Concept drift vs suicide: How one can help prevent the other?

Cédric Maigrot, Sandra Bringay, Jérôme Azé

Concept drift vs suicide: How one can help prevent the other?

Cédric Maigrot[1], Sandra Bringay[2,3], and Jérôme Azé[2]

[1]IRISA Rennes, UMR 6074,
Cedric.Maigrot@irisa.fr, France
[2]LIRMM, CNRS, UMR 5506, France,
{Sandra.Bringay, Jerome.Aze}@lirmm.fr
[3]AMIS, Université de Montpellier Paul Valéry, France

March 18, 2016

Abstract

Suicide has long been a troublesome problem for society and is an event that has far-reaching consequences. 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 behaviours and words that can act as early signs of predisposition to suicide. The objective of this application is to develop a system that semi-automatically detects these markers through social networks. A 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 that training period, we added 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 used this method on synthetic and real data sets issued from the Facebook platform.

1 Introduction

"Every 40 seconds a person dies by suicide somewhere in the world and many more attempt it"[1]. Every part of the world and each age group is concerned, including young people aged 15-29, for whom suicide is the second cause of death globally. We assume in this article that it is possible to prevent some suicide attempts by monitoring the activity of these people using social networks (Twitter, Facebook, etc.).

We propose an extension of the work of [1] to associate a level of risk to a message. In this study, we integrate the time dimension associated with the changing status of the monitored person. For this, we adapted a *concept drift*

[1]Preventing suicide: A global imperative. WHO 2014. ISBN: 978 92 4 156477 9

model [11] to detect these changes of state in the sequence of messages. The goal is to raise an alert as soon as possible when a negative change is detected in the message list, but without over notifying the practitioners, as they are the only authority to make the decision to intervene.

Our process is organized into three steps: 1) Preprocessing and storage of messages with the use of different methods of NLP[2] to extract relevant descriptors of the messages; 2) Detecting the level of risk in the messages from an ensemble learning method (Stacking); 3) Decision to alert the healthcare professional. The alert will be given after a statistical calculation on the modification of a concept, with application of expert rules or with a comparison between ROC curves [13].

The challenges associated with this work are many. NLP methods have to be adapted for specific data from social networks. Messages are written quickly, with different sizes, contain nonconforming grammatical structures, misspellings, abbreviations, specific slang or off topic themes. Our method must be robust enough to understand the vocabulary and the unstructured conversations often used in social media. Additionally, *concept drift* techniques should detect relatively subjective concepts (e.g. anorexia, depression). To our knowledge, there are no sufficiently developed French language resources such as lexicons to capture these concepts. The result of the process must be clearly explained to health professionals to help them in the decision making process.

The originality of this work is located at different levels. Firstly, the work is based on real stories of people with a diagnosed risk behaviour. They are people who posted Facebook messages in thematic groups dealing with risk.

Secondly, the process takes into account a number of levels of description: one evoked risk concepts in a message (e.g expression of loneliness, suicidal thoughts, anorexic behaviour, etc.), the level of risk defined according to the protocol of our partner (the Organisation *Hope for Education*[3] working with people assessed as high risk) and the alert to transmit or not, to the medical team in charge of monitoring them.

Third, the decision to alert the healthcare professional or not, that is to say the information on the evolution of the patient's condition (good or bad) for the doctor, based on three approaches, one based on observation of the disagreement between several classifiers.

The article is organized as follows : Section 2 presents the work related with our study. Section 3 presents the methodology implemented. Section 4 presents the experimental protocol used. Finally, section 5 shows the results.

2 State of the art and motivations

The social web searches for medical perspectives.

Health is an area where the mining of social networks allows us to consider real medical perspectives. Since 2015, more than 2 billion users have been active on social networks. Facebook is the most visited with 1.5 billion users, followed by other networks, including Twitter with about 300 million users. These networks are used to share thoughts, opinions and feelings with friends, especially by young people. Over the past five years, there has been a growing

[2]NLP : Natural Language Processing

[3]`http://www.hopeforeducation.ngo/`

interest in such networks as a tool for public health, for example to analyse the propagation of influenza [20]. [18] use mathematical models to capture symptoms and possible treatments for ailments mentioned on Twitter to define public health measures. By manually examining a large number of tweets, [15] showed that self-reported symptoms are the most reliable signal to predict the onset of a disease.

Mining the social web for the detection of mental illness and suicidal tendencies.

As part of this study, we focus on the potential of social networks to monitor people with a risk of suicide. Other research work on mental illness in general exists [7], [9]. In these applications, a person is considered to be at risk according to their use of social media. For example, the content of their tweets or updates to their Facebook status, are used to classify people in real time according to risk levels. [17] demonstrated that the status updates on Facebook reveal symptoms of major depressive episodes. All of these studies highlight the potential of social media as a signal source for mental illness. Specifically around the theme of suicide, [5] analysed the teenagers messages on MySpace.com to identify subjects at risk (relationships, mental health, substance use / abuse, suicide methods, statements without context). [14] studied the suicide risk factors discussed on Twitter and found a strong correlation between the examined Twitter data and the data on suicide, adjusted for age. [8] point out that interventions are possible, although the validity, feasibility and implementation remain uncertain because very few studies have been conducted to date in real conditions.

3 Overall process of monitoring the individuals

The patient monitoring process is globally inspired by *concept drift* methods that capture the appearance of new concepts over time. We took the [11] architecture and amended it to incorporate the fact that we do not know the true nature of the data (risk level to predict) without the intervention of a health professional and to integrate expert rules to signal the alert. The process is organized in three independent steps described below.

The first two steps are an individual analysis of the messages, the third analyses the latest posts from the same user and decides whether to notify the doctor, or not.

3.1 Step 1: storage of messages and pretreatments

The purpose of this step is to memorize for each person monitored : 1) the theme of the Facebook group to which they belong (e.g attempted suicide, harassment, anorexia, etc.); 2) the content of their messages, date and time of the creation of their messages, the number of *likes* attributed and the number of comments; 3) comments associated with messages. It is important to note that this approach is also applicable to many other social networks (e.g Twitter, Instagram, Ask, etc.).

As indicated by [4], the outcomes of social networks have linguistic peculiarities that may affect the classification performance. For this reason we applied the following pretreatments:

1) replacing usernames by [NAME]; 2) replacing email addresses by [MAIL]; 3) replacing URLs by [URL]; 4) replacing emoticons by a mood word associated with the symbol (cross reference table created for the study); 5) replacing abbreviations with the entire word(s) (cross reference table created for the study); 6) removing accents and uppercases. Indeed, in order to write faster, both writing conventions are not always used. Putting the text in lowercase without accent can restrict the number of N-grams of words generated and add a link between the messages previously having no N-grams in common; 7) lemmatization with the TreeTagger tool [21].

3.2 Step 2 : detecting the level of risk with two sub-steps

We chose to work within the *ensemble learning* paradigm whose goal is to learn more classifiers to solve the same problem. We used this paradigm for the following reasons: 1) the combination of classifiers generally gives better results [22]; 2) the computing power makes learning more achievable and the use of many additional models; 3) the amount of data to be processed may not be learnable by a single classifier; 4) the need to clarify and classify the results to humans involved in the process of decision making. Various forms of learning exist: *Stacking*, *Boosting* and *Bagging*.

In our context, we chose the *Stacking* approach and we gathered a succession of organized classifiers on two levels and aggregated by a majority vote so that each classifier learned new descriptors for redescribing data. We will expand on this in the following two sub-steps:

Sub-step 1: detection of the concepts in a message

The objective of this sub-step is to detect a first level of information in the: presence or absence of a discomfort signal that we will later name *concept*. The list of *concepts* is : *previous suicide attempt*, *suicidal thoughts*, *depression*, *harassment*, *medication*, *anorexia*, *mutilation*, *anger*, *fear*, *loneliness*, *sadness* and *remission*. The remission concept is, in contrast to other, beneficial to the human condition.

This issue list from [19] work has been validated by psychiatric experts. For each concept, a classifier returns the value *yes* if the presence of the concept in the message is validated. We chose descriptors summarized below, supplemented with specific glossaries for each concept: 1) Message content described by the following statistics: a) a set of N-grams in order to allow a comparison of selected messages with *TF-IDF* values to retain only the discriminating words by concept; b) the number of words associated with a concept given by a lexicon created for the study; 2) Number of comments: A post with risk is likely to raise a lot of feedback from other members of the group; 3) Number of *Likes*: Conversely, a message where the victim explains his or her welfare and (his or her) recovery may be accompanied by many *Like* acknowledgements; 4) Message length: two opposite behaviours of the victims are known. The victim can write longer messages because of a need to confide to others or otherwise close in on himself or herself and write less. The output of this stage, a message is associated with a vector of Boolean concepts.

For example, the message "I feel terrible, please help me fast. I have been harassed for 3 years and I am getting beaten everyday. my name is lea, I am 14 years old and I live in Roubaix" is associated with the vec-

tor : ($\overline{suicide\ attempt}$, $\overline{suicidal\ ideation}$, $\overline{depression}$, *harassment*, $\overline{medication}$, $\overline{anorexia}$, $\overline{mutilation}$, $\overline{anger}$, *fear*, $\overline{loneliness}$, *sadness*, $\overline{remission}$).

Sub-Step 2: calculating the risk level of a message

The six classifiers are then applied to predict the level of risk. The predictions are made on the concepts used as descriptors to predict the risk level of the messages. Risk levels are spread over five levels representing non-risk messages (level 0), with a previous risk (level 1), low risk (level 2), moderate risk (Level 3), high risk (Level 4).

The 5 output levels of risk are difficult to interpret in terms of raising the alarm. We also considered a binary prediction after grouping levels 0 and 1 as *low risk* and levels 2, 3 and 4 as *high risk*.

3.3 Step 3 : signalling an alert

For raising an alert, we decided to compare three models: 1) a model of the concepts derived via conventional estimate of losses on *concept drift*; 2) a model by comparing ROC curves; 3) a model based on expert rules as provided by the *Hope for Education* association.

Model based on the loss estimation. Most *concept drift* algorithms [11] consider tasks for which we know, at a certain time, the real value associated with the data to be predicted. For example, for temperature readings, these are predicted values then we make a measurement. The prediction can then be compared to the true value for estimating losses. It raises an alert if the losses are too high (i.e the predictions are often wrong). In our case, the "truth" is not available without the intervention of the health professional. We have therefore adjusted the calculation made by [3] to estimate losses at a given considering the rate of agreement between the classifiers (i.e the number of classifiers having made the same prediction). The intuition works as follows. We use classifiers based on different logics, when a new concept appears (e.g someone starts to regularly discuss death, when before they never had), classifiers usually do not respond at the same time. They are in agreement (before the change), then disagree (at the time of drift) and again in agreement (when they have all detected the change). In the opposite case of the disappearance of a concept (e.g someone who was posting more joyful messages), it is also interesting that the system detects this change in behaviour. Finally, regardless of changes (positive or negative), it must be reported to the doctor (e.g an improvement in the condition can suggest a modification of the regulation).

Let a sliding window $\mathcal{F}$ containing the last N examples of the user (with the arrival of a new message, the oldest is deleted). The 1^{st} and N^{th} examples describe respectively the oldest and the newest message.

The error rate δ in time t, denoted δ_t is given by:

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

where $goodPrediction(i)$ returns the agreement rate of withholding prediction (i.e the number of classifiers have predicted the good level divided by the total

number of classifiers). The δ represents the classifiers agreement rate. It takes the value 1 when all classifiers predict the same value for the level of risk (when there are no errors). Following the calculation checks that the successive values of δ does not exceed a predefined threshold if an alert is raised. It is important to weight each misclassification by its position in the time window. For this, a *progressive forgetfulness* [6] system is set up. Thus, each error is weighted by seniority:

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

The N last values δ are stored, the delta associated with each of the N messages previously named. An alarm is raised if the average of N δ exceeds a threshold of Δ fixed by the user. If this is the case, all values N δ is transmitted to the change detection module to calculate the "date" at which time the change has occurred and initiate a warning procedure with the psychiatrist following the case.

For the interpretation of the health professional, it is important to point to the time that a concept emerged to assist in their analysis. If a drift model is found, he must learn a new model from that date. For this, we seek Ω as:

- Ω is the index of the example of the set of N examples. Let $1 \leq \Omega \leq N$;
- The difference averages estimation values subsets losses defined by the terminals $[1, \Omega - 1]$ et $[\Omega, N]$ is maximum.

Model based on the comparison of ROC curves. The **ROC curve** (Receiver Operating Characteristic) [16] represents a classifier having the ability to perfectly separate the positive from the negative. We use ROGER algorithm (ROC based Genetic Learner) initially proposed in 2003 as part of the prediction of cardiovascular risk [2].

The ROGER algorithm learns functions of the form: $f(\boldsymbol{x}_i) = \sum_j w_j * \boldsymbol{x}_i(j)$ where $\boldsymbol{x}_i(j)$ represents the value of the j^{th} component of the example $\boldsymbol{x}_i$. The algorithm learns weights w_j such as $\sum_i rank_{f(\boldsymbol{x}_i)} * \mathbb{1}_{y_i=+1}$ is minimal (where $rank_{f(\boldsymbol{x}_i)}$ corresponds to the rank of example $\boldsymbol{x}_i$ induced by the function f, and $\mathbb{1}_{y_i=+1}$ corresponds to the indicator function which is +1 if the class of $\boldsymbol{x}_i$ is positive, 0 else).

In our context, ROGER allows the learning of direct messages from users in decreasing risk. Assuming that there is at least one *concept drift* per user, we believe that the greatest risk messages (i.e the message placed in the first position by the learned by ROGER function) is the message corresponding to drift.

Model based on expert rules. We implemented expert rules provided by the association *Hope for Education* for the raising of the alert and we explored the following scenarios. An alarm is raised if there is: 1) a message with a risk level of 4 in the N last posts; 2) an increase in the level of risk between 2 consecutive times; 3) an increase in the level of risk with a gap of at least 2 levels on an N messages window; We will vary N to assess performance for detecting abrupt or slow changes; 4) the oscillations of the level of risk.

We then get the following rules: Let M all N recent messages when a user M_1 corresponds to the oldest message and M_N to the newest. Note R_i the message risk level i.

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

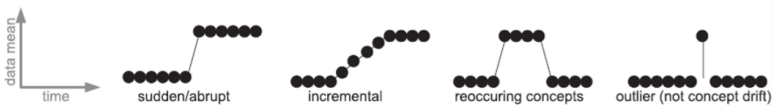

Figure 1: Different *concept drift* types (adapted from [11])

The figure 1 represents different forms of change we want to capture via expert rules: 1) The *sudden* change in an individual who speaks all of a sudden about suicide methods; 2) The *incremental* change corresponds to a situation where the individual speaks more and more about his or her unhappiness; 3) The *recurrent* change is where a user regularly talks about his or her unhappiness; 4) The *outlier* would be an isolated message about suicide.

4 Experimental protocol

4.1 Data used

This study focuses on the analysis of written messages by populations at risk. For this, we used messages from Facebook groups we harvested via the Facebook API[4]. These groups were chosen because they are directly related to the theme of suicide or known risk factors for suicide (*suicide*, *anorexia*, *mutilation* or *harassment*). Authors are usually teenagers between 13 and 18 years old. It is important to note that these groups include older adults who were victims and who testify about their experience as well as parents or relatives who come to seek advice for their child who is possibly being harassed. We collected 4597 messages between March 15^{th} and June 3^{rd} 2015. We are aware that the use of Facebook data is a limitation because Facebook is pretty clear that they will not allow for the redistribution of such data in any form. That means our experiments are not really reproducible. However, Facebook is very popular among young users and its organization in thematic public groups allows us to build scores related to risk factors that are very relevant for health professionals.

As part of our experiments, we manually selected 22 teenagers accounts with between 3 and 14 posts on the groups. They meet the following conditions: 1) The author is a person at risk; 2) The person talks about his well-being; 3) A person is not a group moderator. We then had 168 messages. This low number is due to the difficulty of the annotation data for this study. The *Hope for*

[4] `http://developers.facebook.com`

Education association also provided a protocol used by the volunteers to detect the level of urgency. We have adapted this to our case study which is not limited to cases of harassment. We considered 5 degrees of urgency: 1) *No emergency*: There is no urgency to process the message of the person; 2) *Minimal risk:* The person's problem lies in the past. This situation is unbearable for the person he or she believes that his or her word should be released; 3) *Low Risk* : The person has a problem. They start to isolate themselves; 4) *Significant risk* : The person has a lasting problem that may include violence. An important degree of isolation has occurred; 5) *Absolute Risk* : The person is violent verbally and / or physically. They talk about suicide and / or endangering themselves.

168 messages have been manually added by three people. Every message received a total of 13 annotations: presence or absence of 12 concepts and level of risk. If the size of the dataset is limited, it is important to note that each message has been annotated 13 times (12 concepts and 1 risk level). The annotation of these concepts is a subjective task and a phase of consensus has been applied with several iterations with the annotators. The Fleiss kappa ([10]) that evaluates the consistency in the qualitative assignment objects class to several observers, gives a value of 0.563 for the five levels of risk. The concepts also show interesting Kappa values (eg. 0.702 for the concept *anorexia*). The annotations have been reworked by the annotators to obtain important consensus and therefore a well annotated corpus.

4.2 Evaluations

Detection of the level of risk. In step 1 of the module 2 to detect concepts in a message, we wanted to promote good message classification with a concept. The goal is to find the best configuration that maximizes *recall* class *yes* (i.e minimize the number of messages to be classified as *"yes"* and that are classified as *"no"*). Similarly, the classifier used for the 2^{nd} level must be able to maximize the *recall* class higher level (i.e level 4 in the case of a classification of 5 classes and Level 1 in the case of a binary classification of risk level). It is important to file a message with a high level of risk to be sure to respond to such a message. To achieve the overall learning principle, five classifiers (*J48*, *JRip ASM Naive Bayes*, *IB1*) are used for each prediction thanks to the Weka tool ([12]). The validation is done by cross-validation in Leave-one-out fashion on 168 posts.

Raising the alarm. The warning raised, that is to say, the feedback of information to the health professional making the decision (positive or negative), processes all the messages at the same time for the same user. For this, all the messages associated with a risk level is transmitted in step 3. It is assumed that each user has submitted a risk (i.e an alert should be raised in the sequence of messages). For this, we will impose a rule that the message corresponding to the time of the alert can be raised by the first message posted with no indication of the "normal" state of the person at that moment. Furthermore, if several messages are qualifiable *message* most at risk, the oldest is retained in order to react as quickly as possible.

Loss estimation: The estimation of losses is evaluated using the calculations above. *ROGER:* The algorithm provides the most likely direct message from being at risk. *Expert rules:* The 4 rules are tested on each message. The message having responded positively to the rules maximum is described as the highest risk message.

5 Experiments on real data

5.1 Detecting the level of risk (first sub-step)

Table 1 presents reminder values associated with the detection of the concepts and between parentheses, the values of F-measure. This first layer of detection, which uses information from Facebook and the results of the pretreatments, determines the presence or absence of 12 concepts. Due to the lack of space, we present the results of only 3 concepts *harassment*, *fear* and *loneliness*. The first observation is that each classifier adapts differently to each concept (e.g *Naive Bayes (NB)* is effective in the detection of *harassment*, while the concepts of *fear* and *loneliness* are better captured by the algorithm *J48*). This difference can be explained by specific lexicons to more comprehensive concepts.

	harassment	*fear*	*loneliness*
J48	47,1% (64,8%)	**60,0% (89,6%)**	**61,5% (86,1%)**
NB	**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%)

Table 1: Result classifiers on the detection of harassment, fear and loneliness

5.2 Detecting the level of risk (second sub-step)

The six classifiers are then applied to the second prediction layer (prediction of risk level). For this, four tests are performed: a binary prediction and a prediction separated into five levels of risk, in the case of a prediction based on the concepts and in the case of a classification directly from Facebook messages. Table 2 presents these results by the two measures used to differentiate classifiers: first relying on the highest risk level, and if equal the overall F-measure (in brackets in the table). The first finding is not surprising that the binary classification performs better than the classification in 5 levels of risk. Moreover, we note that our Stacking choice model performs better at prediction than the direct messages from Facebook. Indeed, some classifiers get a relatively high placing on this test (e.g IB1 for direct prediction into five risk levels), but this is achieved by placing a large majority of messages in the class of the highest level, which reveals that it is not very effective in reality, as shown by the F-score associated measurement (e.g 6.6 %)

	2 levels (Stacking)	5 levels (Stacking)	2 levels (Direct)	5 levels (Direct)
J48	92,6% (89,2%)	83,3% (59,2%)	70,4% (57,8%)	23,3% (25,5%)
NB	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%)

Table 2: Result classifiers on the detection of the risk level

5.3 Raising an alert

The three approaches indicate the message corresponding to the passage of the individual in a dangerous state in a series of messages from the individual. Remember that the series of messages were selected because they contained a change of state. Overall, the three approaches are coordinated on the same message in 3 of the 22 individuals (e.g user U_{15}), or very close (e.g user U_6). However, for 6 users, the approach does not work (e.g user U_8). Disagreements arise in the cases of people who alternate between positive and negative messages. It is important to note that none of the three methods are more sensitive than others (i.e systematically before the others). These methods are therefore a complementary way to prevent risks at best.

	U_1	U_2	U_3	U_4	U_5	U_6	U_7	U_8	U_9	U_{10}	U_{11}
Loss Est.	-2	**+1**	+5	**0**	**0**	**0**	**0**	**-4**	+2	+3	+3
ROGER	-2	+4	+6	-6	**0**	-1	-1	+6	**0**	**+2**	+2
Rules	**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}
Loss Est.	**0**	**+1**	**0**	**0**	+2	-1	+4	**0**	**0**	**0**	-3
ROGER	-1	-5	+1	**0**	**0**	+1	**0**	**0**	**0**	-4	-3
Rules	-1	+3	**0**	**0**	+1	**0**	+1	**0**	+1	-2	**-1**

Table 3: Raised alert gap between the message designated by the human and the 3 approaches

6 Conclusion

In this study, we proposed a complete processing operation based on a *Stacking* analysis to assess the level of an individual's risk to commit suicide and if necessary raise the alarm. The first step to this method is to detect 12 concepts in the text of messages, corresponding to known risk factors psychiatrists use. We apply a *concept drift* method to these concepts to detect a change in behaviour. During a raised alert, the concepts are also presented to psychiatrists to help them evaluate the alert and to make a decision whether to intervene or not.

The first limit concerns the size of the data. We plan to collect more messages and in particular to experiment our method with the data provided in the framework of the CLPsych challenge[5] in order to verify the statistical significance of these first results. The second limit concerns the resources used to identify concepts in the messages. A study on the production of French lexicons adapted to each concept and taking into account the familiar vocabulary of individuals on social networks is required to improve the scores of the first-level classifiers. A third limitation concerns the data considered to raise the alarm. We focused here on text messages, but it would also be interesting to consider other signs such as frequency or time of messages and signs of unusual behaviour. We plan to use combinations of classifiers as it seems clear that multiple structures are hidden to highlight this data. Another limit concerns taking into account the type of effect, positive or negative, that can be associated to the appearance or disappearance of a concept (e.g the disappearance of the concept of *joy* or the

[5] `http://clpsych.org/shared-task-2016/`

emergence of the concept of *fear* signals a deterioration in the status, whereas the disappearance of the concept of *sadness* and the emergence of the concept of *joy* indicates an improvement in the state of someone). A final limit of this study lies in the fact that the analysed messages were from non-government groups and completely representative of the general activity of individuals in social networks. We plan to work with a psychiatric service on cases of people who have given their consent, to assess the impact of this approach on the prevention of relapsing. In particular, work on the management of false positives will be realized. Our goal is to minimize them in order to reduce the number of times the doctor is called to false alarms. For this, the raising of alarm threshold must be adjusted. A high threshold (i.e close to 1) is very reactive but would improperly notify the doctor. Conversely, a low threshold (i.e near 0) would rarely ask the doctor to evaluate the alert, but may lead to non identification of critical cases.

References

[1] Amayas Abboute, Yasser Boudjeriou, Gilles Entringer, Jérôme Azé, Sandra Bringay, and Pascal Poncelet. Mining twitter for suicide prevention. In *19th International Conference on Applications of Natural Language to Information Systems*, pages 250–253, 2014.

[2] J. Azé, N. Lucas, and M. Sebag. A new medical test for atherosclerosis detection : Geno. In *Discovery Challenge PKDD 2003*, september 2003.

[3] Stephen Bach and Mark Maloof. A bayesian approach to concept drift. In *Advances in Neural Information Processing Systems*, pages 127–135, 2010.

[4] Alexandra Balahur. Sentiment analysis in social media texts. In *4th workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, pages 120–128, 2013.

[5] Scottye J. Cash, Michael Thelwall, Sydney N. Peck, Jared Z. Ferrell, and Jeffrey A. Bridge. Adolescent suicide statements on myspace. *Cyberpsy., Behavior, and Soc. Networking*, 16(3):166–174, 2013.

[6] Badrish Chandramouli, Jonathan Goldstein, and Abdul Quamar. Scalable progressive analytics on big data in the cloud. In *International Conference on Very Large Databases*, volume 6, pages 1726–1737, August 2014.

[7] Munmun De Choudhury, Scott Counts, Eric Horvitz, and Michael Gamon. Predicting depression via social media. AAAI Conference on Weblogs and Social Media, July 2013.

[8] Helen Christensen, Philip J. Batterham, and Bridianne Dea. E-health interventions for suicide prevention. *International Journal of Environmental Research and Public Health*, 11(8):8193–8212, 2014.

[9] Karthik Dinakar, Emily Weinstein, Henry Lieberman, and Robert Louis Selman. Stacked generalization learning to analyze teenage distress. In *International AAAI Conference on Weblogs and Social Media*, pages 81–90, July 2014.

[10] Joseph L Fleiss. Measuring nominal scale agreement among many raters. *Psychological bulletin*, 76(5):378, 1971.

[11] João Gama, Indrė Žliobaitė, Albert Bifet, Mykola Pechenizkiy, and Abdelhamid Bouchachia. A survey on concept drift adaptation. *ACM Comput. Surv.*, 46(4):44:1–44:37, March 2014.

[12] Mark Hall, Eibe Frank, Geoffrey Holmes, Bernhard Pfahringer, Peter Reutemann, and Ian H Witten. The weka data mining software: an update. *ACM SIGKDD explorations newsletter*, 11(1):10–18, 2009.

[13] James A Hanley and Barbara J McNeil. The meaning and use of the area under a receiver operating characteristic (roc) curve. *Radiology*, 143(1):29–36, 1982.

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

[15] Manuela Krieck, Johannes Dreesman, Lubomir Otrusina, and Kerstin Denecke. A new age of public health: Identifying disease outbreaks by analyzing tweets. In *Proceedings of Health WebScience Workshop, ACM Web Science Conference*, 2011.

[16] Charles E Metz. Basic principles of roc analysis. *Seminars in nuclear medicine*, VIII(4):283–298, Jan 1978.

[17] Megan A. Moreno, Lauren A. Jelenchick, Katie G. Egan, Elizabeth Cox, Henry Young, Kerry E. Gannon, and Tara Becker. Feeling bad on Facebook: depression disclosures by college students on a social networking site., June 2011.

[18] Michael Paul and Mark Dredze. You are what you tweet: Analyzing twitter for public health. *the Fifth International AAAI Conference on Weblogs and Social Media*, pages 265–272, 2011.

[19] Robert Plutchik and Herman M Van Praag. Sucide risk: Amplifiers and attenuators. *Journal of Offender Rehabilitation*, 21(3-4):173–186, 1994.

[20] Adam Sadilek, Henry Kautz, and Vincent Silenzio. Modeling spread of disease from social interactions. In *Sixth AAAI International Conference on Weblogs and Social Media (ICWSM*, pages 322–329, 2012.

[21] Helmut Schmid. Probabilistic part-of-speech tagging using decision trees. In *International Conference on New Methods in Language Processing*, volume 12, pages 44–49, 1994.

[22] Haixun Wang, Wei Fan, Philip S Yu, and Jiawei Han. Mining concept-drifting data streams using ensemble classifiers. In *Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 226–235. ACM, 2003.