Suivi et détection des idéations suicidaires dans les médias sociaux

Bilel Moulahi[1], Jérome Azé[1], Sandra Bringay[1,2]

[1] LIRMM, UNIVERSITÉ DE MONTPELLIER, CNRS
bilel.moulahi@lirmm.fr

[2] AMIS, Université Paul Valéry Montpellier

Résumé : L'utilisation croissante des médias sociaux permet un accès sans précédent aux comportements, aux pensées et aux sentiments des individus. Nous nous intéressons ici à l'évolution des états émotionnels des individus captés au travers des services de microblogging de type Twitter. Notre objectif est de prédire l'apparition d'idéations suicidaires. Dans ce travail, nous avons mis en place une chaîne de traitements permettant d'extraire des caractéristiques à partir des messages reflétant l'état émotionnel. Puis, nous appliquons un modèle basé sur les Conditionnal Random Fields pour prédire un nouvel état. L'originalité de l'approche est de prendre en compte l'historique des états émotionnels pour prédire le nouvel état. Une expérimentation préliminaire nous a permis d'évaluer notre approche sur des cas réels d'utilisateurs de Twitter. Ces type d'approche permet de mieux comprendre les liens entre expressions dans les médias sociaux et idéations suicidaires ainsi que les transitions entre états émotionnels.

Mots-clés : Média sociaux, Suicide, Conditionnal Random Fields.

1 Introduction

Les médias sociaux sont de plus en plus utilisés par les professionnels de santé pour détecter et diagnostiquer des troubles dépressifs majeurs (De Choudhury *et al.*, 2013; O'Dea *et al.*, 2015; Sueki, 2015; Adler *et al.*, 2016; Maigrot *et al.*, 2016). Les plateformes comme Twitter et Facebook facilitent l'auto-présentation sélective de comportements indésirables, tels que l'automutilation, l'anorexie, ainsi que l'expression d'émotions négatives liées aux idéations suicidaires, en particulier chez les jeunes.

Afin de mieux comprendre ces nouvelles pratiques, de nombreuses études comme celles de (Burnap *et al.*, 2015; Colombo *et al.*, 2016) ont porté sur la recherche dans le discours des individus des références à la dépression et aux idéations suicidaires. Des auteurs comme (Burnap *et al.*, 2015; Colombo *et al.*, 2016) soulignent une corrélation entre les taux de tentatives de suicide et le volume de messages liés aux idéations suicidaires publiés dans les médias sociaux. Si les médias sociaux peuvent affecter les individus en répandant des pensées suicidaires, ils peuvent aussi avoir un rôle positif en aidant ces individus à trouver du soutien moral. Par exemple, l'activité intense en ligne, notamment nocturne, est un signe précoce permettant d'anticiper une dégradation de l'état émotionnel d'un individu.

De méthodes efficaces sont désormais disponibles pour analyser les sentiments exprimés dans les réseaux sociaux (Barbosa & Feng, 2010; Kim *et al.*, 2013). Plusieurs études récentes tirent partie de ces travaux pour la détection et la surveillance des idéations suicidaires (Spasic *et al.*, 2012; Gunn & Lester, 2012; Poulin *et al.*, 2014) et d'états dépressifs (Moreno *et al.*, 2011; Karmen *et al.*, 2015). Toutefois, la plupart des méthodes de l'état de l'art prédisent les émotions véhiculées par les utilisateurs au niveau d'un message ou de l'individu. Elles ne prennent pas en compte l'évolution du comportement de l'individu. Or, les idées suicidaires sont incluses

dans un continuum de séquences d'événements influençant les états émotionnels et qui peuvent éventuellement conduire à une tentative de suicide (Adler *et al.*, 2016). Étant donné la nature séquentielle des contenus produits par les individus dans les médias sociaux, nous utilisons dans ces travaux un modèle basé sur les Conditional Random Fields (Lafferty *et al.*, 2001; Sutton & McCallum, 2012) qui permet de capturer l'évolution de l'état émotionnel des individus au fil du temps, en tenant compte du contexte passé et de l'activité du moment. Les état émotionnels sont au préalable inférés à partir des messages, via une analyse de textes permettant d'identifier les facteurs de risque (De Choudhury *et al.*, 2013).

Nous avons évalué notre approche sur un corpus de tweets annotés manuellement, publiés par des individus ayant exprimé des références au suicide. La collection d'individus a été validée par un psychiatre pour n'inclure que les utilisateurs ayant présenté de réels symptômes. Les résultats expérimentaux montrent que le système prédit correctement des séquences d'états mentaux.

Dans le reste de l'article, nous présentons un état de l'art succinct, puis décrivons notre approche pour la détection d'idéations suicidaires. Nous présentons et discutons les résultats, avant de conclure sur des perspectives.

2 Etat de l'art

En France, près de 10 000 personnes mettent fin à leurs jours chaque année, soit environ 25 par jour, selon le dernier rapport de l'OMS [1]. Deuxième cause de mortalité chez les 15-24 ans, après les accidents, le suicide est un fléau qui touche des adolescents souvent fragilisés par cette période charnière de la vie. Avec l'avènement des médias sociaux, les personnes à risque et notamment les jeunes, utilisent des outils comme Facebook, Twitter et Reddit pour exprimer des idéations suicidaires. Il est possible d'utiliser ces médias pour détecter de manière précoce les individus vulnérables et intervenir rapidement. En 2015, Facebook [2] a introduit un nouveau service permettant aux utilisateurs de rapporter un comportement suicidaire. Très récemment, ce service a évolué [3] pour permettre aux personnes qui visionnent un *live-stream* Facebook [4] d'interpeller son auteur ou de faire un signalement.

De nombreuses études ont porté sur les notes laissées par les individus avant un suicide. Ces notes ont été analysées en développant des classifieurs supervisés et non supervisés pour identifier les sujets discutés ainsi que les émotions exprimées par des personnes étant passées à l'acte (Spasic *et al.*, 2012; Pestian *et al.*, 2008).

Plus récemment, plusieurs études se sont intéressées à l'évaluation des facteurs de risque suicidaires dans les médias sociaux avec l'objectif de mieux comprendre ou de prévenir le suicide en détectant les idéations suicidaires de manière précoce. Par exemple, le projet Durkheim [5] étudie les activités des anciens combattants américains sur Twitter, Facebook et LinkedIn. L'objectif de ce projet est d'identifier les marqueurs de comportements à risque. Poulin *et al.* (2014) ont développé des modèles de prédiction en utilisant les textes des notes. Les résultats montrent

1. Observatoire national du suicide `http://www.who.int/topics/suicide/fr/`
2. `https://www.facebook.com/help/suicideprevention`
3. `https://newsroom.fb.com/news/2017/03/building-a-safer-community-with-new-suicide-prevention-tools/`
4. vidéo en direct
5. `http://www.durkheimproject.org/research/`

que les personnes qui se sont suicidées expriment souvent de la peur et une certaine agitation avant de passer à l'acte. Les modèles de prédiction proposés ont montré des taux d'exactitude proche de 65%. Gunn & Lester (2012) ont analysé les messages Twitter d'une jeune fille qui venait de se suicider, publiés les vingt-quatre heures précédant son décès. Ils ont trouvé une augmentation des émotions positives et un changement de la focalisation de soi à d'autres lorsque le moment du décès s'est approché. Les auteurs ont également étudié un éventail plus large de tweets. Pour cela, ils ont utilisé le logiciel Linguistic Inquiry and Word Count (LIWC) [6] pour identifier dans le discours, des mots porteurs d'émotions ainsi que des processus cognitifs. Sueki (2015) ont utilisé un panel en ligne de 250 jeunes d'une vingtaine d'années, utilisant régulièrement Twitter, pour examiner l'association entre les tweets liés au suicide et les passages à l'acte. Les auteurs ont étudié les caractéristiques linguistiques de l'idéation suicidaire et ont identifié les marqueurs les plus fréquents. Par exemple, des phrases comme "*I want to commit suicide*" sont fortement associées aux tentatives de suicide, alors que des phrases suggérant une intention suicidaire, comme "*I want to die*" y sont moins associés. Contrairement aux techniques populaires d'apprentissage, Karmen *et al.* (2015) ont combiné plusieurs méthodes de TAL pour filtrer les utilisateurs de forums et identifier les symptômes de dépression. Ces auteurs ont mis en correspondance les questionnaires traditionnels de dépistage de la dépression avec un ensemble de termes associés aux symptômes. Ensuite, ils détectent ces termes dans les textes et en déduisent un score au niveau du message.

La littérature actuelle manque de modèles efficaces pour prédire les tentatives de suicide. Actuellement, peu d'approches intègrent l'évolution du comportement de l'individu. L'analyse porte généralement sur un message ou sur l'ensemble des messages d'un individu. L'analyse ne permet pas de prédire à quel moment une personne présente un risque suicidaire. Maigrot *et al.* (2016) explorent une approche basée sur les concepts drift pour identifier un temps à risque. Une limite à leur approche est de ne pas expliciter les transitions entre les états émotionnels comme nous souhaitons le faire dans ce travail.

3 Un modèle basé sur le contexte pour le suivi et la détection les idéations suicidaires dans les médias sociaux

Dans cette section, nous reformulons le problème et décrivons le modèle utilisé pour monitorer les idéations suicidaires dans les médias sociaux. Étant donné une séquence de messages pouvant traiter de thèmes jugés à risque tels que la dépression, le suicide, l'automutilation ou l'anorexie, mais également contenir des thèmes sans rapport ou même des blagues dans un intervalle de temps très court, avec quelle précision pouvons-nous prédire qu'un individu présente un réel risque suicidaire ? Un modèle d'analyse de sentiments typique traite ce problème comme une tâche de classification multi-classes et prédit une étiquette pour chaque message indépendamment de la séquence entière. Dans ce travail, nous supposons que l'état émotionnel déduit d'un message au temps t, dépend des états émotionnels précédents. Notre hypothèse principale est que les états émotionnels peuvent être modélisés comme des observations dépendantes et continues, qui peuvent être capturées via des méthodes de traitement automatique de la langue puis prédites. L'état émotionnel peut être représenté soit par un état positif, neutre ou négatif (Barbosa & Feng, 2010), soit par des modèles plus complexes incluant des émotions comme la

6. https://liwc.wpengine.com

tristesse, l'espoir, l'excitation, etc. (Larsen & Diener, 1992; Yik *et al.*, 2011; Kim *et al.*, 2013). Dans la suite, nous considérons trois niveaux d'états émotionnels mais notre approche peut être facilement généralisée indépendamment du nombre d'états émotionnels.

3.1 Description du problème

Soit P = <p_1, p_2, ..., p_n> un flux continu de messages (tweets, messages facebook, etc.) ordonnés dans le temps dans une fenêtre temporelle W. Le problème consiste à prédire un vecteur Y = <y_1, y_1, ..., y_n> d'états émotionnels associés à la séquence de messages observée P. Les observations P en entrée sont représentées par des vecteurs d'attributs. Chaque observation p_j contient différentes informations à propos du message au temps t_j. Chaque variable y_j est un état émotionnel inféré à partir de l'observation p_j. La Figure 1 décrit une séries de messages impliquant un changement de l'état émotionnel.

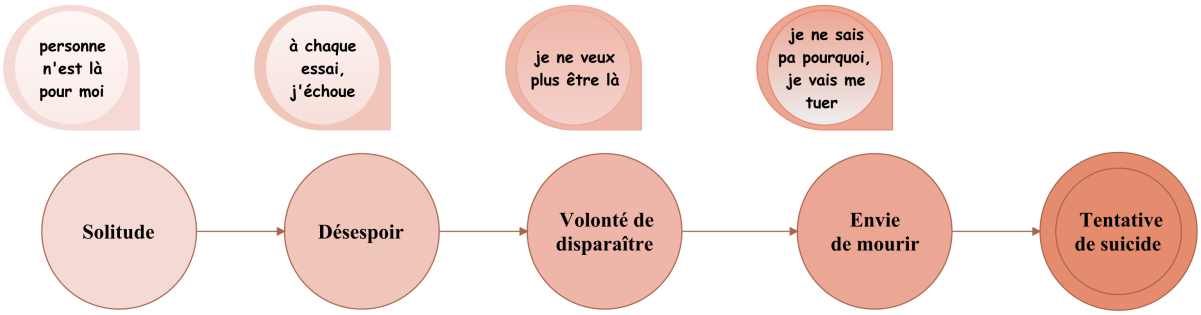

FIGURE 1 – Exemple d'évolution de l'état émotionnel d'un individu.

Ces changements d'états sont inspirés par un travail récent en psychologie cognitive (Adler *et al.*, 2016). Les auteurs ont étudié différents types de comportements suicidaires en explorant des signes cognitifs et les changements de comportements pouvant indiquer une tentative suicide. Par exemple, comme le montre la figure 1, un individu évoque sa *volonté de disparaître* avec le message "*Je ne veux plus être là*". Il est susceptible de voir son état émotionnel évoluer vers un niveau plus risqué comme *Envie de mourir* qu'il exprime avec le message "*Je vais me tuer*", avant de réaliser une *tentative de suicide*. Chaque nœud de la figure représente un état émotionnel qui est calculé en fonction d'une observation, le message de l'utilisateur. Les transitions (arêtes) entre les états encodent la séquence des changements d'état. Les modèles de réseaux de Markov permettent de représenter ces changements émotionnels séquentiels. Une généralisation intéressante est donnée par des modèles graphiques tels que les Conditional Random Fields (CRF) (Lafferty *et al.*, 2001).

Dans la suite, nous présentons les deux étapes de notre méthode. La première étape permet d'associer un ou plusieurs états émotionnels à un message à partir de l'analyse du texte de ce message. La deuxième étape permet de modéliser les interactions entre états émotionnels et peut être utilisée pour prédire le prochain état émotionnel d'un individu.

3.2 Extraction des caractéristiques

Avant d'extraire des caractéristiques des messages, ces derniers sont prétraités (mise en minuscules, suppression des ponctuations multiples, des caractères spéciaux, des mentions à d'autres utilisateurs et des URL). Nous avons extrait des caractéristiques largement utilisées dans la littérature (Spasic *et al.*, 2012).

– Le premier ensemble de caractéristiques inclut les caractéristiques lexicales du texte. Nous utilisons les étiquettes grammaticales (POS) pour capturer l'auto-référence (pronoms personnels à la première personne "*I*", "*My*", "*Je*", "*Mon*", etc), les noms, les verbes et les adverbes. De Choudhury *et al.* (2013) ont montré que l'utilisation de la première personne au singulier ou au pluriel peut révéler le bien-être ou le mal-être mental. Nous prenons également en compte l'intensité des émotions, en considérant la présence d'*intensifieurs* comme *très, complètement, intensément* surtout lorsqu'ils sont utilisés avec des pronoms personnels (eg., "je suis très triste"). Quand une phrase contient une négation suivie d'un symptôme, nous inversons cette caractéristique (eg., "je ne vais pas bien !").
– Le second ensemble de caractéristiques est lié aux lexiques. Nous cherchons dans les messages des termes couramment utilisés par les personnes à risque dans les médias sociaux. Nous considérons la fréquence des termes faisant référence aux émotions négatives, à la dépression, à l'automutilation, à la tristesse, à la santé mentale et au suicide. Pour ce faire, nous nous sommes inspirés des travaux de De Choudhury *et al.* (2013), qui ont exploité le lexique ANEW Bradley & Lang (1999) contenant un classement d'émotions pour un large nombre de mots [7]. Ensuite, nous enrichissons ces caractéristiques en incluant un autre lexique qui se réfère aux mots d'injure. En effet, De Choudhury *et al.* (2013) ont montré que ces caractéristiques véhiculent des informations importantes dans le contexte de l'analyse des états émotionnels.

3.3 Modèle basé sur les Conditional Random Fields

CRF est un type de modèle graphique probabiliste non dirigé qui a été appliqué avec succès dans de nombreux problèmes de traitement de textes et de visualisation (Sutton & McCallum, 2012). Un avantage de ce modèle réside dans sa capacité à capturer les dépendances complexes entre les observations, en plus des interprétations complètes de la relation entre les caractéristiques qu'il fournit. Dans notre contexte, cette propriété est très importante étant donné que la transition d'un état émotionnel à un autre dépend fortement des états observés précédemment. Par exemple, comme représenté dans la figure 1, il est très improbable que l'état émotionnel d'un individu saute soudainement de *solitude* (non risqué) à un état *tentative de suicide* (très risqué). La modélisation CRF est un modèle puissant qui aide à apprendre les comportements des utilisateurs et à prédire la séquence des états mentaux des utilisateurs.

Soit une séquence de messages observée $P = <p_1, p_2, ..., p_n>$ et une séquence d'états émotionnels cachés $Y =< y_1, y_1, ..., y_n>$, CRF modélise la probabilité conditionnelle comme suit :

$$p(Y|P) = \frac{1}{Z(P)} exp(\sum_{i=1}^{n} \sum_{k=1}^{F} w_k f_k(y_{i-1}, y_i, P, i)) \quad (1)$$

7. Ce lien contient le code et les scripts utilisés pour générer ce lexique : `https://github.com/sbma44/begin_anew`.

ou Z est un facteur de normalisation (aussi appelé la fonction de partition) pour que $p(Y|P)$ soit une probabilité valide pour toutes les séquences étiquetées. Z est défini comme la somme de l'exponentielle du nombre de séquences :

$$Z(P) = \sum_{P} exp(\sum_{i=1}^{n} \sum_{k=1}^{F} w_k f_k(y_{i-1}, y_i, P, i)) \quad (2)$$

Les paramètres w_k sont les poids des caractéristiques f_i. Ils sont appris par des techniques d'optimisation comme les approches par gradient. Les fonctions caractéristiques $f_k(y_{i-1}, y_i, P, n)$ prennent en compte une paire d'états émotionnels adjacents y_{i-1}, y_i, la séquence entière de messages P et la position courante dans la séquence i.

Notons que l'utilisation de CRF nous permet de définir un grand nombre de fonctions dépendantes ou indépendantes sans nous soucier de la relation statistique complexe entre ces fonctions. L'utilisation de chaque fonction dépend du poids w_k qui agit comme facteur d'activation de la fonction.

4 Expérimentations

Dans ce qui suit, nous détaillons la préparation du jeu de données utilisé, puis nous analysons les performances de l'approche en utilisant les caractéristiques détaillées dans 3.2. Nous explorons aussi l'importance des caractéristiques extraites à partir des messages et nous montrons les thèmes importants pour chaque état émotionnel en utilisant le modèle Latent Dirichlet Allocation Hoffman *et al.* (2010).

4.1 Préparation des données

En raison de l'absence de base de données librement accessible pour l'évaluation des méthodes de détection des risques de suicide dans les médias sociaux, nous avons utilisé l'API en temps réel Twitter [8] pour collecter des tweets contenant des références à des thèmes tels que la dépression, l'automutilation, l'anorexie et le suicide. La liste de mots clés utilisée pour récupérer les tweets a été définie manuellement à partir de la liste des facteurs de risque définie par l'APA (American Psychological Association [9]) et la liste des signes avant-coureurs définie par l'AAS (American Association of Suicidology [10]).

Parmi les tweets recueillis, nous n'avons conservé que ceux pour lesquels un symptôme lié au suicide a été validé par un psychiatre. 60 individus ont ainsi été choisis. Pour éviter un ajustement excessif du modèle, nous avons également inclus 60 comptes Twitter d'individus non à risque en utilisant les mêmes mots clés. Nous avons sélectionné au hasard les 50 derniers tweets de chacun de ces groupes. La collection totale de données contient 5976 tweets. Huit chercheurs et un psychiatre ont manuellement annoté 507 tweets de la collection pour déterminer les états émotionnels associés. Pour cette étude préliminaire, nous avons considéré trois états émotionnels. Le choix de ces classes est motivé par les travaux de Lehrman *et al.* (2012). Ces classes sont définies comme suit :

8. `https://dev.twitter.com/streaming/overview`
9. `http://www.apa.org/topics/suicide/`
10. `http://www.suicidology.org`

- *Aucune détresse* : le message traite d'événements quotidiens tels que le travail, les sorties, les activités du week-end, etc.
- *Détresse minimale/modéré* : le message exprime un niveau de détresse qui pourrait être considéré comme commun pour la plupart des individus tels que un examen, une présentation pour le travail, une dispute avec un ami, etc.
- *Détresse importante* : le message mentionne des références à l'auto-mutilation, aux idéations suicidaires, des excuses, des sentiments négatifs comme l'inutilité, la haine de soi, la culpabilité, etc.

Chaque tweet a été examiné par au moins deux annotateurs, avec un sous-ensemble de 55 tweets validés par le psychiatre. Nous avons calculé un kappa de Cohen de 69,1%, qui souligne un accord substantiel entre les annotateurs. Nous avons également calculé un kappa pondérée, qui tient compte des différents niveaux de désaccord, de l'ordre de 71,5% qui est un taux largement satisfaisant pour juger l'accord entre les annotateurs. Le processus d'annotation a donné 141 instances de la classe *aucune détresse*, 110 instances de la classe *détresse minimale* et 256 instances de la classe *détresse sévère*.

4.2 Protocole d'évaluation

Afin d'évaluer notre approche, nous avons adopté une méthodologie entièrement automatisée basée sur une validation croisée (k=5) sur l'ensemble des données annotés afin d'apprendre et tester le modèle proposé. Pour ce faire, à chaque itération, nous avons partitionné l'ensemble des 507 tweets annotés en échantillons d'apprentissage (70%) et de test (30%). Chaque instance est constituée par le tweet d'un utilisateur avec comme contexte l'ensemble des tweets formant la séquence des publications de l'utilisateur. L'objectif principal de la phase d'apprentissage est d'apprendre les paramètres de notre modèle ainsi que ceux des méthodes de référence (*baselines*). Nous comparons notre approche avec les deux modèles SVM et Random Forest en utilisant les mêmes ensembles d'apprentissage et de test. Nous avons utilisé les mesures d'évaluation : Rappel, Précision et F-mesure.

4.3 Résultat et discussion

4.3.1 Analyse des caractéristiques

Nous avons exploré l'importance des caractéristiques extraites à partir des messages en fonction des trois états émotionnels considérés dans la figure 2. Cette analyse a été effectuée sur l'ensemble des données d'apprentissage. Il est à noter que cette analyse est exploratoire, et nous ne l'avons pas utilisée pour la sélection des attributs les plus pertinents.

Chaque poids donne l'importance de la caractéristique pour l'état émotionnel. Alors que les poids élevés (positifs ou négatifs) indiquent une association forte, des pondérations nulles indiquent que la caractéristique a peu ou pas d'impact sur l'état émotionnel.

Dans ce travail, nous avons considéré 39 caractéristiques pour chaque message, mais pour des raisons de présentation, nous ne rapportons ici que les caractéristiques les plus importantes. Les *pronoms à la première personne* (lexical), les *mots d'injures* (lexique), les *Symptômes* (lexique) et *Symptômes antérieurs* (lexique) ont un impact négatif sur l'état émotionnel "Aucune détresse". En effet, les individus les moins à risque n'utilisent généralement pas de

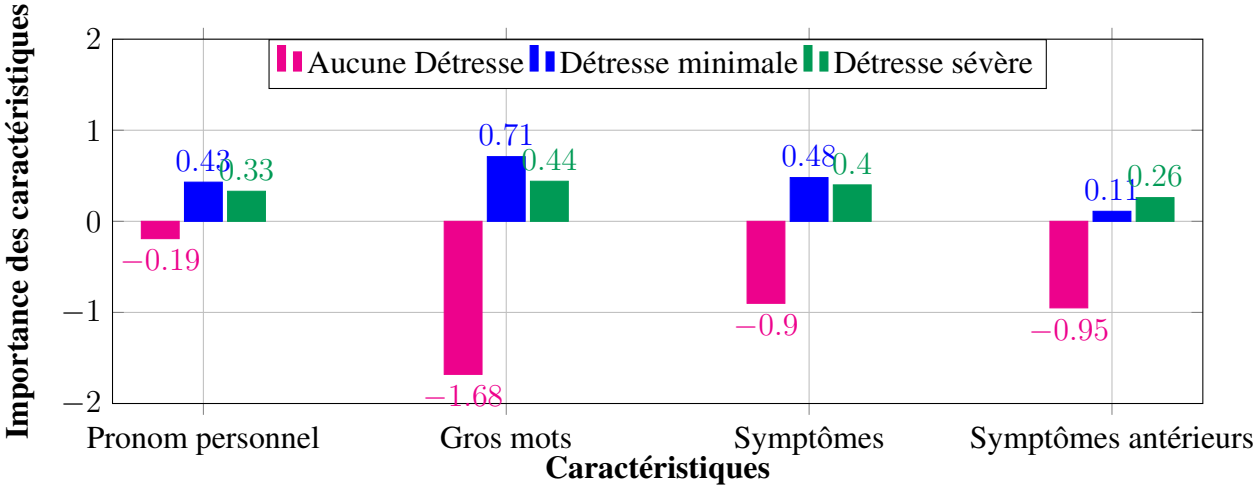

FIGURE 2 – Importance des caractéristiques dans les trois classes.

références aux symptômes ni de mots d'injures dans leurs messages. Au contraire, les injures et les symptômes sont des marqueurs importants pour la classe *détresse minimale* avec une légère différence pour la classe *détresse sévère*. D'autre part, la caractéristique *Symptômes antérieurs* qui reflète les symptômes antérieurs observés, est importante pour la classe *détresse sévère*. En effet, si à l'instant *t-1*, si un individu publie un message ayant une valeur importante pour la caractéristique *Symptôme*, il est probable que l'utilisateur présente une détresse émotionnelle sévère au moment *t*. La possibilité de prendre en compte ce dernier point est un avantage clé des modèles CRF.

Les tableaux 1 et 2 montrent les thèmes extraits des tweets appartenant aux états émotionnels *Aucune détresse* et *Détresse sévère*, en utilisant le modèle Latent Dirichlet Allocation (Hoffman *et al.*, 2010). Par souci de simplicité, nous avons fixé le nombre de thèmes à 2. Dans le tableau 1, les messages de l'état émotionnel *Aucune détresse* portent clairement sur la *famille*, les *voyages*, les *relations*, etc. alors que dans la Table 2, les thèmes abordés sont liés aux pensées suicidaires (e.g. suicide, meurtre, etc.). On remarque l'importance des intensificateurs appliqués aux termes reflétant des idées liées à la fin de vie (e.g. *assez, sans valeur, plus, fin*). Les figures 3 et 4 présentent les termes les plus utilisés pour les deux états émotionnels. Ces dernières corroborent les conclusions du modèle LDA.

4.3.2 Analyse des changements d'états émotionnels

Pour mieux comprendre les changements de comportement des utilisateurs, nous exploitons la puissance des CRF pour analyser les changements entre les états émotionnels. La figure 5 montre les transitions entre les 3 états émotionnels que nous avons considérés, en se basant sur l'ensemble des données d'apprentissage. Cette figure 5 permet d'identifier que les transitions les plus probables entre deux états différents vont de la classe *Aucune détresse* à la classe *Détresse minimale*, avec une probabilité inférieure pour la transition opposée. Les individus passant à un état émotionnel plus risqué sont peu susceptibles de revenir à un état normal. Les utilisateurs dans l'état *Aucune détresse* et *Détresse sévère* tendent à rester dans le même état avec des valeurs

Thème	Description
1	crazy tonight today morning hot **game** road **excitingtimes** haftflmdic **family-travel** arkansas cards poolside be life want up if day good everyone my yourself blessed hidline **girlideas** really **lol** know finally
2	**love** my me **happy** im bed **relationships** can always new better we few little snapchat going wait **excited** school suck drift tired makes exactly pat older attractive **dearly travel** catch

TABLE 1 – Thèmes extraits des tweets de la classe *Aucune détresse*.

FIGURE 3 – Aucune Détresse.

Thème	Description
1	me if would up one really **fucking** will **no** everyone am give **enough** days **everything** good **suicidal** still my seems **suicide** like **never** be **fuck** alive **die** say **help sleep**
2	**want myself tired** can my don **anymore** be wish not **kill worthless** never **feel die** will am **hate** end time **sorry** much me like **enough** live everything many **fight depressionproblems**

TABLE 2 – Thèmes extraits des tweets de la classe *Détresse sévère*.

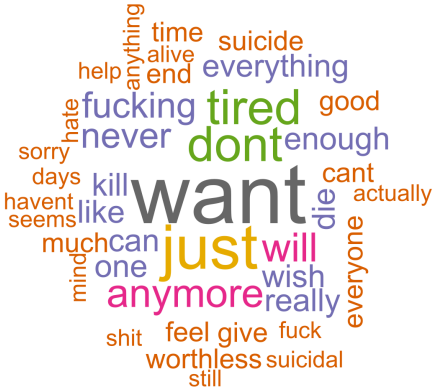

FIGURE 4 – Détresse sévère.

de 0,54 et 1,41, respectivement. La valeur de probabilité très faible pour la transition de *Détresse sévère* à *Détresse minimale* s'explique car l'état émotionnel *Détresse sévère* est généralement atteint lorsque l'individu se focalise sur une possible tentative de suicide (Adler *et al.*, 2016).

La Table 3 fournit les résultats des mesures de précision, de rappel et le F1-score pour toutes les classes, ainsi qu'une comparaison avec les deux modèles d'apprentissage automatique SVM et Random Forest. Dans la phase d'apprentissage, ces modèles sont construits en utilisant les paramètres par défaut de WEKA [11]. Pour apprendre notre modèle CRF, nous exploitons une descente de gradient en utilisant la méthode L-BFGS. Les coefficient de régularisation L1 et L2 du modèle sont fixés par défaut à 0.1 et 0.1, respectivement.
Les meilleurs résultats sont ceux des classes *Aucune détresse* et *Détresse sévère* selon le score F1, en dépit de la valeur importante de précision pour l'état émotionnel *Détresse minimale*. En

11. www.cs.waikato.ac.nz/ml/weka/

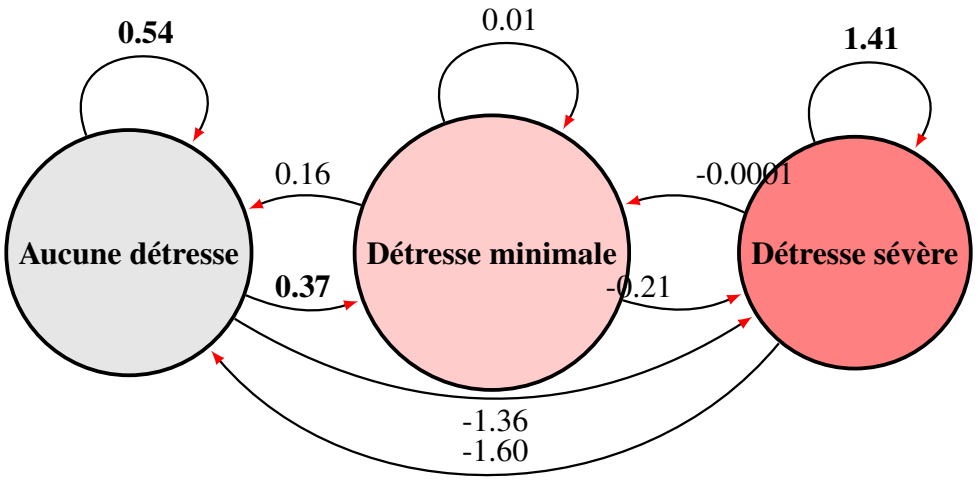

FIGURE 5 – Changement des émotions des utilisateurs selon les trois classes. Les valeurs en gras sont les plus probables. La couleur symbolise la sévérité de l'état émotionnel.

effet, le taux élevé de faux négatifs pour l'état de détresse minimale peut s'expliquer par le fait que : (i) cet état se trouve entre deux états ; (ii) l'auto-transition pour cet état est très faible (0.01) par rapport aux transitions entrantes et sortantes, en particulier depuis l'état *Aucune détresse*. D'autre part, le modèle permet d'identifier tous les messages appartenant à l'état *Aucune détresse*, ce qui n'est pas surprenant compte tenu de l'analyse de l'importance des caractéristique présentée dans la figure 2 (cf. section 4.3.1). Les résultats présentés dans le tableau 3 sont prometteurs car la plupart des méthodes de classification des textes pour des applications liées au suicide ou à la dépression atteignent à peine 0,7 (Burnap *et al.*, 2015; O'Dea *et al.*, 2015). Les résultats obtenus par les méthodes d'apprentissage automatique sont moins importants avec une grande différence en comparaison avec Random Forest. Les valeurs sont plus faibles en termes de Rappel pour les deux méthodes. Cette différence de performance peut être expliquée par l'absence de sélection d'attributs qui pourrait être considéré comme un désavantage dans les tâches de classification de texte.

	Précision	**Rappel**	**F1-score**
Aucune détresse	0.706	**1.000**	0.828
Détresse minimale	**1.000**	0.176	0.300
Détresse sévère	**0.941**	0.571	0.711
Notre approche	**0.816**	**0.752**	**0.711**
SVM	0.446	0.227	0.301
Random Forest	0.500	0.127	0.202

TABLE 3 – Évaluation des résultats du système de monitoring.

5 Conclusion et Perspectives

Dans cet article, nous avons proposé une approche pour le dépistage des idéations suicidaires basée sur un modèle probabiliste appelé Conditional Random Fields qui permet de modéliser et prédire le comportement en ligne de l'individu comme une séquence d'états émotionnels évoluant au fil du temps. Cette représentation permet d'incorporer un ensemble riche de caractéristiques complexes intégrant le contexte des messages précédents. L'efficacité de l'approche a été évaluée sur des données réelles, sur une collection de tweets publiés par des individus ayant montré des symptômes graves liés au suicide. Ces évaluations préliminaires ont montré que notre modèle est capable de fournir des interprétations complètes de la relation entre les états émotionnels et les résultats en termes de prédictions sont encourageants par rapport à la littérature.

Un avantage de l'approche est que nous pouvons facilement incorporer de nouvelles caractéristiques liées au texte en incluant notamment de nouveaux lexiques mais également non liées aux textes comme des informations contextuelles : l'heure de la rédaction du message, sa longueur, etc mais encore des caractéristiques liées à d'autres médias (images, vidéos...) associés aux messages. En effet, le modèle CRF permet d'incorporer un ensemble riche de caractéristiques représentant le contexte sans se soucier de leurs relations a priori (corrélations positives ou négatives). Cette flexibilité nous permet d'intégrer dans le modèle un ensemble de caractéristiques dont les dépendances peuvent être assez complexes et mal connues.

Un point important consiste à filtrer les références réelles liées au suicide par rapport aux messages de support et de condoléances, ou encore les campagnes de prévention du suicide (Burnap *et al.*, 2015). Par ailleurs, nous pouvons intégrer une représentation plus complexe de l'état émotionnel que seulement trois états. Par exemple, dans une étude clinique qui a été menée sur des jeunes étudiants, Moreno *et al.* (2011) ont établi un ensemble de critères cliniques de suicide qui peuvent être présents dans les publications Facebook. Nous citons à titre d'exemple la dépression, perte d'intérêt/plaisir dans les activités, changements d'appétit, problèmes de sommeil, agitation psychomotrice ou retard, perte d'énergie, sentiment d'inutilité ou de culpabilité, diminution de la concentration, et idées suicidaires.

Références

ADLER A., BUSH A., BARG F. K., WEISSINGER G., BECK A. T. & BROWN G. K. (2016). A mixed methods approach to identify cognitive warning signs for suicide attempts. *Archives of Suicide Research*, **20**(4), 528–538.

BARBOSA L. & FENG J. (2010). Robust sentiment detection on Twitter from biased and noisy data. In *Proceedings of the 23rd International Conference on Computational Linguistics : Posters*, COLING 2010, p. 36–44, Beijing, China : Association for Computational Linguistics.

BRADLEY M. M. & LANG P. J. (1999). *Affective norms for English words (ANEW) : Stimuli, instruction manual, and affective ratings*. Rapport interne, Center for Research in Psychophysiology, University of Florida, Gainesville, Florida.

BURNAP P., COLOMBO W. & SCOURFIELD J. (2015). Machine classification and analysis of suicide-related communication on twitter. In *Proceedings of the 26th ACM Conference on Hypertext & Social Media*, HT '15, p. 75–84, New York, NY, USA : ACM.

COLOMBO G. B., BURNAP P., HODOROG A. & SCOURFIELD J. (2016). Analysing the connectivity and communication of suicidal users on twitter. *Computer Communications*, **73, Part B**, 291 – 300. Online Social Networks.

DE CHOUDHURY M., COUNTS S. & HORVITZ E. (2013). Predicting postpartum changes in emotion and behavior via social media. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI'13, p. 3267–3276, New York, NY, USA : ACM.

GUNN J. F. & LESTER D. (2012). Twitter postings and suicide : An analysis of the postings of a fatal suicide in the 24 hours prior to death. *Suicidologi*, **17**(3), 28–30.

HOFFMAN M. D., BLEI D. M. & BACH F. (2010). Online learning for latent dirichlet allocation. In *Proceedings of the 23rd International Conference on Neural Information Processing Systems*, NIPS'10, p. 856–864, USA : Curran Associates Inc.

KARMEN C., HSIUNG R. C. & WETTER T. (2015). Screening internet forum participants for depression symptoms by assembling and enhancing multiple NLP methods. *Computer Methods and Programs in Biomedicine*, **120**(1), 27–36.

KIM S., LI F., LEBANON G. & ESSA I. A. (2013). Beyond sentiment : The manifold of human emotions. In *Proceedings of the Sixteenth International Conference on Artificial Intelligence and Statistics*, p. 360–369.

LAFFERTY J. D., MCCALLUM A. & PEREIRA F. C. N. (2001). Conditional random fields : Probabilistic models for segmenting and labeling sequence data. In *Proceedings of the Eighteenth International Conference on Machine Learning*, ICML '01, p. 282–289, San Francisco, CA, USA : Morgan Kaufmann Publishers Inc.

LARSEN R. J. & DIENER E. (1992). Promises and problems with the circumplex model of emotion. *Review of Personality and Social Psychology*, **13**(13), 25–59.

LEHRMAN M. T., ALM C. O. & PROAÑO R. A. (2012). Detecting distressed and non-distressed affect states in short forum texts. In *Proceedings of the 2012 Workshop on Language in Social Media*, LSM 2012, p. 9–18, Montreal, Canada : Association for Computational Linguistics.

MAIGROT C., BRINGAY S. & AZÉ J. (2016). Concept drift vs suicide : comment l'un peut prévenir l'autre ? In *16ème Journées Francophones Extraction et Gestion des Connaissances, EGC 2016*, volume E-30, p. 219–230.

MORENO M., JELENCHICK L., EGAN K., COX E., YOUNG H., GANNON K. & BECKER T. (2011). Feeling bad on Facebook : depression disclosures by college students on a social networking site. *Depression and Anxiety*, **28**(6), 447–455.

O'DEA B., WAN S., BATTERHAM P. J., CALEAR A. L., PARIS C. & CHRISTENSEN H. (2015). Detecting suicidality on twitter. *Internet Interventions*, **2**(2), 183 – 188.

PESTIAN J. P., MATYKIEWICZ P. & GRUPP-PHELAN J. (2008). Using natural language processing to classify suicide notes. In *Proceedings of the Workshop on Current Trends in Biomedical Natural Language Processing*, BioNLP '08, p. 96–97, Stroudsburg, PA, USA : ACL.

POULIN C., SHINER B., THOMPSON P., VEPSTAS L., YOUNG-XU Y., GOERTZEL B., WATTS B., FLASHMAN L. & MCALLISTER T. (2014). Predicting the risk of suicide by analyzing the text of clinical notes. *PLoS ONE*, **9**(1).

SPASIC I., BURNAP P., GREENWOOD M. & ARRIBAS-AYLLON M. (2012). A Naïve Bayes Approach to Classifying Topics in Suicide Notes. *Biomedical Informatics Insights*, **5**(1), 87–97.

SUEKI H. (2015). The association of suicide-related twitter use with suicidal behaviour : A cross-sectional study of young internet users in japan. *Journal of Affective Disorders*, **170**, 155 – 160.

SUTTON C. & MCCALLUM A. (2012). An introduction to conditional random fields. *Found. Trends Mach. Learn.*, **4**(4), 267–373.

YIK M., RUSSELL J. A. & STEIGER J. H. (2011). A 12-point circumplex structure of core affect. *Emotion*, **11**(4), 705–731.