensées suicidaires auraient comme attribution l'indicatif 1, alors que les mots qui ne sont pas reliés, obtiendraient l'indicatif 0. Une fois le texte converti, ce dernier sera traité par une fonction sigmoïde [22] :

$$\log\left(\frac{y}{1-y}\right) = b_0 + b_1x_1 + b_2x_2 + b_3x_3 + \cdots + b_nx_n \tag{2.1}$$

La valeur « y » indique la probabilité que le texte représente un risque suicidaire (la valeur sera donc égale à 1) ou non (la valeur sera donc égale à 0). Les variables x_1 à x_n représentent les mots du texte, et les valeurs b_0 à b_n représentent le poids de chaque mot. Cette formule permettra donc de déterminer si le texte est en thème ou non avec le thème du suicide, lui accordant une valeur entre 0 et 1.

L'avantage de la régression logistique est qu'il s'agit d'un algorithme très simple à implémenter et ne nécessite pas de grandes ressources d'exécution. Néanmoins, il s'agit d'une solution très simpliste à un problème d'envergure tel que le sujet choisi pour ce mémoire. De plus, cet algorithme ne considère pas le contexte de chaque mot ainsi que l'association entre eux.

Perceptron multicouche

Les perceptrons multicouches (*Multilayer Perceptron - MLP*) sont un type de réseau de neurones souvent utilisés dans des tâches d'apprentissage automatique. Un MLP se compose de plusieurs couches de neurones interconnectées, également appelées nœuds ou neurones artificiels, qui sont organisées en couche d'entrée, en couche cachée et en couche de sortie. Ces couches travaillent ensemble afin de transformer les données d'entrée en prédictions ou en classification, tout comme les réseaux de neurones. [58]

Un MLP est formé à l'aide d'une technique d'apprentissage supervisée appelée rétro propagation et qui ajuste le poids des connexions entre les neurones afin de minimiser la différence entre les prédictions du réseau et les valeurs cibles réelles (c'est-à-dire l'erreur). Ce processus est généralement répété à des fins de performance pendant plusieurs époques (*epochs*) qui sont des passages complets par les données d'apprentissage.

2.3.2 Grand modèle linguistique

Un grand modèle linguistique (*Large language model - LLM*) est un modèle avancé de traitement du langage naturel conçu pour comprendre, interpréter et même générer du langage humain. Il se compose de millions ou même de milliards de paramètres, ce qui lui permet d'apprendre et de représenter des modèles et des structures complexes dans le langage. Ces modèles sont entraînés à l'aide d'énorme quantité de textes, principalement tirées d'internet. Celles-ci leur permettent d'analyser un large éventail de nuances, d'informations contextuelles et de caractéristiques linguistiques.

Le développement des LLM est stimulé par l'impressionnant progrès des nouveaux réseaux de neurones, plus précisément les transformateurs. Ceux-ci sont discutés en détail à la sous-section suivante. Parmi les exemples de grands modèles de langage populaire, on peut citer la série GPT [36] et [15].

Les LLM sont raffinés selon la tâche choisie. Par exemple, la classification de textes par thème, l'analyse des sentiments, la traduction automatique, les résumés, etc. Les avancements technologiques du domaine ont conduit à l'utilisation de plus

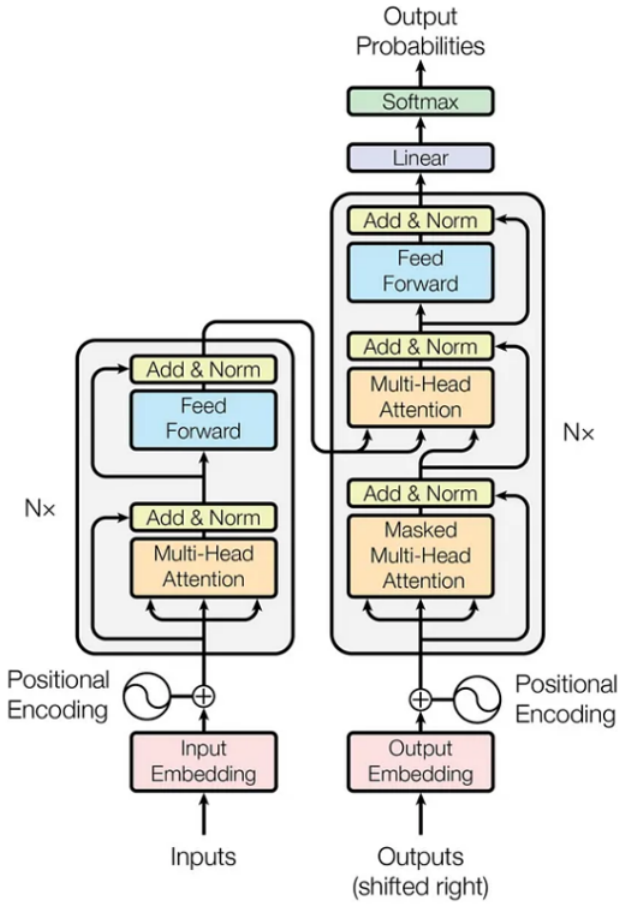
en plus diversifiée de l'intelligence artificielle, dont les robots conversationnels tels que ChatGPT.

2.3.3 Transformateur

Les transformateurs (*Transformers*) sont un nouveau modèle d'apprentissage automatique proposé par Vaswani et al. en 2017 dans l'article « *Attention is All You Need* » [51], obtenant une popularité fulgurante dans le domaine. En effet, les algorithmes les plus performants depuis 2017 utilisent tous les transformateurs. Ces derniers ont une capacité impressionnante à capturer les dépendances à long terme, mais aussi à traiter les données en parallèle. Cette technique permet l'obtention de résultats de performance et de précision très impressionnants [51]. Les transformateurs ont permis la création d'algorithmes très populaires à ce jour, incluant BERT [15] ou GPT [36].

Bien que les transformateurs soient aussi considérés comme un réseau de neurones, la plus grande différence entre ceux-ci et les divers réseaux de neurones présentés est le mécanisme d'attention introduit par Vaswani et al. [51]. Ce mécanisme permet au modèle de se concentrer sur différentes parties de la séquence d'entrée lorsque les prédictions sont faites. Chaque élément obtient un pointage qualifiant la relation avec les autres éléments du texte. Le pointage sert ensuite à pondérer la contribution de chaque élément à la sortie. Voici le schéma des transformateurs présenté par Vaswani et al.

FIGURE 2.2 – Architecture d'un transformateur [51]



Les transformateurs utilisent une architecture encodeur-décodeur à chaque neurone. Pour ce qui est de la partie de gauche, il s'agit de l'encodeur et il peut être utilisé n nombre de fois. Pour ce qui est de la partie de droite, il s'agit du décodeur et il peut être utilisées n nombre de fois. Les encodeurs et les décodeurs ont certaines composantes clés permettant le bon fonctionnement :

1. La première composante est le **mécanisme d'auto-attention** qui permet au modèle de mettre un poids sur les relations entre tous les éléments de

la séquence d'entrée. Ce processus permet au transformateur de capturer les dépendances, peu importe la position de l'élément dans le texte.

2. La deuxième composante est l'**attention multitêtes**. Dans la même optique que les LSTM, cette composante a comme objectif de se concentrer sur divers aspects des données d'entrée. Chaque tête calcule son propre ensemble de pointages d'attention, et les résultats sont rassemblés et transformés linéairement pour produire la sortie finale.
3. La troisième composante est le **codage de position**. Puisque les transformateurs n'ont pas de traitement séquentiel inhérent comme les réseaux de neurones récurrents, ceux-ci doivent encoder la position de l'élément d'entrée afin de tenir compte des positions relatives lors du calcul du pointage d'attention.
4. La quatrième composante est la **normalisation de couches et de connexions résiduelles**. Cette composante permet d'améliorer la stabilité de l'entraînement et de faciliter le flux de gradient. Cette couche normalise les activations des caractéristiques d'entrée, tandis que les connexions résiduelles déterminent si le modèle doit ajouter ou supprimer des informations d'entrée au besoin. Cette composante est comparable au domaine de rétention d'information d'un LSTM.
5. La cinquième composante est la **couche de propagation avant** (*feed-foward*). Cela signifie que contrairement aux réseaux de neurones récurrents, les transformateurs n'ont pas de cycles entre les neurones. L'information circule vers l'avant entre chaque neurone (*feed-foward*).

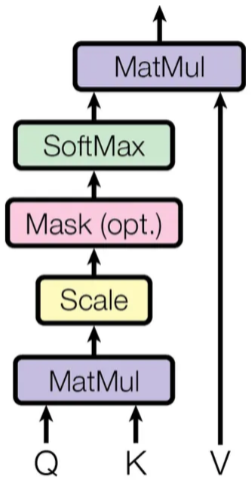
L'élément permettant de différencier ce modèle des autres est le modèle d'attention. Le mécanisme d'attention multitêtes est formé à l'aide de l'opération matricielle d'attention (« *Scaled dot-product attention*)». Celle-ci peut être décrite à l'aide de la formule suivante :

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

[51]

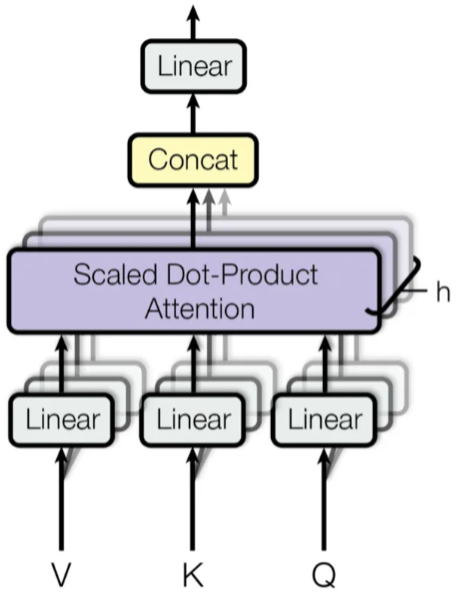
Q représente la matrice de la requête. K représente la matrice des clés, et V est la matrice des valeurs. d_k est la dimension des vecteurs des clés (et des requêtes). La fonction *softmax* est appliquée à la matrice résultante élément par élément, générant les poids d'attention. L'architecture est représentée dans l'article de Vaswani et al. [51] de cette façon :

FIGURE 2.3 – Architecture du produit scalaire avec attention



Ensuite, cette description est utilisée parallèlement à l'aide de plusieurs couches d'attention.

FIGURE 2.4 – Architecture de l'attention multi-tête [51]



Les transformateurs sont présentement la technique d'apprentissage automatique la plus performante et précise en termes de traitement du langage naturel [51, 15], d'où la raison de leur énorme succès dans certains algorithmes fondés sur ce modèle, tel que BERT ou GPT. Selon le modèle d'évaluation GLUE, les modèles utilisant des transformateurs montrent constamment une meilleure performance que les modèles basés sur les réseaux de neurones récurrents ou convolutifs [1]. Il s'agit donc d'un excellent candidat à la résolution de la problématique précédemment définie.

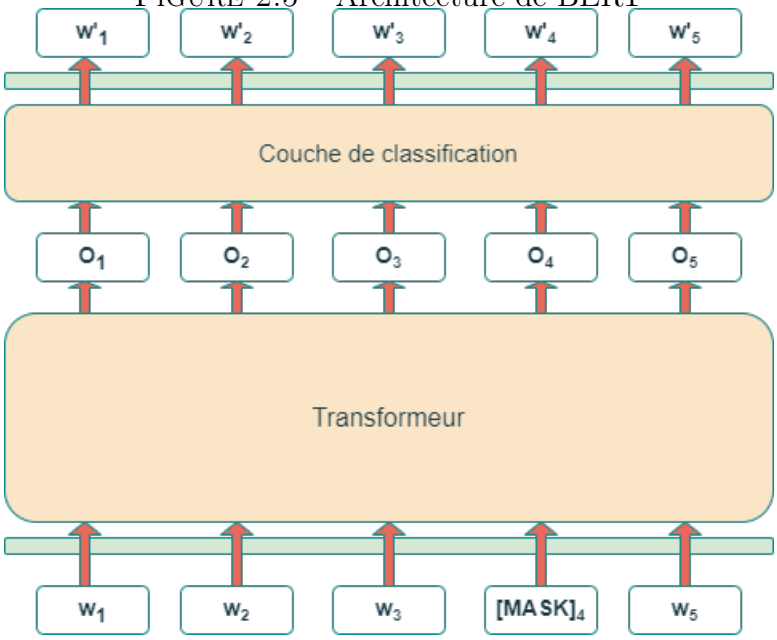
BERT

L'algorithme BERT (*Bidirectional Encoder Representations from Transformers*) est un algorithme gagnant rapidement en popularité auprès des chercheurs en traitement du langage naturel depuis sa création en 2018 par Google. La raison du gain en popularité est causée par ses performances impressionnantes en termes de classification [15]. Tel que mentionné dans le nom, BERT est basé sur l'architecture des transformateurs. Cette nouvelle technique d'apprentissage machine s'avère très ef-

ficace afin de résoudre plusieurs types de problèmes, mais plus particulièrement, le traitement du langage naturel. BERT apporte la fonctionnalité « Bidirectionnel ». En effet, cela signifie que la formule d'encodage et de décodage est appliquée sur les mots ainsi que sur leurs partenaires de gauche et de droite. Par conséquent, BERT offre de meilleures performances qu'un transformateur seul quant à la compréhension du contexte dans lequel se trouve un mot. Le modèle BERT est composé de plusieurs couches de transformateurs empilées, avec des couches d'auto-attention multi têtes à l'intérieur. Les transformateurs permettent au modèle d'apprendre et de représenter efficacement les relations entre les mots dans une séquence, sans recourir à des réseaux de neurones récurrents (RNN) ou convolutifs (CNN). Voici le fonctionnement de l'algorithme :

1. BERT est préalablement entraîné sur deux tâches d'apprentissage non supervisées : la prédiction des mots masqués (*Masked Language Model*, MLM) et la prédiction de la phrase suivante (*Next Sentence Prediction*, NSP).
2. Dans la tâche MLM, un certain pourcentage de mots d'entrée est masqué de manière aléatoire et le modèle doit prédire ces mots masqués en se basant sur le contexte des mots non masqués. Cette approche permet à BERT d'apprendre des représentations bidirectionnelles profondes et de capturer le contexte à partir des deux côtés de la séquence.
3. La tâche NSP vise à apprendre la relation entre les phrases. Le modèle reçoit deux phrases en entrée et doit prédire si la deuxième phrase suit logiquement la première dans un texte original. Cette tâche aide BERT à comprendre la structure des phrases et la relation sémantique entre elles.

FIGURE 2.5 – Architecture de BERT



4. Une fois l'entraînement préalable terminé, BERT peut être adapté à des tâches spécifiques en utilisant le transfert d'apprentissage. Cela implique d'ajouter une couche supplémentaire au-dessus du modèle BERT pré-entraîné et d'entraîner à nouveau cette nouvelle couche ainsi qu'une petite partie du modèle original sur un ensemble de données étiquetées pour la tâche spécifique. Le modèle résultant peut alors être utilisé pour résoudre cette tâche avec une performance élevée, même si l'ensemble de données est relativement petit.

Malgré son succès dans de nombreuses tâches du traitement du langage naturel, BERT présente certaines limitations. Parmi celles-ci, il est possible de citer :

1. Coût d'entraînement élevé : Les modèles BERT nécessitent une grande quantité de ressources de calcul pour l'entraînement. Cette contrainte limite leur utilisation dans des environnements à faibles ressources ou dans des applications en temps réel, tel que ce mémoire par exemple.

2. Taille du modèle : Les modèles BERT ont généralement des millions de paramètres. Cette caractéristique rend leur stockage et leur déploiement difficiles pour les appareils avec des contraintes de ressource telles que la mémoire.
3. Sensibilité aux erreurs et aux adversaires : BERT et d'autres modèles de langage profond peuvent être sensibles aux erreurs d'étiquetage, aux données d'entraînement bruitées et aux exemples adverses conçus pour tromper le modèle.
4. Quantité de données d'entraînement : Lors de l'entraînement de BERT, ce denier nécessite des milliers de données d'entrée manuellement classées par un humain. Selon le problème à résoudre, il s'agit d'une tâche très difficile à performer.

En termes de la problématique adressée, BERT semble a priori la meilleure option. Or, des chercheurs de « *Journal of the American Medical Informatics Association*» relèvent un point très important quant à l'utilisation d'algorithmes d'apprentissage automatique dans un contexte de détection du suicide : « *The prevalence of social media for sharing personal thoughts makes it a viable platform for the assessment of suicide risk. However, deep learning models are not able to capture the diverse nature of linguistic choices and temporal patterns that can be exhibited by a suicidal user on social media and end up overfitting on specific cues that are not generally applicable.*» [41]. En effet, tous les algorithmes et les articles discutés lors de cette recherche n'adressent pas la problématique de pensées suicidaires implicites. De plus, plusieurs d'entre eux simplifient ce problème en classant les textes de façon binaire (suicidaire ou non suicidaire). La question s'est avérée beaucoup plus complexe que cela.

Apprentissage à quelques coups

L'apprentissage à quelques coups (*Few-shot learning*) est une technique d'apprentissage automatique très récente. Par conséquent, le sujet dispose d'une faible source d'information dans la littérature. Il s'agit d'une technique d'apprentissage automatique visant à utiliser une faible quantité de données tout en offrant la même précision.

Cette alternative offre une solution à l'analyse d'un faible volume de données tel que le domaine d'application de ce mémoire. L'apprentissage à quelques coups est inspiré par la capacité des humains à apprendre de nouveaux concepts à partir d'un nombre limité d'exemples. Il s'agit donc d'une approche très intéressante dans les contextes où les données d'entraînement sont limitées ou coûteuses. Cette approche très récente est encore débattue dans certains articles publiés en 2022 proposant divers modèles et technique d'apprentissage. Il s'agit donc d'une branche qui évolue rapidement.

Une des approches proposées face au problème de données d'entraînement limités, est le concept d'apprentissage en contexte (*in-context learning - ICL*). Les ICL furent popularisés par l'algorithme GPT, ou plus précisément ChatGPT [36]. Bien que cette technique soit efficace, une étude par l'Université de Caroline du Nord a établi que l'apprentissage à peu de données est non seulement plus précis lors de la classification, mais demande beaucoup moins de ressources [26]. Ces chercheurs introduisent un algorithme nommé T-Few [26], basé sur le modèle populaire T0 [40], proposant le concept de réglage fin efficace des paramètres (*parameter-efficient fine-tuning - PEFT*).

Peu de temps suivant la publication de cet article, des chercheurs de H*ugging Face* en collaboration avec *cohere.ai*, *Ubiquitous Knowledge Processing Lab*, *Technical University of Darmstadt*, *Emergent AI Lab* ainsi que *Intel Labs*, ont publié un article proposant une approche qui diffère. Ceux-ci reconnaissent l'efficacité de T-Few, mais déplorent les ressources qu'exigent son utilisation, soit 11 milliards de paramètres. C'est pourquoi ceux-ci proposent l'approche basée sur les transformateurs de phrases, utilisant un nombre de paramètres 27 fois plus petit que T-Few, tout en ayant un taux de précision très similaire [49].

Transformateur de phrase

Un transformateur de phrase (*Sentence Transformer*) est un type de réseau de neurones utilisé autant dans le traitement du langage naturel que l'analyse d'image. Cependant, il est spécifiquement conçu pour traiter et comprendre des phrases ou des expressions dans le contexte du traitement du langage naturel. Les transformateurs

de phrases s'appuient sur le concept des transformateurs, un concept introduit à la sous-section 2.1.3. Plus précisément, ils s'appuient sur BERT, un concept également défini au même chapitre. Il s'agit d'un concept très récent ayant très peu de recherches qui s'y attardent. L'article d'origine de Reimers et Gurevych [38] fut publié en 2019, seulement un an suivant l'article initial de BERT [15]. L'article traite une problématique concernant BERT, surtout le temps d'exécution de l'algorithme lors de la comparaison de similarité entre deux phrases. L'article mentionne qu'afin de trouver des paires similaires dans une collection de 10 000 phrases, BERT nécessite environ 50 millions de calculs d'inférence, soit environ 65 heures de calcul [38]. Par conséquent, l'utilisation de BERT pour compléter certaines tâches non supervisées telles que le regroupement de phrases, n'est pas réaliste. Ces chercheurs présentent donc une modification de BERT intitulée « *Sentence-BERT* » (ou SBERT). À l'aide de cette modification d'algorithme, le temps d'exécution de la tâche présentée passe de 65 heures à 5 secondes avec SBERT.

Les auteurs de l'article original de SBERT comparent plusieurs stratégies de calcul de la similarité des phrases. Néanmoins, ceux-ci ont tous le même fondement. Ils utilisent BERT sur une phrase complète plutôt que sur des mots individuels composant une phrase. Ceux-ci ajoutent une opération de mise en commun (*pooling layer*) à la sortie de BERT afin de dériver l'encapsulation d'une phrase de taille fixe (mieux connue sous le nom de *embeddings*). Plusieurs techniques peuvent être utilisées à cette couche offrant des performances différentes. Les auteurs ont pris la décision de faire l'utilisation d'un réseau siamois et d'un réseau de triplet [42].

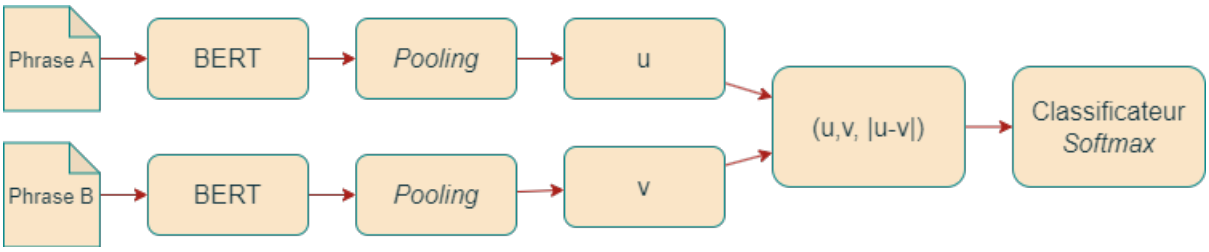
Une fois le type de réseau choisicomme par exemple le réseau siamois, la couche de mise en commun peut aussi utiliser diverses fonctions. Les auteurs font la comparaison de la fonction objective de classification, la fonction objective de régression, ainsi que la fonction objective de triplet. Si nous prenons par exemple la fonction objective de classification, les auteurs concatènent les enchâssements de phrases u et v avec la différence calculée par $|u - v|$, multipliée par le poids d'entraînement $W_t \in \mathbb{R}^{3n \times k}$:

$$o = \text{softmax}\left(W_t(u, v, |u - v|)\right)$$

où n est la dimension des enchâssements de phrases et k est le nombre d'étiquettes. Il suffit d'utiliser ensuite la fonction *softmax* afin d'optimiser l'entropie croisée.

Voici un exemple d'un réseau SBERT utilisant un réseau siamois et une fonction objective de classification :

FIGURE 2.6 – SBERT utilisant un réseau siamois



Les transformateurs de phrases offrent plusieurs avantages :

1. Calcul efficace : Le calcul préalable des enchâssements de phrases réduit le coût de calcul lors de tâches telles que la recherche sémantique où des comparaisons par paires entre les phrases sont nécessaires.
2. Apprentissage par transfert : Les enchâssements de phrases générés par des transformateurs de phrases pré-entraîné peuvent être affinés pour des tâches NLP spécifiques, ce qui permet d'obtenir de meilleures performances en utilisant moins de données d'entraînement.
3. Prise en charge multilingue : Certains transformateurs de phrases sont formés sur des corpus multilingues à grande échelle, ce qui leur permet de générer des enchâssements significatifs pour des phrases dans différentes langues.

Bien que les transformateurs de phrases soient a priori la solution idéale face à notre problématique, une contrainte majeure est toujours en jeu. Il s'agit de la quantité d'information à utiliser afin d'entraîner notre transformateur de phrase. En effet, selon l'article original des transformateurs de phrases, ceux-ci ont procédé à l'élaboration de leur algorithme en utilisant 570 000 pairs de phrases annotées avec les étiquettes de contradiction, d'exagération et de neutralité. Puisque les données

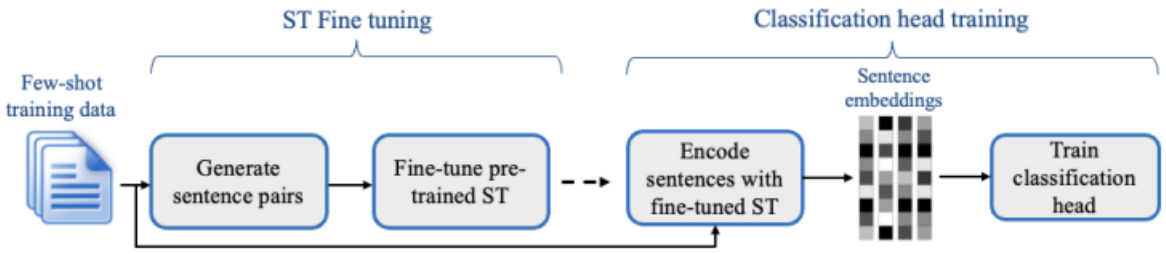
d'entraînement doivent être approuvées par un expert, il est inconcevable de générer des centaines de milliers de données d'entrée. Par conséquent, l'algorithme SetFit devient le choix idéal.

2.3.4 SetFit

L'algorithme SetFit (*Sentence Transformer Fine-tuning*) est un algorithme extrêmement récent développé par la compagnie Hugging Face [2] en collaboration avec Intel Labs [5] et UKP Lab [8]. L'article de recherche présentant cet algorithme a été publié en septembre 2022 [49]. Il s'agit d'une technique d'apprentissage automatique utilisant un transformateur de phrase pré-entraîné, qui va être ensuite ajusté à un problème choisi en utilisant un faible nombre de données d'entrée. Ce type d'approche dans la littérature est nommé l'apprentissage en quelques coups (*Few-shot learning*) [55]. Cette approche a été popularisée par GPT-3 (*Generative Pre-Trained Transformer 3*), plus précisément l'outil ChatGPT développé par OpenAI [12]. Bien que l'objectif de ChatGPT soit de créer un robot de conversation (*Chat bot*) et que l'objectif visé par ce mémoire est de faire de la classification de phrases, la même approche peut être utilisée.

SetFit est basé sur les transformateurs de phrase décrits précédemment qui sont pré-entraînés selon une base de données. Les transformateurs de phrases standards offrent des données de sortie en vecteurs très denses, pouvant ensuite être utilisés par des algorithmes d'apprentissage automatique. Une fois un transformateur de phrase sélectionné, l'algorithme SetFit propose une approche en deux étapes. La première est le raffinement d'un transformateur de phrase, puis la deuxième étape est d'entraîner un classificateur tête (*classifier head*). Ces deux étapes sont illustrées à l'aide de ce graphique :

FIGURE 2.7 – Architecture de SetFit [49]



Les détails du fonctionnement de l'algorithme sont discutés à la section 4.4.

Performance de SetFit

SetFit offre des résultats très impressionnants en termes de précision, mais aussi en termes de la grande taille du modèle. Les auteurs de l'article ont comparé SetFit avec les algorithmes de classification à peu d'essais (*few-shot classification*) les plus populaires présentement à l'aide de RAFT (*Real-World Annotated Few-shot Tasks*) [10, 3]. RAFT est un excellent outil de comparaison de ces modèles et est constamment mis à jour selon le tableau de classement [4]. SetFit Roberta, une variante de SetFit utilisant le transformateur de phrase RoBERTa [27], a réussi à obtenir une précision de 71.3%, soit seulement 2.2% sous le seul de précision d'un humain. Bien que l'algorithme similaire le plus performant soit T-Few [26] avec 75.8% de précision, ce dernier utilise un modèle de 11 milliards de paramètres, comparé à SetFit utilisant 355 millions de paramètres. T-Few, étant 27 fois plus large que SetFit, pose un problème de ressource lorsque vient l'utilisation de cet algorithme. En effet, il est recommandé d'entraîner T-Few utilisant des ordinateurs plus puissants que cette recherche le permet. Bien que SetFit soit 27 fois plus petit que T-Few et ne nécessite aucun *prompt*, il est en mesure d'obtenir des résultats très similaires.

Par conséquent, parmi tous les algorithmes explorés dans cette recherche, SetFit est définitivement l'algorithme de préférence quant à la résolution de notre problématique. La capacité de SetFit à comparer des phrases similaires, ainsi que l'utilisation

de n'importe quel transformateur de phrase et de n'importe quelle couche neuronale en tête de classification offrent une grande flexibilité de création de notre algorithme. Il est donc plus facile d'ajuster ce modèle de base à la problématique si l'approche change en cours de route. Par exemple, si la clientèle cible crée des publications sur les médias sociaux utilisant une autre langue que celle utilisée par cette recherche, il suffit de changer le transformateur de phrase et les données d'entrée. Les fondations de l'algorithme resteront les mêmes.

De plus, le niveau de précision présenté dans l'article original est très prometteur vu l'optique de cette recherche. Puisque l'algorithme ne nécessite pas un grand nombre de données d'entrée, il permet de faire face à une problématique aussi complexe que la santé mentale et le suicide en utilisant des experts du domaine afin d'entraîner l'algorithme.

Enfin, le faible coût d'entraînement permet d'effectuer cette recherche à l'aide des ressources disponibles. Certains algorithmes tels que T-Few offrant une meilleure précision demandent plus de ressources que disponible.

Chapitre 3

État de l'art préliminaire

La prévention du suicide par l'entremise de la technologie et de l'apprentissage automatique n'est pas une idée originale. En effet, il existe plusieurs recherches à son égard, proposant diverses solutions face à cette problématique. Certains chercheurs tentent d'apporter une solution à ce problème en le traitant avec une classification binaire, et d'autres avec une classification en classes multiples. Les solutions varient de simples TF-IDF (*Terme Frequency - Inverse Document Frequency*), en allant jusqu'à une solution plus complexe telle que ASHA (*Adversarial Suicide assessment Hierarchical Attention*). En somme, toutes ces approches méritent une attention en considérant une nouvelle approche plus prometteuse. La section suivante présente les études importantes face à cette problématique.

3.1 Apprentissage machine utilisant des dossiers médicaux

Publiée en 2018, l'étude [14] explore la possibilité de procéder par l'utilisation des réseaux de neurones afin d'analyser des dossiers médicaux pour identifier les personnes présentant des signes précurseurs du passage à l'acte suicidaire. La base de données nommée *Secure Anonymised Information Linkage Databank UK* et utilisée par cette recherche est composée de personnes entrées en contact avec les services de

la santé de 2001 à 2015. Les auteurs ont utilisé des réseaux de neurones de base pour différencier les cas et ont évalué la performance du système par une validation croisée répétée. Les résultats de cette recherche montrent un taux d'erreur moyen de 26,78% avec une sensibilité de 64,57% et une spécificité de 81,86%. L'approche se base sur une classification binaire, indiquant s'il y a un risque ou non (vrai ou faux). L'objectif de ce mémoire est d'utiliser une approche multiclasses en analysant plusieurs facteurs de risque. De plus, l'utilisation de réseaux de neurones plus complexes tels que SetFit apportera une approche supérieure concernant la composante de la flexibilité lors de l'identification d'un risque suicidaire.

3.2 Analyse de Reddit

La recherche publiée en 2018 [25] explore l'intelligence artificielle comme solution potentielle afin de comprendre le risque suicidaire sur les réseaux sociaux. Les recherches de l'Université de Maryland ont utilisé une base de données formée par des messages textuels de Reddit. Les résultats ont permis d'identifier plusieurs signes de risque incluant l'authenticité, les pronoms de la première personne, la négation et la posture sociale. Les chercheurs ont utilisé plusieurs algorithmes d'apprentissage automatique classique incluant le renforcement du gradient [54], les forêts aléatoires [60] ainsi que les machines à vecteur de support [61]. Cette étude montre le potentiel de l'intelligence artificielle dans l'aide aux professionnels de la santé dans leurs tâches d'estimation de la dangerosité du passage à l'acte sur les réseaux sociaux. L'approche utilise des algorithmes d'apprentissage automatique simples, ce qui rend impossible l'analyse de textes incluant les propos indirects. L'utilisation de la solution de ce mémoire prend en compte ces propos implicites dans l'estimation du risque. Or, l'article se base sur le site web Reddit, plus précisément le sous-forum "*Suicide Watch*", ce qui s'avère être une source de données des plus intéressantes.

3.3 NLP sur les réseaux sociaux

La recherche [13] explore l'utilisation d'algorithmes de traitement du langage naturel pour détecter les personnes présentant un risque de suicide en analysant leurs messages sur les médias sociaux. Elle a utilisé une base de données de messages sur les médias sociaux provenant de personnes ayant accepté de partager leurs données. Les auteurs citent le site web "OurDataHelps.com" qui n'est malheureusement plus en fonction. L'algorithme, basé sur un LSTM bidirectionnel (B-LSTM), a été optimisé pour détecter le risque de suicide au niveau des traits, ce qui permet des interventions évolutives et durables des mois précédant une crise. L'étude a montré que le langage et l'analyse linguistique peuvent jouer un rôle important dans la compréhension de la santé mentale. Néanmoins, les auteurs mentionnent un point très important indiquant que les résultats ne peuvent pas s'appliquer à tous les segments de la population. De plus, cette technologie doit être utilisée en conjonction avec les ressources existantes en matière de santé mentale pour être efficace dans la réduction des taux de suicide.

Bien que l'algorithme B-LSTM soit très prometteur, les auteurs de cet article ont traité cette problématique à l'aide de la classification binaire. De plus, tel que discuté lors dans la section ASHA, les LSTM ont tendance à s'attacher à des mots clés et à ne pas prendre en considération les messages implicites d'un individu. Par conséquent, l'utilisation d'un algorithme tel que SetFit pourrait remédier à cette situation.

3.4 Évaluation de la précision

Au cours de ce mémoire, plusieurs techniques d'apprentissage automatique seront évaluées afin d'offrir la solution optimale. L'évaluation par le biais de la mesure F1 [30] mérite un rappel. Cette métrique est définie par la formule suivante :

$$F_1 = \frac{2 \times (P \times R)}{(P + R)} \tag{3.1}$$

La précision (P) représente la proportion de prédictions vraies positives (VP - Vrai positifs) parmi le nombre total d'instances positives prédites. Elle indique la capacité du classificateur à identifier correctement les instances positives et à éviter les faux positifs (FP). La précision est calculée comme suit :

$$Precision = P = \frac{VP}{(VP + FP)} \tag{3.2}$$

Le rappel (R), également connu sous le nom de sensibilité ou de taux de vrais positifs (TVP), représente la proportion de vraies prédictions positives (VP) parmi le nombre total d'instances positives réelles. Il indique la capacité du classificateur à identifier toutes les instances positives de l'ensemble de données et à éviter les faux négatifs (FN). Le rappel est calculé comme suit :

$$Recall = R = \frac{VP}{(VP + FN)} \tag{3.3}$$

Cette technique d'évaluation permettra de choisir le meilleur transformateur de phrase ainsi que le meilleur classificateur tête.

3.5 Risque suicidaire selon les publications des réseaux sociaux

L'étude menée par Shing, H.-C. et al. [47] traite de l'utilisation de techniques de traitement du langage naturel pour évaluer le risque de suicide sur la base du contenu des médias sociaux. Les auteurs ont utilisé un ensemble de données de messages provenant du sous-forum *SuicideWatch* sur Reddit. Ils ont extrait diverses caractéristiques telles que l'enquête linguistique et le nombre de mots, les caractéristiques des émotions et le lexique des maladies mentales. Cette base de données a été analysée par quatre experts du domaine, ce qui aurait la possibilité d'être une excellente source de données, mais est malheureusement privée et est conservée par *the American Association of Suicidology.*

Ils ont ensuite utilisé des algorithmes d'apprentissage automatique supervisé tels que les machines à vecteurs de support et un réseau neuronal convolutif pour classer les utilisateurs comme présentant un risque élevé ou faible de suicide. Le modèle SVM a obtenu un score F1 de 0,46 pour l'évaluation du risque, tandis que le modèle CNN a obtenu un score F1 de 0,42.

Encore une fois, les algorithmes utilisés ici ne sont pas adéquats à la compréhension d'un message implicite d'un individu sur les réseaux sociaux. Un algorithme tel que SetFit offre un meilleur potentiel.

3.6 ASHA

ASHA (*Adversarial Suicide assessment Hierarchical Attention*) [41] est l'algorithme proposant la meilleure solution face à cette problématique délicate. Dans cette étude, les auteurs proposent un nouveau cadre d'apprentissage profond appelé apprentissage multitâches contradictoire pour améliorer les performances de l'évaluation du risque de suicide. Ce cadre combine l'apprentissage multitâches avec l'apprentissage contradictoire, ce qui permet au modèle de mieux se généraliser et d'éviter le surajustement (*Overfitting*). De cette façon, l'algorithme proposé ne va pas simplement se fier au contenu explicite de l'utilisateur de médias sociaux, mais plutôt chercher les remarques implicites sur le suicide.

ASHA se base sur les deux algorithmes analysés précédemment ayant le plus de potentiel à résoudre notre problématique, soit BERT et LSTM. Le processus se compose de trois parties principales : Post-intégration (*Post embedding*), Modélisation du contexte de l'utilisateur (*User Context Modeling*), et entraînement contradictoire (*Adversarial Training*).

Concernant le post-intégration, les auteurs utilisent le modèle BERT pour la création de jetons et pour encoder chaque message historique. Ils ajoutent un jeton [CLS] au début de chaque message et utilisent l'état caché correspondant à ce jeton comme représentation globale du message. Le processus d'encodage est représenté par l'équation suivante [41] :

$$e_t^i = \text{BERT}\left(p_t^i\right) \tag{3.4}$$

Quant à la modélisation du contexte de l'utilisateur, les auteurs codent séquentiellement les messages historiques d'un utilisateur en appliquant une couche LSTM bidirectionnelle (B-LSTM) pour capturer le contexte passé et futur d'un message. Cela permet d'extraire des modèles temporels et de prendre en compte les dépendances à long terme dans les messages. La couche B-LSTM met en correspondance les encodages historiques des messages avec les représentations contextuelles, comme le montrent les équations suivantes [41] :

$$\begin{aligned} &\overrightarrow{h_t^i} = \mathrm{LSTM}\left(e_t^i, \quad b_{t-1}^i\right), \quad \overline{h_t^i} = \mathrm{LSTM}\left(e_t^i, \quad b_{t-1}^{\bar{i}}\right) \\ &\mathbf{h}_\mathrm{t}^\mathrm{i} = \left[\overrightarrow{h_t^i}, \overrightarrow{h_t^i}\right] \end{aligned} \tag{3.5}$$

Ensuite, les auteurs utilisent un mécanisme d'attention temporelle pour appliquer des pondérations adaptatives à la représentation contextuelle de chaque message, récompensant les messages présentant des marqueurs indicatifs du risque de suicide. La représentation est désignée par a_i que l'on appelle l'exemple propre (*clean example*). La régression ordinale est utilisée pour le réseau discriminant final afin de produire des valeurs de confiance de classification pour tous les niveaux de risque de suicide, en prenant l'entrée a_i et en produisant $\widehat{y}i = Wya_i + b_y$. Le modèle est formé en minimisant une perte de régression ordinale.

Quant à l'entraînement contradictoire et dans le but d'atténuer l'impact du surajustement des mots spécifiques des idées de suicide, les auteurs appliquent l'apprentissage contradictoire. L'approche génère des exemples contradictoires en ajoutant des perturbations intentionnelles aux exemples propres dans la direction qui conduit au plus grand changement dans la prédiction du modèle. L'exemple contradictoire a_i' est généré comme suit :

$$a_i' = a_i + \varepsilon \frac{k_i}{|k_i|}, \quad k_i = \frac{\partial l\left(y_i, \widehat{y}i\right)}{\partial ai} \tag{3.6}$$

où ε est l'échelle de perturbation qui est un hyper paramètre et $\frac{k_i}{|k_i|}$ est un vecteur dans la direction qui conduit au plus grand changement dans la prédiction du modèle.

Le modèle est ensuite optimisé sur la base d'une fonction objective qui prend en compte à la fois les exemples propres et ceux contradictoires :

$$L = \sum_{i=1}^{n} L_{\text{ord}} \ (y_i, \widehat{y}i) + \gamma \sum i = 1^n L_{\text{ord}} \ \left(y_i, \widehat{y_i'}\right) \tag{3.7}$$

où $\widehat{y_i'}$ est la confiance de classification des exemples contradictoires et γ est un paramètre de contrôle pour équilibrer la perte entre les exemples propres et les exemples contradictoires qui sont des hyperparamètres.

En somme, les auteurs présentent une approche utilisant les algorithmes populaires BERT et B-LSTM afin d'estimer le risque suicidaire d'un utilisateur de médias sociaux. Or, ces deux algorithmes doivent se fonder sur une grande base de données d'entraînement. Il est difficile de se procurer une aussi vaste base de données considérant la problématique choisie. Par conséquent, l'algorithme SetFit offre une solution à ce problème tout en considérant les bienfaits de l'entraînement contradictoire approchés par ASHA.

Chapitre 4

Approche proposée

Tel que spécifié tout au long de cette proposition de mémoire, l'objectif est d'analyser des documents textuels à l'aide de nouvelles techniques d'apprentissage automatique afin de procéder à l'estimation de la dangerosité du passage à l'acte sur les réseaux sociaux. Ce chapitre propose une approche basée non seulement sur les recherches effectuées dans le domaine, mais aussi sur les discussions effectuées avec divers experts et organismes du domaine. Il présente donc les signes de risques suicidaires, l'algorithme choisi, l'architecture de l'entraînement et l'architecture de la solution.

4.1 Signe de risque suicidaire

Tel que discuté à la section 2.2, il existe une vaste variété de grilles qui proposent une estimation de la dangerosité du passage à l'acte. Par conséquent, le choix de la grille à suivre est discuté avec divers travailleurs sociaux et organismes au Québec. Ceux-ci sont d'un commun accord que la grille présentée à la section 2.2, dont la « Grille d'estimation de la dangerosité du passage à l'acte », est la meilleure option. Cette grille d'estimation (et non d'évaluation) est la norme suivie par l'ensemble des travailleurs sociaux et des organismes du Québec. Il est tout de même important de noter que cette grille est conçue dans l'objectif d'une discussion entre des travailleurs

sociaux et des personnes nécessitant de l'aide. Par conséquent, elle n'est pas conçue pour tirer seulement de l'information d'une source et d'offrir une estimation. Elle est plutôt conçue pour diriger une conversation entre un intervenant et une personne dans le besoin, pour guider l'intervenant dans sa ligne de questionnements. Cela étant dit, la solution doit baser ses critères d'estimation sur une source de données. Ce mémoire va donc s'inspirer de cette grille en tant que fondation. Une fois la solution testée, celle-ci pourra facilement être ajustée à une grille ou à des critères différents.

Les critères qui seront évalués par la solution sont donc :

- Planification du suicide,
- Antécédent de tentative de suicide,
- Capacité à espérer un changement,
- Consommation de substances,
- Impulsivité,
- Solitude et isolement,
- Capacité à prendre soin de soi.

Tel que mentionné à la section 2.2.1, la grille utilisée ici applique en outre un système de couleur, ajoutant quatre classes par signe. Néanmoins, puisque l'utilisation de cette grille dans sa totalité est complexe et requiert une formation, le système de couleur ne sera pas utilisé dans ce mémoire. Comme précisé précédemment, aux fins de ce mémoire, la grille sert simplement de base à la solution proposée afin de tenir compte de réels critères utilisés par les professionnels de la santé. Dans des travaux futurs, une équipe de chercheurs, incluant des experts sur le suicide et des experts en intelligence artificielle, pourra se rassembler afin de créer une grille spécialement adaptée pour l'analyse de profil de médias sociaux. Aux fins de cette recherche, les signes sont simplifiés sous forme de présence/absence (1 ou 0), ce qui dénote tout de même une plus grande complexité que ce que la plupart des études de l'état de l'art.

À l'aide d'un travailleur social, des phrases clés seront identifiées pour chacun des critères d'évaluation présents. Ceux-ci seront basés sur des cas réels dont le travailleur social a déjà fait l'estimation.

4.2 Algorithme de traitement du langage naturel

Pour donner suite à de nombreuses considérations, l'algorithme choisi à l'implémentation de la solution est SetFit [49]. Tel qu'expliqué dans l'étude présentant l'algorithme ASHA à la section 3.5, l'enjeu le plus important de notre problématique est la capacité de la solution donnée à détecter des intentions implicites dans certaines phrases. ASHA est une étude clé à la comparaison des objectifs et des limites du traitement du langage naturel dans le domaine d'application choisi, soit la prévention du suicide. Tandis que la majorité des études du domaine se concentre sur des mots clés situés dans une phrase afin de déterminer le risque du passage à l'acte, ASHA se concentre sur les intentions implicites d'une phrase donnée. L'étude présentant ASHA sert d'inspiration et de guide à la réalisation de cette solution. L'objectif de ce mémoire est donc d'entraîner un algorithme à détecter les intentions implicites d'un individu qu'un intervenant normal serait en mesure de comprendre, sans passer par des mots clés.

Par exemple la phrase : « Je vais me suicider vendredi ». Les intentions de cette phrase sont très évidentes et n'importe quel algorithme discuté dans l'état de l'art au troisième chapitre serait en mesure de comprendre cette phrase. Toutefois, si la phrase choisie à des intentions cachées telles que : «Vendredi, c'est la fin pour moi ». Tous les lecteurs sont en mesure de comprendre l'intention de cette phrase, étant la même que la première. Bien que cette deuxième phrase exprime exactement les mêmes intentions que la première, la majorité des algorithmes discutée dans l'état de l'art ne serait en mesure de discerner les intentions de cette phrase comme étant un risque. Dans la première phrase, les algorithmes classiques seront en mesure de détecter le mot « suicide », tandis qu'aucun mot clé est présent dans la deuxième phrase.

L'objectif de ASHA est justement de capter ces intentions implicites à l'aide de l'entraînement contradictoire. C'est pour cette raison que l'algorithme SetFit est idéal pour ce type d'apprentissage automatique. Cet algorithme est en mesure d'analyser les phrases au complet à l'aide des transformateurs de phrases afin de classer les deux phrases présentées en une seule catégorie, celle du « passage à l'acte dans un

temps rapproché », se catégorisant dans le premier critère présenté dans la section précédente.

Afin de résoudre la problématique de la sorte, un deuxième problème se présente. L'utilisation d'un ensemble de données d'entraînement d'un tel algorithme ne peut être très grande, vu la complexité de rédaction des phrases. Afin d'assurer la qualité de l'entraînement de la solution, il est important de confirmer chacune des phrases sélectionnées par un expert du domaine. Par conséquent, il n'est pas possible de créer un ensemble de données contenant des milliers d'exemples par classes. Donc, SetFit présente une fois de plus une solution idéale. Cet algorithme ne requiert qu'une dizaine d'exemples par classe afin de raffiner le transformateur de phrase [49]. Par conséquent, chacune des phrases utilisées à l'entraînement va être vérifiée par un expert de la grille, ayant précédemment procédé à l'analyse de situations réelles.

4.3 Collecte et prétraitement des données

Tel que mentionné plus haut, le grand avantage de SetFit, contrairement à certains algorithmes tels que BERT, est sa capacité d'entraîner un algorithme du traitement du langage naturel en utilisant un ensemble de données très réduite. Par conséquent, grâce à la collaboration avec un intervenant du milieu, l'algorithme est entraîné sur des phrases exemples utilisées dans un réel contexte d'estimation de la dangerosité du passage à l'acte. L'intervenant en question fait plusieurs estimations utilisant la grille, et possède donc des exemples pouvant être repris pour l'entraînement de la solution.

Les phrases vont suivre la procédure de collecte et de prétraitement suivante :

- **Collecte des données** : Obtenir les données de Reddit à l'aide de l'API Reddit. Il est possible de spécifier les sous-reddits, l'intervalle de temps et d'autres paramètres pour collecter les messages et les commentaires pertinents.
- **Nettoyage du texte** : Nettoyage des données textuelles collectées afin de supprimer les éléments inutiles et le bruit susceptible d'entraver les performances de la classification à l'aide des étapes suivantes :
  — Suppression des balises HTML, des URL et des caractères spéciaux.

— Tokenisation du texte en mots ou en phrases individuelles.

- **Traitement des cas particuliers** : Les données de Reddit peuvent contenir des éléments spécifiques qui nécessitent un traitement particulier. Aux fins de cette étude, bien que la plateforme Reddit soit déjà une plateforme anonyme, le remplacement des noms d'utilisateurs, par une chaine de caractère aléatoire, est appliqué.
- **Traitement des classes déséquilibrées** : Si l'ensemble de données comporte des classes déséquilibrées, l'utilisation de techniques telles que le sur-échantillonnage de la classe minoritaire ou le sous-échantillonnage de la classe majoritaire seront appliquées à l'ensemble des données.

Lors du déploiement, les données sont préalablement modifiées de façon similaire. En effet, lors de l'extraction des données expliquée à la sous-section suivante, les données seront normalisées en utilisant la même technique, respectant la documentation de SetFit. L'objectif est de simplement conserver une phrase pleine aussi similaire que possible aux données d'entraînement. Les détails de la source de données pour la solution finale sont fournis à la section 4.5.

4.3.1 Ensemble de données d'entraînement

Les données d'entraînement sont composées à l'aide d'exemples fournis par le travailleur social, contenant chacun des signes à détecter par l'algorithme. Selon l'article original de SetFit, seulement dix exemples de phrases par catégorie sont suffisants. Par conséquent, l'ensemble de données utilisé dans ce mémoire comprend 15 exemples fois les sept signes à détecter, soit seulement 105 phrases clés (dix pour l'entraînement et cinq pour les tests). Puisque la solution nécessite seulement 105 exemples, chacun d'entre eux est annoté par un expert du domaine. Cette tâche serait très difficile sous l'utilisation d'un autre type d'algorithme tel que BERT, nécessitant des milliers d'exemples. Une huitième catégorie nommée « autre » est ajoutée à l'ensemble de données d'entraînement afin de filtrer les phrases n'appartenant pas à une catégorie. En effet, il est possible que, lors du déploiement de la solution, une phrase ne représente pas les caractéristiques nécessaires pour appartenir à une

des catégories offertes. Par conséquent, la huitième catégorie sert de filtre afin de retirer toutes les données inopportunes à l'analyse du texte. Sans cela, SetFit force le placement de ce type de phrases et cela fausse les résultats. Cette huitième catégorie est composée de 15 phrases n'ayant aucun contexte ou aucun lien quelconque avec les autres catégories. Les huit classes sont les suivantes :

- Planification suicidaire.
  - Exemple : « J'veux pu être là, j'veux pu endurer ça. J'pense sérieusement à m'suicider. »
- Tentative antérieure de suicide.
  - Exemple : « J'ai déjà essayé d'me suicider une fois, ça s'est passé lors d'une situation semblable. »
- Présence d'un proche ou solitude.
  - Exemple : « Je n'ai personne à qui parler ou personne qui me comprend. »
- Capacité à prendre soin de soi.
  - Exemple : « C'est difficile d'maintenir une routine ou d'assumer mes responsabilités, j'ai complètement décroché(e). »
- Consommation de drogue ou alcool.
  - Exemple : « Des fois, j'me drogue pour échapper à toute cette misère. »
- Capacité à espérer du changement ou désespoir.
  - Exemple : « C'est décourageant, j'pense pas pouvoir changer les choses en mieux. »
- Capacité à se contrôler ou impulsivité.
  - Exemple : « J'ai pu d'contrôle sur moi-même, j'fais n'importe quoi. »
- Autre
  - Exemple : « J'ai récemment découvert un nouveau groupe formidable, dont la musique est unique et inspirante. »

Les exemples présentés sont donc traduits en anglais pour entraîner l'algorithme, puisque le déploiement de celui-ci est effectué sur un forum anglophone. Toutefois, il est important de noter que la solution peut facilement être ajustée à n'importe quel

langage. Il suffit de sélectionner ou d'entraîner un transformateur de phrase avec la langue choisie et générer des phrases dans la même langue.

4.4 Utilisation de SetFit

L'utilisation de SetFit s'avère assez simple. Tel que discuté en détail à la sous-section 2.3.4, il s'agit de raffiner un transformateur de phrase préalablement entraîné, puis utiliser un classificateur tête afin d'émettre des prédictions. L'approche propose de suivre les étapes ci-après :

- Création d'un ensemble de données d'entraînement approuvé par le travailleur social.
- Sélection d'un transformateur de phrase approprié au domaine d'application.
- Entraînement du classificateur tête et production de résultats.

Ces étapes sont discutées aux sections suivantes.

4.4.1 Transformateur de phrase

Ensuite, l'étape qui suit sera d'obtenir un transformateur de phrase adéquat. Puisqu'il s'agit d'un mémoire et que les ressources sont limitées, la solution fera l'utilisation d'un transformateur de phrase déjà créé. Heureusement, le site Hugging Face offre une plateforme ouverte au public contenant plusieurs types de transformateurs de phrases [6]. À l'aide de la base de données, les phrases seront évaluées par les transformateurs de phrases afin de repérer un transformateur de phrase fonctionnant adéquatement avant même le raffinement. Grâce à l'API (*Application programming interface*) de Hugging Face, plusieurs transformateurs de phrases seront testées simultanément.

Différents modèles

Chaque transformateur de phrase est entraîné d'une manière distincte et peut mener à divers résultats selon le contexte défini. Il s'agit d'une étape importante afin de sélectionner un transformateur de phrase offrant déjà des résultats prometteurs

sans même le raffinement de SetFit. La combinaison d'un transformateur de phrase adéquat et de l'ajustement de SetFit est la partie la plus importante de cette recherche. Parmi les transformateurs de phrases préalablement entraînés et offerts par Hugging Face, certains d'entre eux ont été sélectionnés, ayant le potentiel de répondre à nos besoins. Chacun d'entre eux offre une différente approche, ce qui peut influer sur un résultat. La sous-section suivante est un sous-ensemble des transformateurs de phrases les plus populaires.

Paraphrase-mpnet-base-v2

Le transformateur de phrase *paraphrase-mpnet-base-v2* [43] est le modèle contenant le plus de dimensions vectorielles comparé aux autres modèles testés. En effet, ce modèle utilise un espace vectoriel dense de 768 dimensions. Il est particulièrement intéressant puisqu'il se concentre sur la paraphrase, ce qui se rapproche de l'objectif visé. La solution cherche différentes phrases qui expriment les mêmes intentions. Le modèle n'offre pas plus de détails sur la configuration ni sur les données d'entraînement.

All-MiniLM-L6-v2

Le transformateur de phrase *all-MiniLM-L6-v2* [19] est un des modèles les plus populaires sur la plateforme *Hugging Face*. Avec, en moyenne 1.5 million de téléchargements par mois, ce modèle contient 384 dimensions vectorielles denses. Il a été entraîné en utilisant un milliard de paires d'entraînement, incluant les ensembles de données de *Reddit* [20], de *S2ORC* [28], *WikiAnswers* [16], et plusieurs autres. La liste exhaustive peut être retrouvée sur la page officielle du modèle.

Roberta-base-nli-stsb-mean-tokens

Le transformateur de phrase *roberta-base-nli-stsb-mean-tokens* [44] est entraîné sur la base de RoBERTa [27] avec le principe de S-BERT [38]. Il utilise aussi 768 dimensions vectorielles.

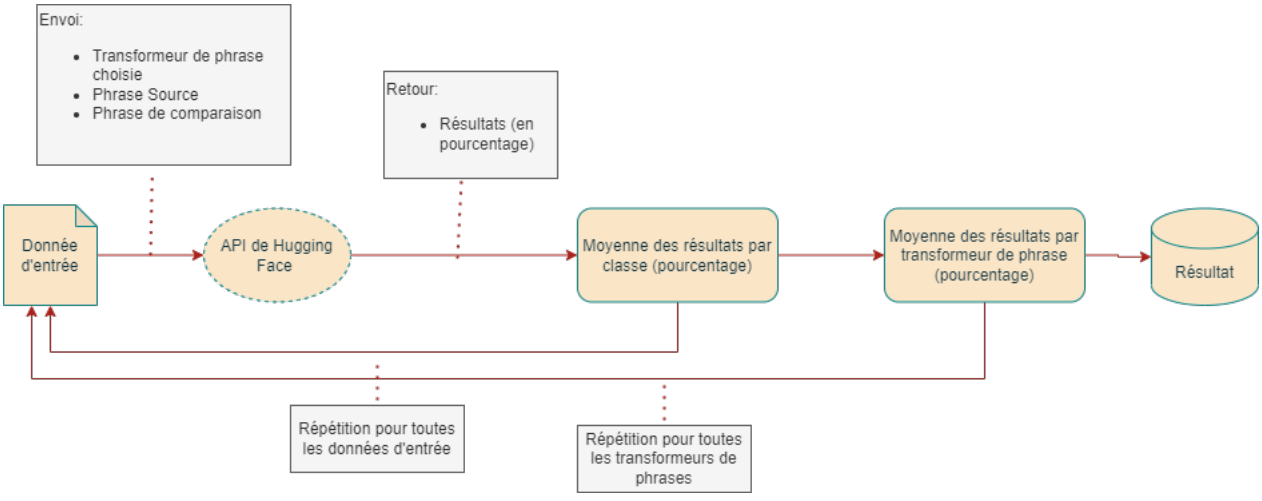
Multi-qa-distilbert-cos-v1

Le transformateur de phrase *multi-qa-distilbert-cos-v1* [45] est fondé sur 768 dimensions vectorielles et utilise *DistilBERT* comme base. Il a été entraîné sur environ 215 millions de paires de phrases.

4.4.2 Sélection d'un transformateur de phrase

Grâce à *Hugging Face,* l'évaluation des candidats sélectionnés s'avère simple puisque chaque phrase peut être téléchargée et utilisée à l'aide d'un appel API. Les données fournies par le travailleur social sont d'abord transformées dans un format JSON identifiant une phrase source par classe et une dizaine de phrases de comparaison. La phrase source est envoyée au transformateur de phrase, puis les phrases de comparaison sont envoyées une à une. Le transformateur de phrase retourne un résultat en pourcentage selon le niveau de similarité avec la phrase source. Les dix phrases sont envoyées, puis une moyenne est compilée afin de mesurer le taux de précision du transformateur de phrase avant même le raffinement de SetFit. La méthode peut être représentée par le schéma suivant :

FIGURE 4.1 – Évaluation des transformateurs de phrases



Les résultats sont ensuite compilés dans le tableau qui suit :

FIGURE 4.2 – Résultats de l'évaluation des transformateurs de phrases

Catégories\Modèle	paraphrase mp-net base v2	all MiniLM L6 v2	paraphrase MiniLM L6 v2	roberta-base-nli-stsb-mean-tokens	all-mpnet-base-v2	all-distilroberta-v1	all-MiniLM-L12-v2	multi-qa-distilbert-cos-v1
Planification du suicide	0.38	0.32	0.31	0.37	0.37	0.38	0.35	0.36
Antécédent de tentative de suicide	0.54	0.37	0.39	0.44	0.41	0.39	0.39	0.33
Capacité à espérer un changement	0.54	0.33	0.42	0.62	0.50	0.44	0.42	0.46
Consommation de substances	0.55	0.43	0.47	0.59	0.47	0.41	0.47	0.46
Impulsivité	0.41	0.41	0.42	0.24	0.44	0.43	0.37	0.45
Solitude et isolement	0.54	0.40	0.34	0.46	0.48	0.52	0.45	0.46
Capacité à prendre soin d'elle	0.54	0.39	0.46	0.47	0.40	0.43	0.38	0.44
Moyenne cumulative	0.50	0.38	0.40	0.46	0.44	0.43	0.40	0.42

Ainsi, le transformateur de phrase le plus adéquat aux fins de cette recherche est "*paraphrase mpnet base v2*" [39], obtenant un résultat d'environ 50% sans raffinement par SetFit. Cette performance peut être attribuée à la qualité de l'algorithme utilisant non seulement un vecteur à 768 dimensions, mais également due à l'utilisation de MPNet (*Masked and Permuted Pre-training for Language Understanding*) [48]. Au cours de la prochaine étape, la précision de la solution sera augmentée suivant le raffinement de SetFit.

Ajustement du transformateur de phrase

Tel que discuté en détail à la sous-section 2.3.4, le transformateur de phrase est ensuite raffiné en utilisant un réseau siamois [11]. Soit un ensemble de K exemples dans $D = (x_i, y_i)$ où x_i représentent des phrases, et y_i représentent les classes. Pour chaque classe $c \in C$, l'algorithme génère un groupe de R triplets positifs $T_c^p = (x_i, x_j, 1)$ où x_i et x_j sont des paires choisies au hasard dans la même classe c de sorte que $(y_i = y_j = c)$. Par la suite, l'algorithme utilise un groupe de triplets négatifs R de la même façon $T_c^n = (x_i, x_j, 0)$ où x_i sont des phrases de la classe c et x_j ont des phrases sélectionnées aléatoirement parmi des classes différentes, satisfaisant $(y_i = c, y_j \neq c)$. Ensuite, le groupe entraîné T est produit par le regroupement des triplets positifs et négatifs des étiquettes de classe : $T = (T_0^p, T_0^n), (T_1^p, T_1^n), \ldots, (T_{|C|}^p, T_{|C|}^n)$, où $|C|$ dénote le nombre d'étiquettes de classe, $|T| = 2 \times R \times |C|$ est le nombre total

de paires dans T, et R est un hyperparamètre. L'article présente SetFit en utilisant $R = 20$. [49]

4.4.3 Sélection du classificateur tête

Une fois le transformateur de phrase déterminé, le classificateur tête devrait, d'autant plus, être sélectionné. L'article d'origine de SetFit [49] mentionne que plusieurs types de classificateurs peuvent être utilisés à ce niveau, donc nombreux d'entre eux sont analysés afin d'obtenir le meilleur résultat possible. Par défaut, la régression logistique est utilisée. Toutefois, cette dernière est une meilleure solution lorsqu'il s'agit d'une classification binaire. C'est pourquoi elle est comparée avec : les perceptrons multicouches (MLP), les forêts aléatoires, les réseaux bayésiens, les machines à vecteur de support, le renforcement du gradient et les perceptrons linéaires. La mesure F1 est utilisée lors de l'évaluation des divers transformateurs de phrases et de classificateurs têtes, [30], expliquée en détail à la sous-section 3.4.

Lors de la sélection du classificateur tête, une tendance a été observée où la précision du modèle pouvait varier de 72% à 97% avec le même classificateur tête et les mêmes réglages. Cet écart peut être attribué au faible nombre de données d'entraînement. Puisque le nombre est réduit, la solution peut être raffinée de différentes façons à chaque exécution selon les décisions de l'algorithme et offrir des résultats très différents selon une exécution donnée. Par conséquent, afin d'obtenir un taux de précision plus réaliste, chaque classificateur tête a été entraîné 100 fois avec les mêmes réglages et le même ensemble de données. Une moyenne fut cumulée des 100 entraînements par le classificateur tête, pour un total de 700 entraînements de la solution. Voici les résultats :

FIGURE 4.3 – Résultats de l'évaluation des classificateurs tête

Phrases\Classes	Régression logistique	MLP	Forêt aléatoire	Réseau bayésien	SVM	GradientBoosting	Perceptron
Exactitude (accuracy)	0.83875	0.88625	0.859	0.8365	0.848	0.824	0.825
F1	0.83875	0.88625	0.859	0.8365	0.848	0.824	0.825

Tel qu'on peut le constater, le classificateur tête offrant la meilleure performance pour la problématique identifiée est celui des perceptrons multicouches avec une valeur de F1 d'environ 89% en moyenne.

Classificateur tête

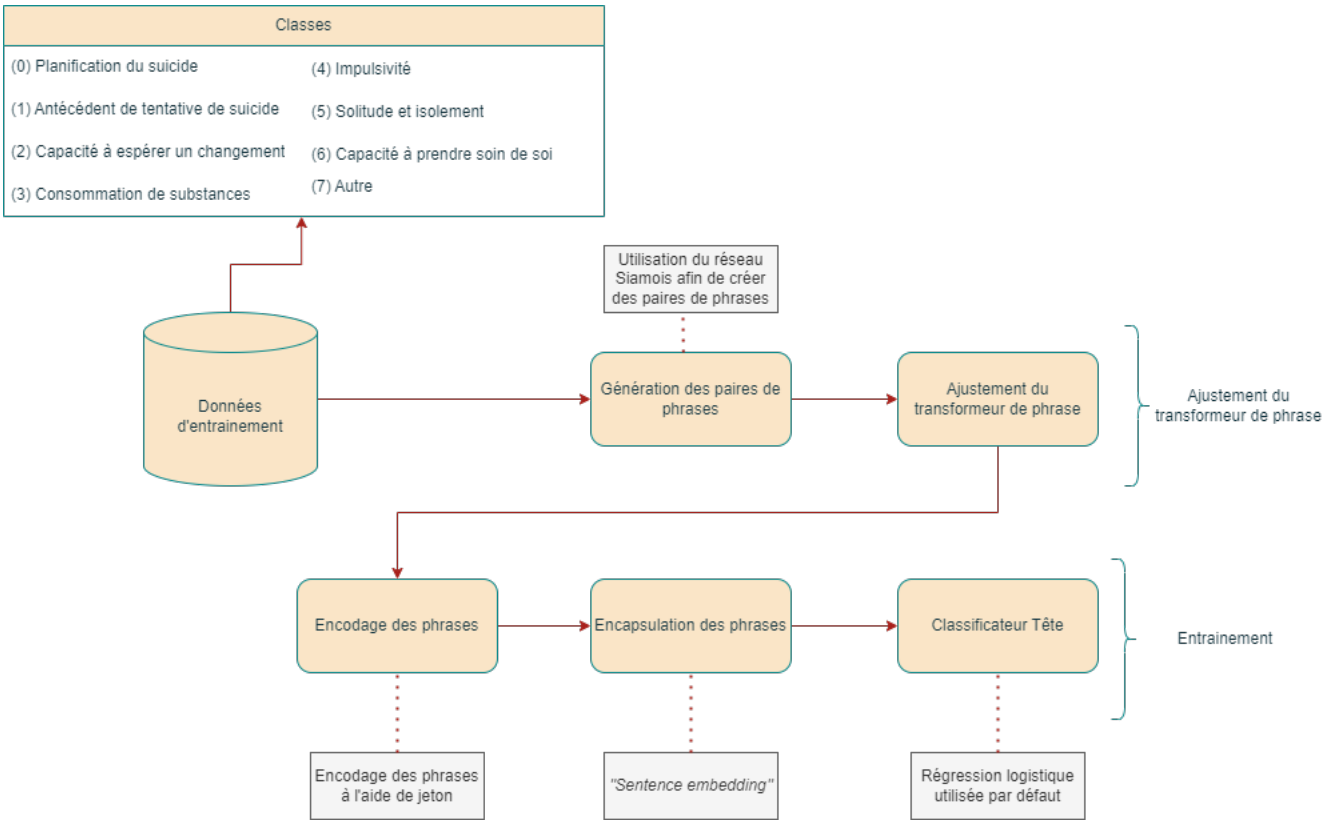
Une fois le classificateur tête et le transformateur de phrase sélectionnés, l'étape suivante est l'entraînement du classificateur tête et le raffinement du transformateur de phrase. Le ST (*Sentence Transformer*) raffiné traite les données d'apprentissage étiquetées initiales x_i, générant une représentation unique de phrase pour chaque échantillon d'apprentissage : $Emb_{x_i} = ST(x_i)$, où $ST()$ symbolise la fonction ST ajustée. Ces représentations, accompagnées de leurs étiquettes de classe, forment l'ensemble d'apprentissage pour la tête de classification $T_{CH} = (Emb_{x_i}, y_i)$, avec $|T_{CH}| = |D|$ [49]. Bien que l'article présentant cet algorithme utilise la régression logistique comme tête de classification, les auteurs mentionnent que plusieurs types de couches neuronales peuvent être utilisés à cette étape [21]. Tel que mentionné au paragraphe précédent, ce mémoire utilise les perceptrons multicouches comme classificateur tête.

4.4.4 Entraînement de SetFit

L'entraînement de SetFit survient lorsque les étapes précédentes sont terminées (création de l'ensemble de données, sélection d'un transformateur de phrase et sélec-

tion d'un classificateur tête). Par conséquent, les étapes menant à l'entraînement de SetFit ressemblent à ceci :

FIGURE 4.4 – Entraînement de SetFit



Il est important de rappeler que les détails du fonctionnement de l'algorithme de SetFit sont décrits à la sous-section 2.3.4. À la suite de l'entraînement, la solution a obtenu une précision et une valeur de F1 d'environ 89%. Ces résultats sont une grande amélioration au transformateur de phrase non raffiné ayant une précision d'environ 50%. Il s'agit d'une précision significative considérant que chaque classe contient seulement dix exemples.

4.5 Expérimentation

4.5.1 Comparaison de SetFit

Une fois la solution préparée, il s'impose une comparaison avec des algorithmes contemporains. La solution incluant chaque classificateur tête est comparée dans un tableau avec le classificateur de base seul. De plus, le classificateur LSTM a été ajouté à des fins de comparaison, puisqu'il est souvent mentionné dans l'état de l'art. Chaque classificateur de base est préalablement traité par TF-IDF [62], puisque cette technique revient aussi souvent dans l'état de l'art.

FIGURE 4.5 – Pointage F1 de SetFit contre les modèles traditionnels

Type\Classificateur	Régression logistique	MLP	Forêt aléatoire	Réseau bayésien	SVM	Renforcement de gradient	Perceptron linéaire	LSTM
Modèle traditionnel	0.54	0.58	0.42	0.54	0.38	0.46	0.38	0.08
SetFit	0.84	0.89	0.86	0.84	0.85	0.82	0.83	NA

La différence entre la précision de SetFit et celle des classificateurs de base est significative. En effet, de tels résultats peuvent être attribués à la quantité de données d'entraînement. SetFit est préparé à l'entraînement avec un faible nombre de données, tandis que les algorithmes de base ne le sont pas. Voilà la raison pour laquelle un algorithme avec une si bonne réputation tel que LSTM offre une précision de seulement 8% dans notre contexte. En effet, l'algorithme LSTM est adapté pour un entraînement utilisant un plus grand ensemble de données, et ne peut donc pas avoir une bonne performance avec un volume réduit de données.

La force de la solution présentée dans ce mémoire réside dans l'utilisation d'un LLM pré-entraîné et d'un classificateur tête. Ces derniers travaillent en synergie afin d'obtenir des résultats significatifs avec si peu de données. Dans le contexte de cette problématique, la combinaison de *paraphrase mp-net base v3* et MLP améliore grandement les résultats.

FIGURE 4.6 – Meilleure performance par modèle

Résultat\Modèle	MLP	paraphrase mp-net base v3	SetFit
Exactitude (accuracy)	58%	50%	89%
F1	58%	48%	89%

4.5.2 Expérimentation préliminaire

Afin de tester la solution avant son déploiement, une série de tests préliminaires est effectuée sur des textes tirés du sous-forum *Suicide Watch* de Reddit. Voici les textes sélectionnés à la main :

FIGURE 4.7 – Distance des classes dans une expérience préliminaire

Phrases\Classes	Planification du suicide (0)	Antécédent de tentative de suicide (1)	Capacité à espérer un changement (2)	Consommation de substances (3)	Impulsivité (4)	Solitude et isolement (5)	Capacité à prendre soin d'elle (6)	Autre (7)
I was quite proud of my plans, I was going to make it look like an accident.	0.085	0.3358	0.1028	0.0744	0.2339	0.041	0.0667	0.0602
I hate being a liability to my loved ones, and I hate feeling like this	0.0443	0.0837	0.1777	0.0481	0.2066	0.2021	0.1949	0.0426
Sometimes I imagine what my life would be like if I were no longer here.	0.277	0.1275	0.1879	0.0613	0.119	0.08	0.0931	0.0541
I'm writing this to remind myself I will eventually lose my control and take my own life if I continue to let people in	0.0632	0.0888	0.114	0.0812	0.4182	0.1231	0.0772	0.0342
Should I just give up and drink everyday?	0.0456	0.0612	0.1411	0.2828	0.1961	0.0534	0.1795	0.0402
I'll never be happy with the life I live, no matter what I do	0.0466	0.0631	0.4239	0.0524	0.1548	0.0867	0.144	0.0287
I recently stopped taking my antidepressants bc of financial reasons	0.037	0.1365	0.1536	0.1267	0.1307	0.0751	0.2882	0.0523
I loved the spiderman movie, it was so good	0.0401	0.122	0.0714	0.0646	0.0493	0.0394	0.038	0.5751

La solution accorde une table de probabilités à chacune des classes offertes. Il ne s'agit pas d'un pourcentage indépendant par classe, mais bien d'une répartition de la probabilité. Par conséquent, le pourcentage donné par phrase se situera toujours entre 0 et 1. L'addition de ces pourcentages donne toujours 1. Il est possible de constater que la première phrase est classée dans la catégorie « Antécédent de tentative de suicide » obtenant un pointage d'environ 34%. La classe sélectionnée est attribuée à la phrase ayant obtenu le pourcentage le plus élevé. Il est important de rappeler que les phrases sont écrites en anglais puisque le déploiement de la solution finale est sur un forum anglophone. Les traductions ont été effectuées à l'aide du travailleur social.

4.6 Conclusion de l'approche proposée

En conclusion du chapitre 4, l'ensemble de données fourni par l'expert du domaine a permis la sélection du transformateur de phrase offert par *Hugging Face*, ainsi que la sélection d'un classificateur tête. Ces éléments offrent une précision de 50% et 58% respectivement, avant le raffinement de SetFit. Ensuite, les expériences préliminaires montrent un taux de précision lors de l'entraînement de SetFit d'environ 89%, ce qui est très significatif vu la quantité restreinte de données d'entraînement. Par conséquent, cette approche proposée offre une analyse du traitement du langage naturel réaliste face à la problématique donnée, tout en utilisant des données d'entraînement vérifiées par un expert du domaine.

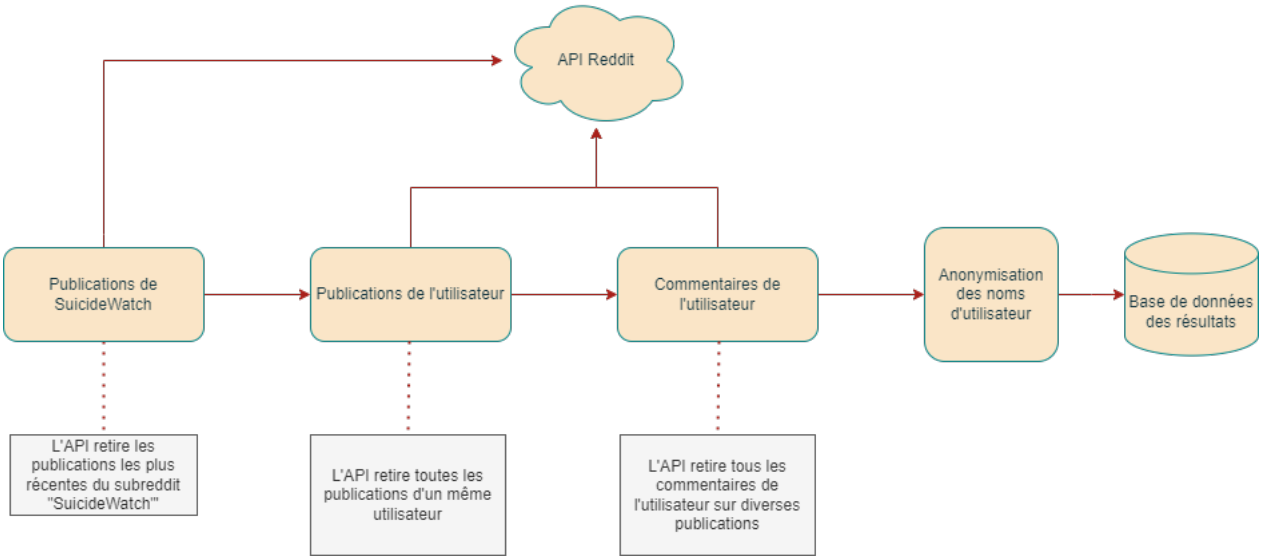
Chapitre 5

Application et quantification : analyse sur des données réelles

5.1 Création de l'ensemble de données

Une fois l'algorithme entraîné, il est intéressant d'appliquer l'algorithme sur des données réelles afin de mesurer son efficacité. Pour ce faire, l'utilisation de l'API de Reddit permettra d'obtenir des publications sur le sous-forum « *SuicideWatch* » [7]. Il est important de rappeler que ce sous-forum est utilisé à plusieurs reprises dans l'état de l'art en raison de sa grande réserve de publications relatant le suicide. Voici ci-après comment les textes sont sélectionnés.

FIGURE 5.1 – Utilisation de l'API de Reddit

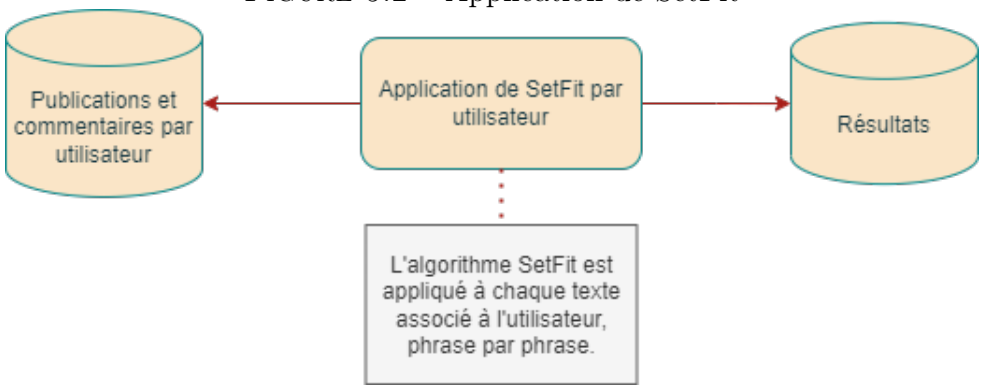


Tout d'abord, l'ensemble de données est le résultat de la compilation de publications sur le sous-forum *SuicideWatch*, selon les publications récentes, à l'aide de l'API de Reddit. Par la suite, l'API cherche toutes les publications et les commentaires laissés par l'auteur sur la même plateforme, à travers les divers sous-forums. De cette façon, l'ensemble de données comprend non seulement les publications effectuées sur le forum choisi, mais aussi toutes les autres publications de l'auteur pouvant se trouver sur d'autres forums pertinents. Cette procédure pourrait engendrer la découverte de nouveaux forums qui mériteraient une attention. Par exemple, le sous-forum "*Depression*" revient fréquemment et pourrait faire l'objet d'analyses plus précises dans le futur.

Bien que la plateforme Reddit soit composée de données publiques et anonymes, la compilation des données procède tout de même à l'anonymisation des noms d'utilisateurs. Chaque nom d'utilisateur est remplacé par une série aléatoire de chiffres et de lettres lors de l'extraction même des données par l'API. Par conséquent, les noms d'utilisateurs ne sont jamais conservés.

Une fois l'ensemble de données collecté, chaque phrase est séparée et est classée individuellement par SetFit afin de retourner le résultat final. Le traitement est illustré par la figure ci-après.

FIGURE 5.2 – Application de SetFit



L'ensemble de données comprend les informtions suivantes :

- Le type de publication.
- L'identifiant unique du texte.
- Le nom anonymisé du profil original.
- Le titre.
- Le texte complet.
- Le subreddit (sous-forum) impliqué.
- Le pointage de Reddit (nombre de votes positifs sur la publication/commentaire).
- L'URL de la publication.
- La date de publication.
- La phrase sélectionnée
- Le signe détecté (0 à 7).
- Le signe détecté en texte.

À l'aide de ce nouvel ensemble de données, la solution peut être appliquée à des données réelles afin d'évaluer sa performance dans le contexte de la problématique abordée. Celle-ci facilite la vérification individuelle de chaque estimation par un professionnel qui le désire.

5.2 Analyse sur des données réelles

5.2.1 Statistique globale

La solution a donc été appliquée suivant les indications de la section précédente. En somme, l'ensemble de données contient 5 142 auteurs distincts et 39 813 publications. Puisque la solution cherche les publications d'un auteur donné sur ces diverses plateformes, cette recherche a permis la découverte de plusieurs sous-forums pouvant être intéressants pour l'analyse de données. Au total, 5 147 sous-forums ont été identifiés au travers de l'ensemble de données créé. Parmi ceux-ci, la figure 5.3 représente les 100 sous-forums ayant le plus grand nombre de publications, et le tableau 5.1 met de l'avant les détails des dix sous-forums les plus populaires.

FIGURE 5.3 – 100 sous-forums les plus populaires

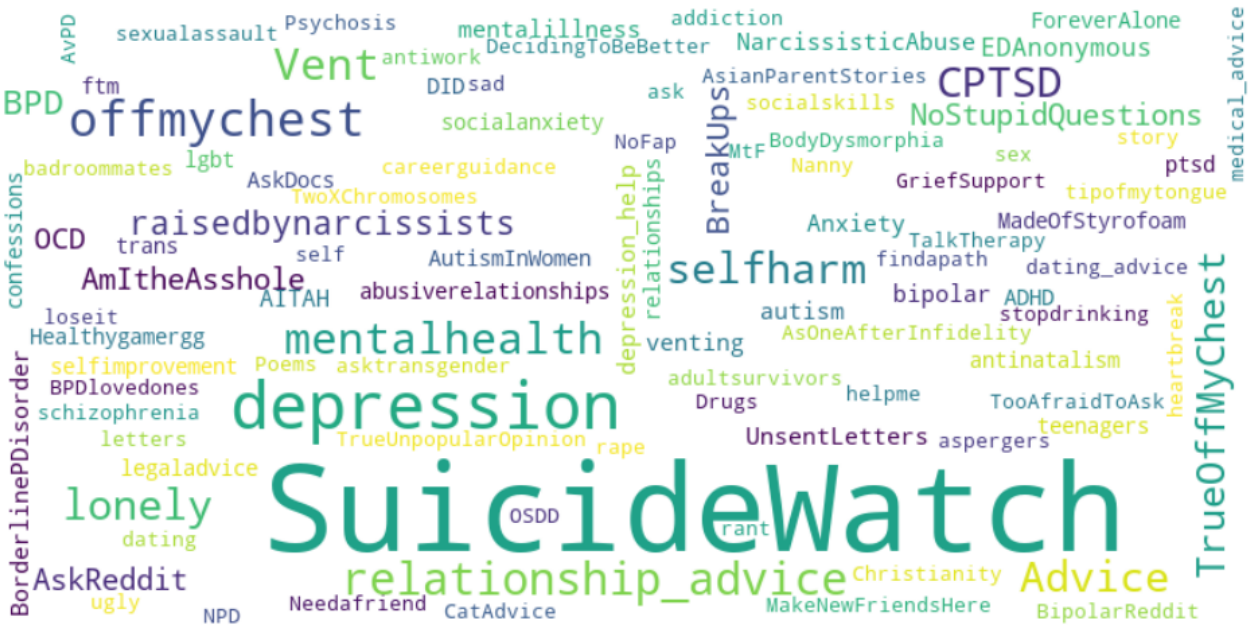


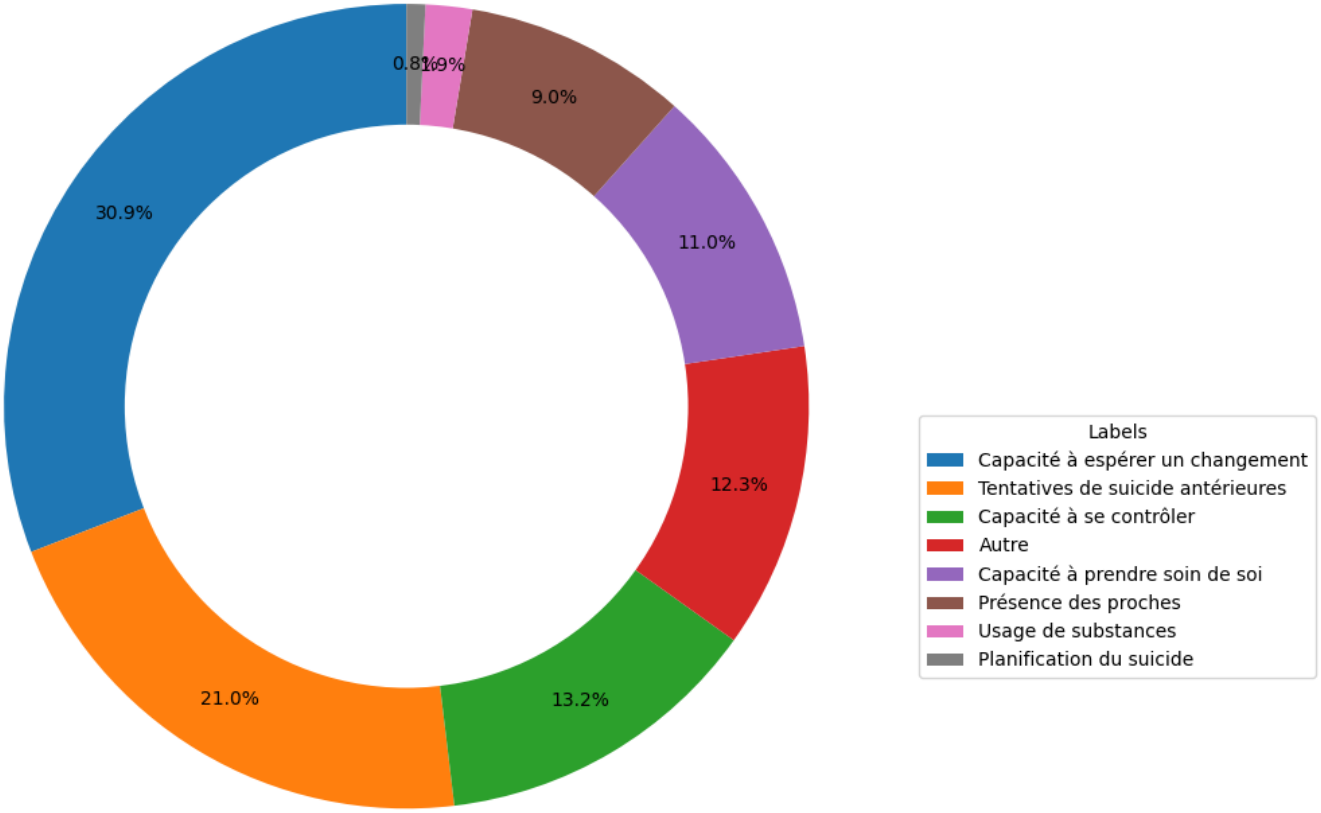
TABLE 5.1 – 10 sous-forums les plus populaires

Subreddit	Nombre de publications
SuicideWatch	73086
depression	6127
offmychest	2959
relationship_advice	2610
mentalhealth	2491
Vent	2488
Advice	2290
lonely	2250
CPTSD	2242
selfharm	2168

Ces manipulations permettent de découvrir plusieurs sous-forums connexes où les utilisateurs ont tendance à publier des informations similaires aux premiers sous-forums. Les noms de ces « *subreddits* » montrent des sujets similaires à l'original, tel que « *depression* », « *off my chest* » ou « *mental health* ». Par conséquent, ceux-ci pourraient être considérés comme candidats potentiels lors d'analyses futures.

De plus, l'analyse complète de ce nouvel ensemble de données permet d'examiner des tendances sur les réseaux sociaux quant aux signes précurseurs du passage à l'acte. La figure 5.4 illustre la répartition des classes détectées lors de l'analyse de l'ensemble de données collecté.

FIGURE 5.4 – Distribution des classes



La solution montre que le signe le plus répandu au travers des publications est la « capacité à espérer un changement », avec environ 30.9%. La « planification du suicide » est en fait le signe le moins fréquent d'entre tous avec seulement 0.8%. De plus, seulement 12.3% des phrases analysées font partie de la classe « autre », étant des phrases n'ayant aucun lien avec les signes recherchés.

5.2.2 Un auteur spécifique

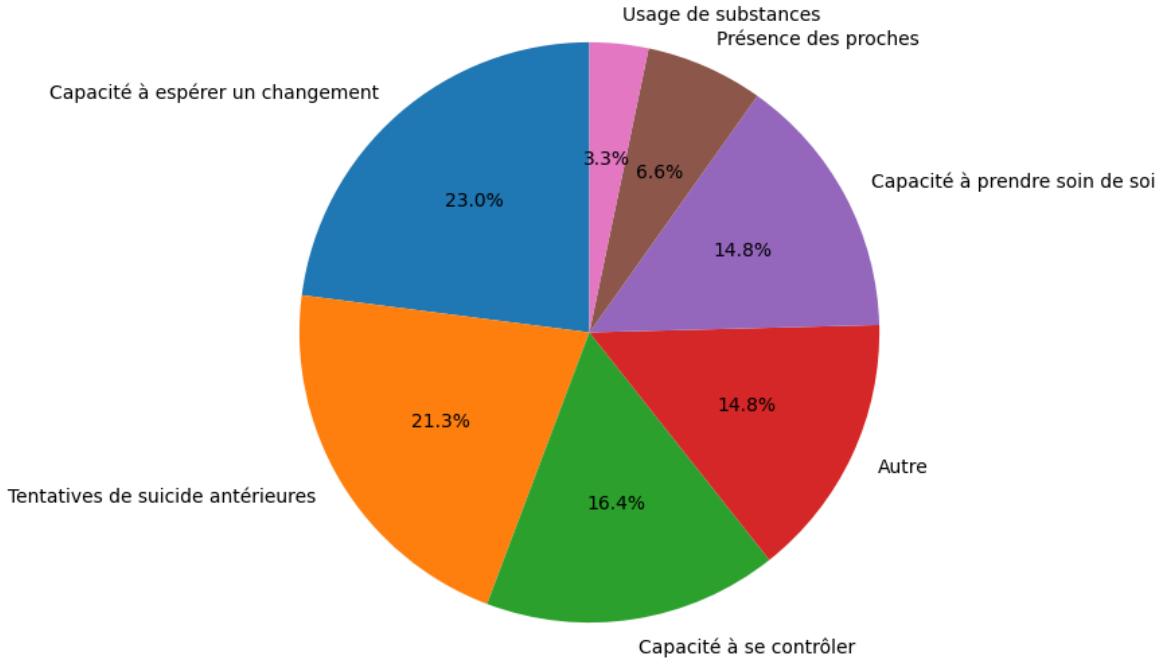
La solution prévoit aussi la capacité à analyser les publications d'un individu spécifique. À des fins d'analyse, l'auteur « *69f2597b* » (le nom original de l'auteur est anonymisé par la solution) a été choisi aléatoirement. Le tableau 5.2 montre un résumé des textes obtenus originaires du même auteur.

TABLE 5.2 – Résumé de l'auteur anonyme *69f2597b*

Auteur	Titre de la publication	ID	Classe
69f2597b	"Think about how other people will feel."	15ab7yx	[2, 4]
69f2597b	*Sight :** You gain telekinetic powers, and are...	14zgcwn	[7, 4, 6, 2, 4, 4, 4, 7, 7, 4, 3, 6, 7, 7, 4, ...]
69f2597b	About a month ago I had some really bad person...	10z66yg	[6, 1, 1, 4, 6, 6, 4, 5]
69f2597b	I moved towns for this job. Moved into a shitt...	15fl1x0	[1, 6, 6, 1, 2, 2, 2, 6, 2, 2, 1]
69f2597b	It's a Saturday night, a student nights out	11429q1	[5, 5, 2, 2]
69f2597b	My sister's partner has been doing home haircu...	145xqb2	[7, 5]
69f2597b	Never had to put so much of my energy to just ...	10phhlr	[1]

De plus, la solution permet la représentation visuelle des signes précurseurs détectés chez l'auteur parmi toutes les publications publiques. La figure 5.5 illustre la répartition des classes de cet auteur.

FIGURE 5.5 – Distribution des classes de l'auteur *69f2597b*



Tel qu'illustré par la figure 5.5, l'auteur *69f2597b* montre plusieurs signes de difficulté à espérer un changement et souligne également des tentatives de suicide antérieures. À cette étape, des démarches pourraient être entreprises à l'égard de l'auteur selon les recommandations des professionnels de la santé.

Les démarches exactes à prendre dans une telle situation relèvent des experts en santé mentale et sont hors de la portée de ce mémoire. Par contre, certains experts rencontrés proposent, par exemple, des publicités incitatives d'organismes qui sont situés près de l'individu et qui sont, par le fait même, en lien avec le type de signe détecté. Dans le cas de l'auteur sélectionné, des publicités ciblées concernant des organismes tels que le centre d'aide 24/7 [9] seraient beaucoup plus utiles à l'auteur qu'une publicité des AA (Alcoolique Anonyme) [35], puisque ce dernier n'a que 3.3% des publications en lien avec l'usage de substances. Selon la solution, l'auteur pré-

sente des signes de difficulté à espérer un changement et de tentatives antérieures de suicide. Il s'agit d'une situation dans laquelle le centre d'aide 24/7 pourrait prendre en charge l'individu. Un auteur présentant des signes de planification du suicide pourrait recevoir des publicités incitatives de lignes d'urgence dans une situation semblable, telle que le 911.

Appliquer cette solution en temps réel sur une plateforme cible pourrait procéder de la même façon sur des millions d'auteurs simultanément. En effet, une fois la solution entraînée, celle-ci peut analyser un texte de 1200 phrases (environ 20 000 mots) en seulement 2.12 secondes, sur une carte graphique *RTX 3090*. Les coûts d'un déploiement à grande échelle pourraient être le sujet d'une seconde étude incluant du matériel plus diversifié.

Chapitre 6

Conclusion

En somme, ce mémoire aborde une question de société cruciale, à savoir la prévention du suicide. À l'aide de l'algorithme SetFit, nous avons été en mesure de créer une solution utilisant des techniques de traitement du langage naturel pour développer un modèle de classification multiclasses. Les données d'entraînement sont basées sur la grille d'estimation de la dangerosité du passage à l'acte, fournie par le gouvernement du Québec.

L'objectif principal de ce mémoire est de créer un filtre pour les professionnels de la santé, permettant ainsi d'analyser rapidement les médias sociaux. Dans le futur, les critères d'estimation ainsi que la plateforme cible pourront être ajustés à la demande.

Dans un premier temps, un rappel des concepts de base nécessaires à la compréhension de la solution finale est présenté dans ce mémoire, ainsi qu'une description du domaine d'application. L'exploration de diverses approches par différents chercheurs est décrite afin d'orienter la solution de ce mémoire vers une nouvelle approche.

La contribution de cette recherche réside dans l'offre d'une solution à la problématique du suicide annoncé sur les réseaux sociaux, grâce aux nouveaux algorithmes développés récemment en apprentissage automatique par la communauté de recherche en traitement du langage naturel. Le résultat final est une solution réaliste aux organismes du domaine d'application, tout en étant flexible à la modification vers une différente plateforme, une différente langue ou une différente grille d'estimation.

L'approche proposée offre une bonne et prometteuse précision d'environ 89%, considérant la quantité réduite des données d'entraînement. À des fins de validation, elle est comparée à divers classificateurs tels que TF-IDF, la régression logistique, les forêts aléatoires, la classification naïve bayésienne, LSTM et SVM. De plus, le travail comprend la création d'un ensemble de données à l'aide de l'API de *Reddit* rassemblant plusieurs textes publiés par divers utilisateurs dont l'anonymat est garanti. Ensuite, on utilise SetFit sur ces données afin d'offrir une estimation de la dangerosité du passage à l'acte par profil en analysant phrase par phrase.

Les travaux futurs pourraient considérer la création d'une grille d'estimation de la dangerosité du passage à l'acte mise en place spécifiquement pour l'analyse de profils de médias sociaux en collaboration avec des experts en santé mentale. Une telle grille permettrait une classification plus adaptée que celle utilisée dans ce mémoire. De plus, l'élaboration d'un transformateur de phrase en français pourrait être planifiée en utilisant la technique de *paraphrase-mpnet-base-v2*, ce qui permettrait de faire l'analyse de textes francophones. Par ailleurs, le déploiement de cette solution dans un contexte réel devra être fait en collaboration avec des experts d'un organisme du domaine en suivant leurs indications et leurs conseils du fait que le sujet est délicat et les experts sont bien qualifiés pour l'application de cette solution dans le monde réel.

Même si notre démarche a été appliquée à l'estimation du risque suicidaire à partir de textes postés sur les réseaux sociaux qui sont de plus en plus utilisés pour exprimer, entre autres, des actes futurs, nous croyons qu'elle peut être appliquée à d'autres situations sociales comme la prédiction d'un acte malveillant, d'un décrochage scolaire, d'un événement familial (ex. fugue, divorce), etc.

Bibliographie

[1] Glue benchmark leaderboard. https://gluebenchmark.com/leaderboard, 2023.

[2] Hugging face – the ai community building the future. https://huggingface.co/, 2023.

[3] Raft - elicit. https://raft.elicit.org/, 2023.

[4] Raft leaderboard - hugging face spaces. https://huggingface.co/spaces/ought/raft-leaderboard, 2023.

[5] Research at intel. https://www.intel.com/content/www/us/en/research/overview.html, 2023.

[6] Sentence transformers on hugging face. https://huggingface.co/sentence-transformers, 2023.

[7] Suicidewatch. https://www.reddit.com/r/SuicideWatch/, 2023.

[8] Ukp lab - ubiquitous knowledge processing lab. https://www.informatik.tu-darmstadt.de/ukp/ukp_home/index.en.jsp, 2023.

[9] 24/7, C. D. Centre d'aide 24/7 | crise suicidaire. https://centredaide247.com/, 2023. Accessed : 2023-08-23.

[10] Alex, N., Lifland, E., Tunstall, L., Thakur, A., Maham, P., Riedel, C. J., Hine, E., Ashurst, C., Sedille, P., Carlier, A., Noetel, M., and Stuhlmüller, A. RAFT : A real-world few-shot text classification benchmark. *CoRR abs/2109.14076* (2021).

[11] Bromley, J., Guyon, I., LeCun, Y., Säckinger, E., and Shah, R. Signature verification using a "siamese" time delay neural network. In *Advances in Neural Information Processing Systems* (1993), J. Cowan, G. Tesauro, and J. Alspector, Eds., vol. 6, Morgan-Kaufmann.

[12] Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D. M., Wu, J., Winter, C., Hesse, C., Chen, M., Sigler, E., Litwin, M., Gray, S., Chess, B., Clark, J., Berner, C., McCandlish, S., Radford, A., Sutskever, I., and Amodei, D. Language models are few-shot learners. *CoRR abs/2005.14165* (2020).

[13] Coppersmith, G., Leary, R., Crutchley, P., and Fine, A. Natural language processing of social media as screening for suicide risk. *Biomedical Informatics Insights 10* (2018), 1178222618792860. PMID : 30158822.

[14] DelPozo-Banos, M., John, A., Petkov, N., Berridge, D. M., Southern, K., LLoyd, K., Jones, C., Spencer, S., and Travieso, C. M. Using neural networks with routine health records to identify suicide risk : Feasibility study. *JMIR Mental Health 5*, 2 (2018).

[15] Devlin, J., Chang, M., Lee, K., and Toutanova, K. BERT : pre-training of deep bidirectional transformers for language understanding. *CoRR abs/1810.04805* (2018).

[16] Fader, A., Zettlemoyer, L., and Etzioni, O. Open question answering over curated and extracted knowledge bases. In *Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (New York, NY, USA, 2014), KDD '14, Association for Computing Machinery, p. 1156–1165.

[17] for Disease Control, C., and Prevention. Increase in suicide mortality in the united states, 1999–2019. `https://www.cdc.gov/nchs/products/databriefs/db433.htm`, 2022.

[18] Gao, T., Yao, X., and Chen, D. Simcse : Simple contrastive learning of sentence embeddings. *CoRR abs/2104.08821* (2021).

[19] Henderson, M., Budzianowski, P., Casanueva, I., Coope, S., Gerz, D., Kumar, G., Mrksic, N., Spithourakis, G., Su, P., Vulic, I., and Wen, T. A repository of conversational datasets. *CoRR abs/1904.06472* (2019).

[20] Henderson, M., Budzianowski, P., Casanueva, I., Coope, S., Gerz, D., Kumar, G., Mrkšić, N., Spithourakis, G., Su, P.-H., Vulić, I., and Wen, T.-H. A repository of conversational datasets, 2019.

[21] HuggingFace. Efficient few-shot learning with sentence transformers. `https://www.youtube.com/watch?v=8h27lV8v8BU&t=960s`, 2022. Accessed : Minute 16 :00.

[22] JavaTpoint. Linear regression vs logistic regression in machine learning. `https://www.javatpoint.com/linear-regression-vs-logistic-regression-in-machine-learning`, 2023.

[23] Lafleur, C., and Séguin, M. *Intervenir en situation de crise suicidaire*. Les Presses de l'Univertisé Laval, 2008.

[24] Lane, J., Archambault, J., Collins-Poulette, M., and Camirand, R. Guide de bonnes pratiques en prévention du suicide à l'intention des intervenants des centres de santé et de services sociaux. `https://publications.msss.gouv.qc.ca/msss/fichiers/2010/10-247-02.pdf`, 2010.

[25] Lao, C., Lane, J., and Suominen, H. Analyzing suicide risk from linguistic features in social media : Evaluation study. *JMIR Form Res 6*, 8 (Aug 2022), e35563.

[26] Liu, H., Tam, D., Muqeeth, M., Mohta, J., Huang, T., Bansal, M., and Raffel, C. Few-shot parameter-efficient fine-tuning is better and cheaper than in-context learning, 2022.

[27] Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., and Stoyanov, V. Roberta : A robustly optimized BERT pretraining approach. *CoRR abs/1907.11692* (2019).

[28] Lo, K., Wang, L. L., Neumann, M., Kinney, R., and Weld, D. S2ORC : The semantic scholar open research corpus. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics* (Online, July 2020), Association for Computational Linguistics, pp. 4969–4983.

[29] Mann, J. J., Apter, A., Bertolote, J., Beautrais, A., Currier, D., Haas, A., Hegerl, U., Lonnqvist, J., Malone, K., Marusic, A., et al. Suicide prevention strategies : a systematic review. *Jama 294*, 16 (2005), 2064–2074.

[30] Manning, C. D., Raghavan, P., and Schütze, H. *Introduction to Information Retrieval*. Cambridge University Press, 2008.

[31] Mohammad, S., Kiritchenko, S., and Zhu, X. NRC-Canada : Building the state-of-the-art in sentiment analysis of tweets. In *Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 2 : Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013)* (Atlanta, Georgia, USA, June 2013), Association for Computational Linguistics, pp. 321–327.

[32] of Canada, M. H. C., and Institute, C. P. S. Suicide risk assessment toolkit. `https://suicideprevention.ca/wp-content/uploads/2021/08/MHCC-CPSI-Suicide-Risk-Assessment-Toolkit-EN.pdf`, 2021.

[33] Organization, W. H. Preventing preventing suicide suicide a global imperative, 2014.

[34] Organization, W. H. World health statistics 2016 : monitoring health for the sdgs sustainable development goals, 2016.

[35] Québec, A. Aa québec | aide téléphonique. `https://aa-quebec.org/aaqc_wp/`, 2023. Accessed : 2023-08-23.

[36] Radford, A., Narasimhan, K., Salimans, T., and Sutskever, I. Improving language understanding by generative pre-training.

[37] Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., and Liu, P. J. Exploring the limits of transfer learning with a unified text-to-text transformer. *CoRR abs/1910.10683* (2019).

[38] Reimers, N., and Gurevych, I. Sentence-bert : Sentence embeddings using siamese bert-networks. *CoRR abs/1908.10084* (2019).

[39] Reimers, N., and Gurevych, I. Sentence-bert : Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing* (11 2019), Association for Computational Linguistics.

[40] Sanh, V., Webson, A., Raffel, C., Bach, S. H., Sutawika, L., Alyafeai, Z., Chaffin, A., Stiegler, A., Scao, T. L., Raja, A., Dey, M., Bari, M. S., Xu, C., Thakker, U., Sharma, S., Szczechla, E., Kim, T., Chhablani, G., Nayak, N. V., Datta, D., Chang, J., Jiang, M. T., Wang, H., Manica, M., Shen, S., Yong, Z. X., Pandey, H., Bawden, R., Wang, T., Neeraj, T., Rozen, J., Sharma, A., Santilli, A., Févry, T., Fries, J. A., Teehan, R., Biderman, S., Gao, L., Bers, T., Wolf, T., and Rush, A. M. Multitask prompted training enables zero-shot task generalization. *CoRR abs/2110.08207* (2021).

[41] Sawhney, R., Joshi, H., Gandhi, S., Jin, D., and Shah, R. R. Robust suicide risk assessment on social media via deep adversarial learning. *Journal of the American Medical Informatics Association 28*, 7 (2021), 1497–1506.

[42] Schroff, F., Kalenichenko, D., and Philbin, J. Facenet : A unified embedding for face recognition and clustering. *CoRR abs/1503.03832* (2015).

[43] sentence transformers. sentence-transformers/paraphrase-mpnet-base-v2. `https://huggingface.co/sentence-transformers/paraphrase-mpnet-base-v2`, 2021. [Online ; accessed 14-August-2023].

[44] sentence transformers. sentence-transformers/roberta-base-nli-stsb-mean-tokens. `https://huggingface.co/sentence-transformers/roberta-base-nli-stsb-mean-tokens`, 2021. [Online ; accessed 14-August-2023].

[45] sentence transformers. sentence-transformers/multi-qa-distilbert-cos-v1. `https://huggingface.co/sentence-transformers/multi-qa-distilbert-cos-v1`, 2023. [Online ; accessed 14-August-2023].

[46] Severyn, A., and Moschitti, A. Twitter sentiment analysis with deep convolutional neural networks. In *Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval* (2015), ACM, pp. 959–962.

[47] Shing, H.-C., Nair, S., Zirikly, A., Friedenberg, M., Daumé III, H., and Resnik, P. Expert, crowdsourced, and machine assessment of suicide risk via online postings. In *Proceedings of the fifth workshop on computational linguistics and clinical psychology : from keyboard to clinic* (2018), pp. 25–36.

[48] Song, K., Tan, X., Qin, T., Lu, J., and Liu, T.-Y. Mpnet : Masked and permuted pre-training for language understanding, 2020.

[49] Tunstall, L., Reimers, N., Jo, U. E. S., Bates, L., Korat, D., Wasserblat, M., and Pereg, O. Efficient few-shot learning without prompts, 2022.

[50] van den Oord, A., Li, Y., and Vinyals, O. Representation learning with contrastive predictive coding. *CoRR abs/1807.03748* (2018).

[51] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., and Polosukhin, I. Attention is all you need. *CoRR abs/1706.03762* (2017).

[52] Wikipedia. Artificial neural network — Wikipedia, the free encyclopedia. `https://en.wikipedia.org/wiki/Artificial_neural_network`, 2023.

[53] Wikipedia. Byte pair encoding — Wikipedia, the free encyclopedia. `https://en.wikipedia.org/wiki/Byte_pair_encoding`, 2023.

[54] Wikipedia. Gradient boosting. `https://en.wikipedia.org/wiki/Gradient_boosting`, 2023.

[55] Wikipedia. In-context learning (natural language processing) — Wikipedia, the free encyclopedia. `https://en.wikipedia.org/wiki/In-context_learning_(natural_language_processing)`, 2023.

[56] Wikipedia. Logistic regression. `https://en.wikipedia.org/wiki/Logistic_regression`, 2023.

[57] Wikipedia. Long short-term memory. https://en.wikipedia.org/wiki/Long_short-term_memory, 2023.

[58] Wikipedia. Multilayer perceptron. https://en.wikipedia.org/wiki/Multilayer_perceptron, 2023.

[59] Wikipedia. Natural language processing. https://en.wikipedia.org/wiki/Natural_language_processing, 2023. [Online; accessed 14-May-2023].

[60] Wikipedia. Random forest. https://en.wikipedia.org/wiki/Random_forest, 2023.

[61] Wikipedia. Support vector machine — Wikipedia, the free encyclopedia. https://en.wikipedia.org/wiki/Support-vector_machine, 2023.

[62] Wikipedia. Term frequency–inverse document frequency — Wikipedia, the free encyclopedia. https://en.wikipedia.org/wiki/TfâĂŞidf, 2023.

[63] Wikipedia. Triplet loss. https://en.wikipedia.org/wiki/Triplet_loss, 2023.