# Concept drift vs suicide: comment l'un peut prévenir l'autre?

Cédric Maigrot*, Sandra Bringay**,***, Jérôme Azé**

*IRISA Rennes, UMR 6074
Cedric.Maigrot@irisa.fr
**LIRMM, CNRS, UMR 5506
{Sandra.Bringay, Jerome.Aze}@lirmm.fr
***AMIS, Université de Montpellier Paul Valéry

**Résumé.** Le suicide devient d'année en année une problématique plus préoccupante. Les organismes de santé tels que l'OMS se sont engagés à réduire le nombre de suicides de 10% dans l'ensemble des pays membres d'ici 2020. Si le suicide est généralement un geste impulsif, il existe souvent des actes et des paroles qui peuvent révéler un mal être et représenter des signes précurseurs de prédispositions au suicide. L'objectif de cette étude est de mettre en place un système pour détecter semi-automatiquement ces comportements et ces paroles au travers des réseaux sociaux. Des travaux précédents ont proposé la classification de messages issus de Twitter suivant des thèmes liés au suicide : tristesse, blessures psychologiques, état mental, etc. Dans cette étude, nous ajoutons la dimension temporelle pour prendre en compte l'évolution de l'état des personnes monitorées. Nous avons implémenté pour cela différentes méthodes d'apprentissage dont une méthode originale de *concept drift*. Nous avons expérimenté avec succès cette méthode sur des données réelles issues du réseau social Facebook.

## 1 Introduction

Toutes les 40 secondes, une personne se suicide dans le monde [1]. Toutes les régions et tous les âges sont touchés, notamment les jeunes âgés de 15 à 29 ans, pour qui le suicide est la deuxième cause de mortalité à l'échelle mondiale. Nous faisons l'hypothèse dans cet article qu'il est possible de prévenir certains suicides en monitorant l'activité de ces personnes sur les réseaux sociaux (Twitter, Facebook, etc.) qu'ils utilisent pour la plupart quotidiennement.

Nous proposons une extension des travaux de Abboute et al. (2014) qui permettent d'associer un niveau de risque à un message. Dans cette étude, nous intégrons la dimension temporelle associée à l'évolution de l'état de la personne monitorée qui est capturée au travers de sa séquence de messages. Pour cela, nous avons adapté un modèle de *concept drift* Gama et al. (2014) pour détecter ces changements d'états. L'objectif est de lever une alerte au plus tôt lorsque l'on détecte une évolution des messages interprétée comme négative, mais sans solliciter abusivement le médecin qui sera le seul habilité à prendre la décision d'intervenir.

1. Prévention du suicide, L'état d'urgence mondial. OMS 2014. ISBN : 978 92 4 256477 8

Notre processus est organisé en trois étapes : 1) Prétraitement et stockage des messages avec l'utilisation de différentes méthodes de TAL [2] pour extraire des descripteurs pertinents des messages ; 2) Détection du niveau de risque des messages à partir d'une méthode ensembliste de classification (Stacking). 3) Levée d'une alerte à destination du professionel de santé. La levée d'alerte est issue d'un calcul statistique sur la dérive d'un concept ou sur l'application de règles expertes ou encore sur la comparaison de courbes ROC Hanley et McNeil (1982).

Les challenges associés à ces travaux sont nombreux. Tout d'abord, les méthodes de TAL doivent être adaptées pour traiter les données particulières issues des réseaux sociaux. Les messages sont écrits de manière peu rigoureuse, sont de taille variable, contiennent des structures grammaticales non conformes, des fautes d'orthographe, des abréviations, des mots d'argot spécifiques ou non aux thèmes des forums. Notre méthode doit donc être robuste au vocabulaire, au discours ni structuré, ni prévisible dans les médias sociaux. De plus, les techniques de *concept drift* doivent détecter des concepts relativement subjectifs (*e.g.* anorexie, dépression). À notre connaissance, il n'existe pas de ressources linguistiques francophones tel que des lexiques étoffés pour capturer ces concepts. Le résultat du processus doit être clairement explicité aux professionnels de santé afin de les aider dans le processus de décision.

L'originalité de ce travail se situe à plusieurs niveaux. Tout d'abord, le travail réalisé se base sur les récits de personnes ayant des comportements à risques avérés. Il s'agit de personnes ayant posté des messages dans des groupes Facebook traitant de thématiques à risque. Par ailleurs, le processus mis en place prend en compte plusieurs niveaux de description : celui des concepts à risque évoqués dans un message (e.g. expression de la solitude, idéation suicidaire, récit de comportement anorexique, etc.), celui du niveau de risque défini selon le protocole de notre partenaire (l'Organisation *Hope for Education* travaillant avec des personnes harcelées très à risque) et celui de l'alerte à transmettre ou non à l'équipe médicale chargée du monitoring. Troisièmement, la levée d'alerte, c'est-à-dire la remontée d'une information sur l'évolution de l'état du patient (en bien ou en mal) pour le médecin, repose sur trois approches dont une basée sur l'observation du désaccord entre plusieurs classifieurs.

L'article est organisé comme suit. La section 2 présente les travaux réalisés en lien avec notre étude. La section 3 présente la méthodologie mise en place. La section 4 présente le protocole expérimental mit en place. Enfin, la section 5 présente les résultats obtenus.

# 2 État de l'art et motivations

**Fouille du web social pour des perspectives médicales.** La santé est un domaine où la fouille des réseaux sociaux permet d'envisager de réelles perspectives médicales. Depuis 2013, 2 milliards d'utilisateurs sont actifs dans les réseaux sociaux. Facebook est le plus visité avec 1,2 milliards d'individus, suivi par d'autres réseaux, dont Twitter avec 225 millions. Ces réseaux sont très utilisés pour partager des pensées, opinions et émotions avec ses proches, notamment par les populations jeunes. Au cours des cinq dernières années, il y a eu un intérêt croissant pour exploiter ces réseaux comme un outil pour la santé publique, par exemple pour analyser la propagation de la grippe Sadilek et al. (2012). Paul et Dredze (2011) utilisent des modèles thématiques pour capturer les symptômes et les traitements possibles pour des maux évoqués sur Twitter afin de définir des mesures de santé publique. En examinant manuellement un grand

2. TAL : Traitement Automatique de la Langue

nombre de tweets, Krieck et al. (2011) ont montré que les symptômes auto-déclarés sont le signal le plus fiable pour prévoir l'apparition d'une maladie.

**Fouille du web social pour la détection des maladies mentales et des personnes suicidaires.** Dans le cadre de cette étude, nous nous concentrons sur le potentiel des réseaux sociaux pour surveiller les populations à risque suicidaire. D'autres recherches portant sur les maladies mentales en général existentChoudhury et al. (2013), Dinakar et al. (2014). Dans ces applications, une personne est considérée comme à risque selon son utilisation des médias sociaux. Par exemple, le contenu de leurs tweets, les mises à jour de leurs statuts Facebook, sont utilisés pour classer en temps réel les personnes selon des niveaux de risque. Moreno et al. (2011) ont démontré que des mises à jour de statut sur Facebook révèlent des symptômes d'épisodes dépressifs majeurs. Tous ces travaux soulignent le potentiel des médias sociaux comme une source de signaux pour la maladie mentale. Plus précisément, autour de la thématique du suicide, Cash et al. (2013) analysent les messages d'adolescents sur MySpace.com afin de déterminer les sujets à risque (relations, santé mentale, toxicomanie/abus, méthodes de suicide, déclarations sans contexte). Jashinsky et al. (2014) étudient les facteurs de risque de suicide évoqués sur Twitter et trouvent une forte corrélation entre les données Twitter dérivées et les données sur le suicide, ajustées selon l'âge. Christensen et al. (2014) soulignent que des interventions sont possibles, bien que la validité, la faisabilité et la mise en œuvre restent incertaines car peu d'études ont été menées à ce jour dans des conditions réelles.

# 3 Processus global de surveillance des individus

Le processus de surveillance des patients est inspiré globalement des méthodes de *concept drift* qui permettent de capturer l'apparition de nouveaux concepts au cours du temps. Nous avons repris l'architecture de Gama et al. (2014) que nous avons modifiée pour intégrer le fait que nous ne connaissons pas la véritable nature des données (niveau de risque à prédire) sans intervention du professionnel de santé et pour intégrer des règles expertes pour la levée d'alerte. Le processus est organisé en 3 étapes indépendantes décrites ci dessous.

Les deux premiers modules font une analyse individuelle des messages, le troisième analyse les derniers messages d'un même utilisateur pour lever ou non une alerte.

## 3.1 Étape 1 : mémorisation des messages et prétraitements

L'objectif de cette étape est de mémoriser pour chaque personne monitorée : 1) la thématique du groupe Facebook auquel elle appartient (e.g. tentative de suicide, harcèlement, anorexie, ...) ; 2) le contenu de ses messages, la date et l'heure de l'écriture des messages, le nombre de mentions *likes*, le nombre de commentaires ; 3) les commentaires associés à un message. Il est important de noter que cette démarche est aussi applicable à de nombreux autres réseaux sociaux (e.g Twitter, Instagram, Ask, ...).

Comme indiqué par Balahur (2013), les textes issus des réseaux sociaux ont des particularités linguistiques qui peuvent influencer les performances de la classification. Pour cette raison nous avons appliqué les prétraitements suivants : 1) remplacement des noms d'utilisateurs par [NOM] ; 2) remplacement des adresses mails par [MAIL] ; 3) remplacement des adresses URL

par [URL] ; 4) remplacement des émoticônes par un mot d'humeur associé (table de correspondance créée pour l'étude) ; 5) remplacement des abréviations par le(s) mot(s) complet(s) (table de correspondance créée pour l'étude) ; 6) suppression des accents et des majuscules. En effet, afin d'écrire plus vite, ces deux conventions d'écriture ne sont pas toujours utilisées. Leur normalisation en minuscules sans accent permet de restreindre le nombre de N-Grammes de mots générés et ainsi faire le lien entre des messages ne présentant précédemment aucun N-Grammes en commun ; 7) lemmatisation avec l'outil TreeTagger Schmid (1994).

## 3.2 Étape 2 : détection du niveau de risque en 2 sous étapes

Nous avons choisi de travailler en respectant le paradigme de l'*ensemble learning* qui consiste à apprendre plusieurs classifieurs pour résoudre le même problème. Nous nous sommes placés dans ce paradigme pour les raisons suivantes : 1) la combinaison de classifieurs donne en général de meilleurs résultats Wang et al. (2003) ; 2) la puissance de calcul à notre disposition rend accessible l'apprentissage et l'utilisation de nombreux modèles ; 3) les masses de données à traiter pourraient s'avèrer non apprenables par un unique classifieur ; 4) le besoin d'expliciter les résultats de la classification pour les humains impliqués dans le processus de prise de décision. Diverses formes d'ensemble learning existent : *Stacking*, *Boosting* et *Bagging*. Dans notre contexte, nous avons choisi l'approche *Stacking* qui consiste à apprendre une succession de classifieurs organisés en deux niveaux et agrégés selon un vote majoritaire et tels que chaque classifieur apprenne de nouveaux descripteurs permettant de redécrire les données. Nous détaillons dans la suite ces deux sous-étapes.

**Sous étape 1 : détection des concepts dans un message.** Il s'agit de repérer dans les messages un premier niveau d'information : la présence ou l'absence d'un signal de mal-être que nous nommerons par la suite *concept*. La liste des *concepts* considérés est : *précédente tentative de suicide*, *idéations suicidaires*, *dépression*, *harcèlement*, *prise de médicaments*, *problème d'alimentation* (anorexie ou boulimie), *auto-mutilation*, *colère*, *peur*, *solitude*, *tristesse* et *rémission*. Cette liste issue des travaux de Plutchik et Van Praag (1994) a été validée par des experts psychiatres. Pour chaque concept, un classifieur renvoie la valeur *oui* pour associer la présence du concept dans le message. Nous avons choisi les descripteurs résumés ci-dessous, complétés avec des lexiques spécifiques pour chaque concept : 1) Contenu du message décrit par les mesures statistiques suivantes : a) un ensemble de N-grammes pour permettre une comparaison entre les messages sélectionnés selon la mesure *TF-IDF* pour ne conserver que les mots discriminants par concepts ; b) le nombre de mots associés à un concept donné par un lexique réalisé pour l'étude ; 2) Nombre de commentaires : un message à risque est suceptible de soulever beaucoup de réactions auprès des autres membres du groupe ; 3) Nombre de mentions *Likes* : à l'inverse, un message où la victime évoque son bien-être ou son rétablissement peut être accompagné de beaucoup de mentions *Likes* ; 4) Longueur du message : deux comportements opposés des victimes sont connus. La victime peut écrire des messages plus longs par besoin de se confier aux autres ou au contraire se refermer sur elle-même et moins écrire.

En sortie de cette étape, un message est associé à un vecteur de concepts booléens. Par exemple, le message "*Je vais tres mal, aidez moi vite svp. Je me fais harcelee depuis 3ans et me fais frapper tous les jours. Je mappelle lea, jai 14ans et vis a roubaix*" est associé au vecteur ( *tentative de suicide*, *idéations suicidaires*, *dépression*, *harcèlement*, *médicaments*, *anorexie*, *mutilation*, *colère*, *peur*, *solitude*, *tristesse*, *rémission* ).

**Sous étape 2 : calcul du niveau de risque pour un message** Les six classifieurs sont ensuite appliqués à la prédiction du niveau de risque. Les prédictions réalisées sur les concepts sont utilisées comme descripteurs pour prédire le niveau de risque des messages. Les niveaux de risque sont répartis sur cinq niveaux représentant les messages non à risque (niveau 0), présentant un ancien risque (niveau 1), à risque faible (niveau 2), à risque modéré (niveau 3), à risque élevé (niveau 4). Les 5 niveaux de risque en sortie sont difficiles à interpréter en termes de levée d'alerte. Nous avons aussi considéré une prédiction binaire après regroupement des niveaux 0 et 1 en *risque faible* et 2, 3 et 4 en *risque élevé*.

## 3.3 Étape 3 : levée d'une alerte

Pour la levée d'une alerte, nous avons décidé de comparer 3 modèles : 1) un modèle de dérive des concepts via l'estimation classique des pertes en *concept drift* ; 2) un modèle par comparaison de courbes ROC ; 3) un modèle basé sur des règles expertes telles que fournies par l'association *Hope for Education*.

**Modèle basé sur l'estimation des pertes.** La plupart des algorithmes de *concept drift* Gama et al. (2014) considèrent des tâches pour lesquelles on connait, à un certain moment, la vraie valeur associée aux données à prédire. Par exemple, pour des relevés de températures, on prédit des valeurs puis on les mesure. On peut alors comparer la prédiction à la vraie valeur pour estimer les pertes. On lève une alerte si les pertes sont trop élevées c-à-d. si les prédictions sont trop souvent fausses. Dans notre cas, la "vérité" n'est pas disponible sans intervention du professionnel de santé que l'on souhaite minimiser. Nous avons donc adapté le calcul réalisé par Bach et Maloof (2010) pour estimer les pertes à un instant donné en considérant le taux d'accord entre les classifieurs (i.e le nombre de classifieurs ayant réalisé la même prédiction). L'intuition est la suivante. Nous utilisons des classifieurs reposant sur des logiques différentes. Lorsqu'un nouveau concept apparaît (e.g. quelqu'un se met à évoquer régulièrement la mort alors qu'il ne le faisait pas auparavant), les classifieurs ne réagissent généralement pas au même moment. Ils sont en accord (avant le changement), puis en désaccord (au moment de la dérive) puis de nouveau en accord (quand ils ont tous détecté le changement). Dans le cas inverse de la disparition d'un concept (e.g. quelqu'un qui postait des messages joyeux n'en poste plus), il est également intéressant que le système détecte ce changement de comportement. Finalement, quel que soit l'évolution (positive et négative), elle doit être signalée au médecin (e.g. une amélioration de l'état peut suggérer une modification de la prescription).

Soit une fenêtre glissante $\mathcal{F}$ contenant les $N$ derniers exemples de l'utilisateur (lors de l'arrivée d'un nouveau message, le plus ancien est supprimé). Le $1^{er}$ et le $N^e$ exemples décrivent le message le plus ancien et le plus récent. Le taux d'erreur $\delta$ au temps $t$, noté $\delta_t$ est donné par :

$$\delta_t = \frac{\sum_{i=1}^{N} bienClasse(i)}{N}$$

où $bienClasse(i)$ retourne le taux d'accord de la prédiction retenue (i.e le nombre de classifieurs ayant prédit le niveau retenu sur le nombre de classifieurs). Le $\delta$ représente le taux de classifieurs en accord. Il prend la valeur 1 lorsque tous les classifieurs prédisent la même valeur

pour le niveau de risque (quand il n'y a aucune erreur). La suite du calcul vérifie en effet que les valeurs successives de $\delta$ ne dépassent pas un seuil prédéfini sinon une alerte est levée.

Il est important de pondérer chaque erreur de classement par sa position dans la fenêtre temporelle considérée. Pour cela, un système d'*oubli progressif* Chandramouli et al. (2014) est mis en place. Ainsi, chaque erreur est pondérée en fonction de son ancienneté :

$$\delta_t = \frac{\sum_{i=1}^{N} (bienClasse(i) * \frac{i}{N})}{\frac{N+1}{2}}$$

Les $N$ dernières valeurs $\delta$ sont mémorisées, soit le $\delta$ associé à chacun des $N$ messages précédemment nommés. Une alerte est levée si la moyenne des $N$ $\delta$ dépasse un seuil $\Delta$ fixé par l'utilisateur. Si c'est le cas, l'ensemble des $N$ valeurs $\delta$ est transmis au module de détection de changement pour calculer la "date" à laquelle le changement a eu lieu et lancer une procédure d'alerte auprès du psychiatre référent.

Pour l'interprétation du professionnel de santé, il est important de lui pointer le temps où un concept est apparu pour l'aider dans son analyse. Si une dérive du modèle est constatée, il faut réapprendre un nouveau modèle à partir de cette date. Pour cela, on cherche $\Omega$ tel que :

- $\Omega$ soit l'indice de l'exemple parmi l'ensemble des $N$ exemples. Soit $1 \leq \Omega \leq N$ ;
- La différence des moyennes des valeurs d'estimation des pertes des sous-ensembles définie par les bornes $[1, \Omega - 1]$ et $[\Omega, N]$ soit maximale.

**Modèle basé sur la comparaison de courbes ROC.** La **courbe ROC** (Receiver Operating Characteristic) Metz (1978) représente un classifieur ayant la capacité de séparer parfaitement les positifs des négatifs. Nous utilisons l'algorithme ROGER (ROc based Genetic learnER) initialement proposé en 2003 dans le cadre de la prédiction du risque cardio-vasculaire Azé et al. (2003).

L'algorithme ROGER apprend des fonctions de la forme : $f(\boldsymbol{x}_i) = \sum_j w_j * \boldsymbol{x}_i(j)$ où $\boldsymbol{x}_i(j)$ représente la valeur de la $j^{eme}$ composante de l'exemple $\boldsymbol{x}_i$. L'algorithme apprend les poids $w_j$ tels que $\sum_i rang_f(\boldsymbol{x}_i) * \mathbb{1}_{y_i=+1}$ soit minimale (où $rang_f(\boldsymbol{x}_i)$ correspond au rang de l'exemple $\boldsymbol{x}_i$ induit par la fonction $f$, et $\mathbb{1}_{y_i=+1}$ correspondant à la fonction indicatrice qui vaut $+1$ si la classe de $\boldsymbol{x}_i$ est positive et 0 sinon).

Dans notre contexte, ROGER permet d'apprendre à ordonner les messages des utilisateurs par risque décroissant. Sous l'hypothèse qu'il existe au moins un *concept drift* par utilisateur, nous considérons que le message le plus à risque (c-à-d. le message placé en première position par la fonction apprise par ROGER) est le message correspondant au drift.

**Modèle basé sur les règles expertes** Nous avons implémenté des règles expertes fournies par l'association *Hope for Education* pour la levée de l'alerte et nous avons exploré les scénarios suivants. Une alerte est levée si il y a : 1) un message avec un risque de niveau 4 dans les $N$ derniers messages ; 2) une augmentation du niveau de risque entre 2 temps consécutifs ; 3) une augmentation du niveau de risque avec un écart d'au moins 2 niveaux sur une fenêtre de $N$ messages ; Nous ferons varier $N$ pour évaluer les performances pour détecter des changements abrupts ou lents ; 4) des oscillations du niveau de risque.

Nous obtenons alors les règles suivantes : Soit $M$ l'ensemble des $N$ derniers messages d'un utilisateur où $M_1$ correspond au message le plus ancien et $M_N$ au plus récent. On notera $R_i$ le niveau de risque du message $i$.

- $\exists i, 0 \leq i \leq N | R_i = 4$
- $\exists i, 0 \leq i \leq N-1 | R_i < R_{i+1}$
- $\exists i, 0 \leq i \leq N-1 | R_i < (R_{i+1} - 1)$
- $\exists i \exists j, 0 \leq i \leq N-1, 0 \leq j \leq N-1 | R_i < R_{i+1}, R_j > (R_{j+1})$

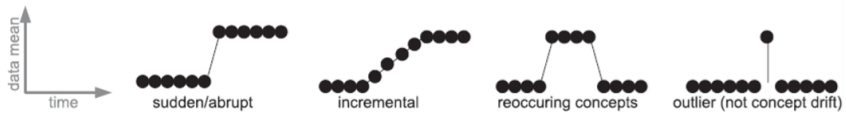


FIG. 1 – *Les différentes formes de* concept drift *(adaptés de Gama et al. (2014))*

Le schéma 1 représente les différentes formes de changement que nous souhaitons capturer via les règles expertes : 1) Le changement *immédiat* correspond à un individu qui parle du jour au lendemain de méthodes de suicide ; 2) Le changement *incrémental* correspond au cas où l'individu parle de plus en plus de son mal être ; 3) Le changement *récurrent* est un utilisateur qui régulièrement parle de son mal être ; 4) Le *bruit* serait un message isolé parlant de suicide.

# 4 Protocole expérimental

## 4.1 Données utilisées

Cette étude a porté sur l'analyse de messages écrits par des populations à risque. Pour cela, nous avons utilisé des messages issus de groupes Facebook que nous avons récoltés via l'API Facebook [3]. Ces groupes ont été choisis car directement liés à la thématique du suicide ou à des facteurs de risque connus pour le suicide (*suicide*, *anorexie*, *mutilation* ou *harcèlement*). Les auteurs sont généralement des adolescents entre 13 et 18 ans. Il est important de noter que ces groupes comprennent des adultes dans la situation d'anciennes victimes qui viennent témoigner de leur expérience ainsi que des parents ou des proches qui viennent demander des conseils pour leur enfant possiblement harcelé. Nous avons récolté 4597 messages entre le 15 Mars et le 3 Juin 2015.

Dans le cadre de nos expérimentations, Nous avons sélectionné manuellement 22 comptes d'adolescents ayant entre 3 et 14 messages postés sur les groupes. Ils respectent les conditions suivantes : 1) L'auteur est une personne à risque ; 2) La personne évoque son bien/mal-être ; 3) La personne n'est pas un modérateur du groupe. Nous avons alors obtenu 168 messages. Ce faible nombre s'explique par la difficulté que représente l'annotation des données pour cette étude. L'association *Hope for Education* nous a également fourni un protocole utilisé par les volontaires pour détecter le niveau d'urgence. Nous l'avons adapté à notre cas d'étude qui ne se limite pas au cas des personnes harcelées. Nous considérons 5 degrés d'urgence : 1) *Pas d'urgence* : Il n'y a aucune urgence à traiter le message de la personne ; 2) *Risque minimal* :

3. http ://developers.facebook.com

Le problème de la personne se situe dans le passé. Cette situation est encore difficilement supportable pour la personne, elle estime que sa parole doit être libérée ; 3) *Risque intermédiaire* : La personne a un problème. L'isolement commence à s'installer ; 4) *Risque important* : La personne a un problème durable pouvant inclure de la violence. Un isolement important s'installe ; 5) *Risque absolu* : La personne est violente verbalement et/ou physiquement. Elle tient des propos suicidaires et/ou se met en danger elle-même.

Les 168 messages ont été annotés manuellement par trois personnes. Chaque message a reçu au total 13 annotations : présence ou absence des 12 concepts et un niveau de risque. Le kappa de Fleiss (Fleiss (1971)) qui évalue la concordance lors de l'assignation qualitative d'objets en catégorie pour plusieurs observateurs, donne une valeur de 0,563 pour les cinq niveaux de risque. Les concepts montrent aussi des valeurs de Kappa intéressantes (e.g. 0,702 pour le concept *anorexie*). Les annotations ont été retravaillées par les annotateurs afin d'obtenir un consensus important et donc un corpus bien annoté.

## 4.2 Évaluations

**Détection du niveau de risque.** Lors de l'étape 1 du module 2 pour détecter les concepts dans un message, nous avons souhaité favoriser une bonne classification des messages possédant un concept. L'objectif est de trouver la meilleure configuration qui maximise le *rappel* de la classe *oui* (i.e minimiser le nombre de messages devant être classés comme *"oui"* et qui sont classés comme *"non"*). De la même manière, le classifieur utilisé pour le $2^e$ niveau doit être capable de maximiser le *rappel* de la classe de plus haut niveau (i.e niveau 4 dans le cas d'une classification sur 5 classes et niveau 1 dans le cas d'une classification binaire du niveau de risque). Il est important de bien classer un message présentant un haut niveau de risque pour être sûr de réagir à un tel message. Afin de réaliser le principe d'ensemble learning, cinq classifieurs (*J48*, *JRip*, *SMO*, *Naive Bayes*, *IB1*) sont utilisés pour chaque prédiction grâce à l'outil Weka (Hall et al. (2009)). La validation est effectuée par validation croisée en mode Leave-one-out sur les 168 messages.

**Levée d'une alerte.** La levée d'alerte, c'est-à-dire la remontée d'une information au professionnel de santé prenant la décision (positive ou négative), traite tous les messages d'un utilisateur en même temps. Pour cela, l'intégralité des messages associés à un niveau de risque est transmis à l'étape 3. On admet que chaque utilisateur a présenté un risque (i.e. une alerte doit être levée dans la séquence de messages). Pour cela, nous imposerons que le message correspondant à l'instant de la levée d'alerte ne peut être le premier message posté n'ayant aucune indication de l'état "normal" de la personne à cet instant précis. De plus, dans le cas où plusieurs messages seraient qualifiables de *message le plus à risque*, le plus ancien est retenu afin de réagir le plus vite possible. *Estimation des pertes :* L'estimation des pertes est évaluée grâce aux calculs effectués précédemment. *ROGER :* L'algorithme présente directement le message le plus probable d'être à risque. *Règles expertes :* Les 4 règles sont testées sur chaque message. Le message ayant répondu positivement au maximun de règles est qualifié de message le plus à risque.

# 5 Expérimentations sur des données réelles

## 5.1 Détection du niveau de risque (1ère étape)

Le tableau 1 présente les valeurs de rappel associées à la détection des concepts, ainsi qu'entre parenthèses les valeurs de F-Mesure. Cette première couche de détection, qui utilise les informations provenant de Facebook ainsi que les résultats des prétraitements, détermine la présence ou non des 12 concepts. Par manque de place, nous présentons les résultats des concepts *harcèlement*, *peur* et *solitude*. La première constatation est que chaque classifieur s'adapte différement selon les concepts (e.g. *Naive Bayes* est efficace sur la détection du *harcèlement*, alors que les concepts de *peur* et *solitude* sont mieux captés par l'algorithme *J48*). Cette différence peut s'expliquer par des lexiques spécifiques aux concepts.

| | *harcèlement* | *peur* | *solitude* |
|---|---|---|---|
| J48 | 47,1% (64,8%) | **60,0% (89,6%)** | **61,5% (86,1%)** |
| Naive Bayes | **54,9% (70,0%)** | 52,0% (82,4%) | 53,8% (88,2%) |
| IB1 | 05,8% (59,1%) | 24,0% (82,0%) | 30,8% (82,5%) |
| JRip | 27,5% (62,0%) | 40,0% (84,0%) | 53,8% (88,7%) |
| SMO | 37,3% (70,2%) | 56,0% (90,5%) | 34,6% (83,8%) |
| Vote | 29,4% (69,8%) | 48,0% (88,9%) | 46,2% (86,7%) |

TAB. 1 – *Résultat des classifieurs sur la détection du harcèlement, de la peur et de la solitude*

## 5.2 Détection du niveau de risque (2ème étape)

Les six classifieurs sont ensuite appliqués à la seconde couche de prédiction (prédiction du niveau de risque). Pour cela, 4 tests sont réalisés : une prédiction binaire et une prédiction en cinq niveaux de risque dans le cas d'une prédiction à partir des concepts et dans le cas d'une classification directement à partir des messages Facebook. Le tableau 2 présente ces résultats par les deux mesures retenues pour différencier les classifieurs : en premier le rappel sur le niveau le plus à risque et en cas d'égalité la F-Mesure globale (entre parenthèse dans le tableau). Une première constatation non surprenante est que la classification binaire obtient de meilleurs résultats que la classification en 5 niveaux de risque. De plus, on remarque que notre choix de modèle Stacking obtient de meilleurs résultats que la prédiction directe depuis les messages Facebook. En effet, certains classifieurs obtiennent un rappel très élevé sur ce test (e.g. IB1 pour la prédiction directe en cinq niveaux de risque) mais cela est obtenu en classant une grande majorité des messages dans la classe de niveau le plus élevé, ce qui ne se révèle pas efficace en réalité comme le montre le score de F-Mesure associé (e.g. 6,6 %)

## 5.3 Levée d'une alerte

Les trois approches désignent le message correspondant au passage de l'individu dans un état à risque dans la série de messages de cet individu. Rappelons que les séries de messages ont été choisies car elles contiennent toutes un changement d'état. De manière globale, les 3 approches se coordonnent sur le même message, soit 3 individus sur les 22 (e.g. utilisateur $U_{15}$)

| | 2 niveaux (Stacking) | 5 niveaux (Stacking) | 2 niveaux (Direct) | 5 niveaux (Direct) |
|---|---|---|---|---|
| J48 | 92,6% (89,2%) | 83,3% (59,2%) | 70,4% (57,8%) | 23,3% (25,5%) |
| Naive Bayes | 88,9% (85,7%) | 73,3% (44,5%) | **99,1% (50,0%)** | 26,7% (29,8%) |
| IB1 | 88,0% (80,1%) | 73,3% (50,1%) | 47,2% (53,2%) | **96,7% (06,6%)** |
| JRip | 92,6% (87,5%) | **83,3% (65,4%)** | 70,4% (62,5%) | 23,3% (25,4%) |
| SMO | 90,7% (87,5%) | 83,3% (51,8%) | 80,6% (62,5%) | 06,7% (16,5%) |
| Vote | **96,6% (89,0%)** | 83,3% (57,1%) | 78,7% (63,7%) | 92,3% (13,4%) |

TAB. 2 – *Résultat des classifieurs sur la couche de détection du niveau de risque*

ou de façon très proche soit 13 (e.g. utilisateur $U_6$). Toutefois, pour 6 utilisateurs, les approches sont en désaccord (e.g. utilisateur $U_8$), Les désaccords surviennent dans le cas de personnes qui alternent des messages positifs et des messages négatifs. Il est important de noter qu'aucune des trois méthodes n'est plus sensible que les autres (i.e. systématiquement avant les autres). Ces méthodes se veulent donc complémentaires afin de prévenir au mieux les risques.

| | $U_1$ | $U_2$ | $U_3$ | $U_4$ | $U_5$ | $U_6$ | $U_7$ | $U_8$ | $U_9$ | $U_{10}$ | $U_{11}$ |
|---|---|---|---|---|---|---|---|---|---|---|---|
| EdP | -2 | **+1** | +5 | **0** | **0** | **0** | **0** | **-4** | +2 | +3 | +3 |
| ROGER | -2 | +4 | +6 | -6 | **0** | -1 | -1 | +6 | **0** | **+2** | +2 |
| Règles | **0** | +4 | **+4** | -1 | **0** | -1 | **0** | +5 | **0** | **+2** | **+1** |
| | $U_{12}$ | $U_{13}$ | $U_{14}$ | $U_{15}$ | $U_{16}$ | $U_{17}$ | $U_{18}$ | $U_{19}$ | $U_{20}$ | $U_{21}$ | $U_{22}$ |
| EdP | **0** | **+1** | **0** | **0** | +2 | -1 | +4 | **0** | **0** | **0** | -3 |
| ROGER | -1 | -5 | +1 | **0** | **0** | +1 | **0** | **0** | **0** | -4 | -3 |
| Règles | -1 | +3 | **0** | **0** | +1 | **0** | +1 | **0** | +1 | -2 | **-1** |

TAB. 3 – *Écart de levée d'alerte entre le message désigné par l'humain et les 3 approches*

# 6 Conclusion

Dans cette étude, nous avons proposé une chaîne de traitements complète basée sur une analyse de type *Stacking* qui permet d'évaluer le niveau de risque suicidaire d'un individu et de lever une alerte. Une première étape de cette méthode consiste à détecter 12 concepts dans les messages textuels, correspondant à des facteurs de risque connus des psychiatres. Nous appliquons une méthode de *concept drift* sur ces concepts pour détecter un changement de comportement. Lors d'une levée d'alerte, les concepts sont également présentés aux psychiatres afin de les aider à interpréter l'alerte et à prendre une décision d'intervention.

Une première limite concerne les ressources utilisées pour détecter les concepts dans les messages. Un travail sur la production de lexiques francophones adaptés à chaque concept et prenant en compte le vocabulaire familier des individus sur les réseaux sociaux est nécessaire pour améliorer les scores des classifieurs de premier niveau. Une deuxième limite concerne les données considérées pour lever l'alerte. Nous nous sommes ici focalisés sur les textes des messages mais il serait également intéressant de considérer d'autres signes comme la fréquence ou l'heure des messages, signes de comportements atypiques. Nous envisageons d'utiliser des combinaisons de classifieurs tant il semble évident que de multiples structures cachées sont à mettre en exergue sur ces données. Une autre limite concerne la prise en compte du type

d'orientation, positive ou négative, que l'on peut associer à l'apparition ou à la disparition d'un concept (e.g. la disparition du concept de *joie* ou l'apparition du concept de *peur* signale une dégradation de l'état alors que la disparition du concept de *tristesse* et l'apparition du concept de *joie* signale une amélioration de l'état de la personne). Une dernière limite de ce travail réside dans le fait que les messages analysés proviennent de groupes publics non complètement représentatifs de l'activité générale des individus dans les réseaux sociaux. Nous prévoyons de travailler avec un service de psychiatrie sur des comptes de personnes ayant donné leur consentement, pour évaluer l'impact de cette méthode sur la prévention de la récidive. En particulier, un travail sur la gestion des faux positifs sera réalisé. Notre objectif est de les minimiser afin de ne pas solliciter abusivement le médecin avec de fausses alertes. Pour cela, le seuil de levée d'alerte doit être ajusté. Un seuil haut (i.e. proche de 1) sera très réactif mais sollicitera abusivement le médecin. À l'inverse, un seuil bas (i.e. proche de 0) demandera rarement au médecin de valider l'alerte mais pourrait laisser passer des cas critiques.

# Références

Abboute, A., Y. Boudjeriou, G. Entringer, J. Azé, S. Bringay, et P. Poncelet (2014). Mining twitter for suicide prevention. In *19th International Conference on Applications of Natural Language to Information Systems*, pp. 250–253.

Azé, J., N. Lucas, et M. Sebag (2003). A new medical test for atherosclerosis detection : Geno. In *Discovery Challenge PKDD 2003*.

Bach, S. et M. Maloof (2010). A bayesian approach to concept drift. In *Advances in Neural Information Processing Systems*, pp. 127–135.

Balahur, A. (2013). Sentiment analysis in social media texts. In *4th workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, pp. 120–128.

Cash, S. J., M. Thelwall, S. N. Peck, J. Z. Ferrell, et J. A. Bridge (2013). Adolescent suicide statements on myspace. *Cyberpsy., Behavior, and Soc. Networking 16*(3), 166–174.

Chandramouli, B., J. Goldstein, et A. Quamar (2014). Scalable progressive analytics on big data in the cloud. In *International Conference on Very Large Databases*, Volume 6, pp. 1726–1737.

Choudhury, M. D., S. Counts, E. Horvitz, et M. Gamon (2013). Predicting depression via social media. AAAI Conference on Weblogs and Social Media.

Christensen, H., P. J. Batterham, et B. Dea (2014). E-health interventions for suicide prevention. *International Journal of Environmental Research and Public Health 11*(8), 8193–8212.

Dinakar, K., E. Weinstein, H. Lieberman, et R. L. Selman (2014). Stacked generalization learning to analyze teenage distress. In *International AAAI Conference on Weblogs and Social Media*, pp. 81–90.

Fleiss, J. L. (1971). Measuring nominal scale agreement among many raters. *Psychological bulletin 76*(5), 378.

Gama, J., I. Žliobaitė, A. Bifet, M. Pechenizkiy, et A. Bouchachia (2014). A survey on concept drift adaptation. *ACM Comput. Surv. 46*(4), 44 :1–44 :37.

Hall, M., E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, et I. H. Witten (2009). The weka data mining software : an update. *ACM SIGKDD explorations newsletter 11*(1), 10–18.

Hanley, J. A. et B. J. McNeil (1982). The meaning and use of the area under a receiver operating characteristic (roc) curve. *Radiology 143*(1), 29–36.

Jashinsky, J., S. H. Burton, C. L. Hanson, J. West, C. Giraud-Carrier, M. D. Barnes, et T. Argyle (2014). Tracking suicide risk factors through twitter in the us. *Crisis 35*(1), 51–59.

Krieck, M., J. Dreesman, L. Otrusina, et K. Denecke (2011). A new age of public health : Identifying disease outbreaks by analyzing tweets. In *Proceedings of Health WebScience Workshop, ACM Web Science Conference*.

Metz, C. E. (1978). Basic principles of roc analysis. *Seminars in nuclear medicine VIII*(4), 283–298.

Moreno, M. A., L. A. Jelenchick, K. G. Egan, E. Cox, H. Young, K. E. Gannon, et T. Becker (2011). Feeling bad on Facebook : depression disclosures by college students on a social networking site.

Paul, M. et M. Dredze (2011). You are what you tweet : Analyzing twitter for public health. *the Fifth International AAAI Conference on Weblogs and Social Media*, 265–272.

Plutchik, R. et H. M. Van Praag (1994). Sucide risk : Amplifiers and attenuators. *Journal of Offender Rehabilitation 21*(3-4), 173–186.

Sadilek, A., H. Kautz, et V. Silenzio (2012). Modeling spread of disease from social interactions. In *Sixth AAAI International Conference on Weblogs and Social Media (ICWSM*, pp. 322–329.

Schmid, H. (1994). Probabilistic part-of-speech tagging using decision trees. In *International Conference on New Methods in Language Processing*, Volume 12, pp. 44–49.

Wang, H., W. Fan, P. S. Yu, et J. Han (2003). Mining concept-drifting data streams using ensemble classifiers. In *Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 226–235. ACM.

# Summary

Suicide has long been a worrisome problem for society and is an event that has far-reaching effects. Health organizations such as the World Health Organization (WHO) and the French National Observatory of Suicide (ONS) have pledged to reduce the number of suicides by 10% in all countries by 2020. While suicide is a very marked event, there are often behaviors and words that can act as early signs of predisposition to suicide. The objective of this internship is to develop a system that semi-automatically detects these markers through social networks. Previous work has proposed the classification of Tweets using vocabulary in topics related to suicide: sadness, psychological injuries, mental state, depression, fear, loneliness, proposed suicide method, anorexia, insults, and cyber bullying. During this training period, we add a new dimension, time to reflect changes in the status of monitored people. We implemented it with different learning methods including an original *concept drift* method. We have successfully experienced this method on synthetic and real data sets issued from the Facebook networks.