Detection of suicide-related posts in Twitter data streams

M. Johnson Vioulès
B. Moulahi
J. Azé
S. Bringay

Suicidal ideation detection in online social networks is an emerging research area with major challenges. Recent research has shown that the publicly available information, spread across social media platforms, holds valuable indicators for effectively detecting individuals with suicidal intentions. The key challenge of suicide prevention is understanding and detecting the complex risk factors and warning signs that may precipitate the event. In this paper, we present a new approach that uses the social media platform Twitter to quantify suicide warning signs for individuals and to detect posts containing suicide-related content. The main originality of this approach is the automatic identification of sudden changes in a user's online behavior. To detect such changes, we combine natural language processing techniques to aggregate behavioral and textual features and pass these features through a martingale framework, which is widely used for change detection in data streams. Experiments show that our text-scoring approach effectively captures warning signs in text compared to traditional machine learning classifiers. Additionally, the application of the martingale framework highlights changes in online behavior and shows promise for detecting behavioral changes in at-risk individuals.

Introduction

According to the World Health Organization (WHO), it is estimated that 800,000 people worldwide die by suicide each year with at least as many suicide attempts [1]. The grief felt in the aftermath of such an event is compounded by the fact that a suicide may be prevented. This reality of suicide has motivated WHO member states to commit themselves to reducing the rate of suicide by a significant percent by 2020 [2].

In an effort to educate the public, the American Foundation for Suicide Prevention (AFSP) [3] has identified characteristics or conditions that may increase an individual's risk. The three major risk factors are: 1) health factors (e.g., mental health, chronic pain), 2) environmental factors (e.g., harassment, and stressful life events), and 3) historical factors (e.g., previous suicide attempts and family history). Additionally, the time period preceding a suicide can hold clues to an individual's struggle. The AFSP categorizes these warning signs as follows: 1) talk (e.g., mentioning being a burden or having no reason to live), 2) behavior (e.g., withdrawing from activities or sleeping too much or too little), and 3) mood (e.g., depression or rage).

Identifying these risk factors is the first step in suicide prevention. However, the social stigma surrounding mental illnesses means that at-risk individuals may avoid professional assistance [4]. In fact, they may be more willing to turn to less formal resources for support [5]. Recently, online social media networks have become one such informal resource. Research has shown that at-risk individuals are turning to contemporary technologies (forums or micro-blogs) to express their deepest struggles without having to face someone directly [6, 7]. As a result, suicide risk factors and warning signs have been seen in a new arena. There are even instances of suicide victims writing their final thoughts on Twitter, Facebook, and other online communities [8, 9].

We believe that this large amount of data on people's feelings and behaviors can be used successfully for early detection of behavioral changes in at-risk individuals and may even help prevent deaths. Social computing research

Digital Object Identifier: 10.1147/JRD.2017.2768678

has focused on this topic in recent years [6, 9, 10]. However, few initiatives have been concerned with the real-time detection of suicidal ideation on Twitter. Previously proposed detection methods rely heavily on manually annotated speech, which can limit their effectiveness due in part to the varying forms of suicide warning signs in at-risk individuals [6, 11, 12]. Many of these methods also focus on the messages published by individuals at a specific time, independent of the whole context, which may be represented by the sequence of publications over time.

In this paper, we address the challenge of real-time analysis of Twitter posts and the detection of suicide-related behavior. To process the stream of an individual's online content, we implement a martingale framework, which is widely used for the detection of changes in data stream settings. The input into this framework is a series of behavioral features computed from each individual Twitter post (tweet). These features are compared to previously seen behavior, in order to detect a sudden change in emotion that may indicate an elevated risk of suicide.

The main contributions of this paper are twofold. First, using research from the field of psychology, we design and develop behavioral features to quantify the level of risk for an individual according to his online behavior on Twitter (speech, diurnal activities, size of social network, etc.). In particular, we create a feature for text analysis called the *Suicide Prevention Assistant* (*SPA*) text score. Second, we monitor the stream of an individual Twitter user and his behavioral features using an innovative application of a martingale framework to detect sudden behavioral changes.

Literature review

The definition and identification of risk factors and warning signs lie at the core of suicide prevention efforts. In this paper, we have chosen to reference the risk factors defined by the American Psychiatric Association (APA) [13] and the warning signs identified by the American Association of Suicidology (AAS) [14]. These resources represent a level of consensus between mental health professionals and also provide a rich discussion of the differences between suicide risk factors and warning signs. For further reading, we direct the reader to the work of [14].

As highlighted by [14], warning signs signify increased imminent risk for suicide (i.e., within minutes, hours, or days). According to the APA, suicide warning signs may include talking about dying, significant recent loss (death, divorce, separation, or broken relationship), change in personality, fear of losing control, suicide plan, suicidal thoughts, or no hope for the future. As discussed in the following paragraphs, recent research has shown the emergence of such signs on social networking sites.

Most of the research at the intersection of behavioral health disorders and social media has focused on depression detection in online communities, specifically Major Depressive Episodes (MDE). However, the risk factors for suicide defined by the APA [13] go far beyond depression alone. It is important to remember that depression does not necessarily imply suicidal ideation. Rather, suicide should be thought of as a potential end symptom of depression.

While mental health issues such as depression, suicidal ideation, and self-mutilation are defined medically as separate illnesses with overlapping symptoms, the approaches proposed to detect them online can be quite similar. The approaches vary in the data they are treating, i.e., Facebook posts, Twitter tweets, Reddit forum threads, etc., and the specific event they are attempting to predict. Moreno et al. [7] first demonstrated that social networking sites could be a potential avenue for identifying students suffering from depression. The prevalence rates found for depression disclosed on Facebook corresponded to previous works in which such information was self-reported. On a larger scale, Jashinsky et al. [15] showed correlation between Twitter-derived and actual United States per-state suicide data. Together, these works established the presence of depression disclosure in online communities and opened up a new avenue for mental health research.

De Choudhury et al. [6] explored the potential to use social media to detect and predict major depressive episodes in Twitter users. Using crowd-sourcing techniques, the authors built a cohort of Twitter users scoring high for depression on the CES-D (Center for Epidemiologic Studies Depression Scale) scale and for other users scoring low. Studying these two classes, they found that what is known from traditional literature on depressive behavior also translates to social media. For example, users with a high CES-D score posted more frequently late at night, interacted less with their online friends, and had a higher use of first-person pronouns. Additionally, online linguistic patterns match previous findings regarding language use of depressed individuals [16]. More recently, De Choudhury et al. [10] have shown that linguistic features are important predictors in identifying individuals transitioning from mental discourse on social media to suicidal ideation. The authors showed a number of markers characterizing these shifts, including social engagement, manifestation of hopelessness, anxiety, and impulsiveness based on a small subset of Reddit posts.

Coppersmith et al. [17] examined the data published by Twitter users prior to a suicide attempt and provided an empirical analysis of the language and emotions expressed around their attempt. One of the interesting results found in this study is the increase in the percentage of tweets expressing sadness in the weeks prior to a suicide attempt, which is then followed by a noticeable increase in anger and sadness emotions the week following a suicide attempt. In the same line of research, O'Dea et al. [18] confirmed that Twitter is used by individuals to express suicidality and demonstrated that it is possible to distinguish the level of concern among

suicide-related tweets, using both human coders and an automatic machine classifier. These insights have also been investigated by Braithwaite et al. [19], who demonstrated that machine learning algorithms are efficient in differentiating people who are at a suicidal risk from those who are not. For a more detailed review of the use of social media platforms as a tool for suicide prevention, the reader may refer to the recent systematic survey by Robinson et al. [20].

These works have shown that individuals disclose their depression and other struggles to online communities, which indicates that social media networks can be used as a new arena for studying mental health. Despite the solid foundation, the current literature is missing potential key factors in the effort to detect depression and predict suicide. Currently, few works analyze the evolution of an individual's online behavior. Rather, the analysis is static and may take into consideration one post or tweet at a time while ignoring the whole context. Additionally, an individual's online "speech" is often compared to other individuals and not to their own linguistic style. This is a disadvantage because two individuals suffering the same severity of depression may express themselves very differently online.

A general framework for detecting suicide-related posts in social networks

In this section, we present the proposed framework for the analysis and real-time detection of suicide-related posts on Twitter. First, we introduce the real-time detection problem. Then, we define our online proxy measurements (behavior features) for suicide warning signs. Finally, we describe the approach we implement for detecting behavioral change points.

Problem statement

Sudden behavioral change is one of the most important suicide warning signs. As reported by the AFSP, a person's suicide risk is greater if a behavior is new or has increased, especially if it is related to a painful event, loss, or change. Considering this in conjunction with social media, where users constantly publish messages and deliberately express their feelings, we address suicide warning sign detection as a real-time data stream mining problem. Given a series of observations over time (tweets, messages, or blog posts), the task is to detect an abrupt change in a user behavior that may be considered as a suicide warning sign. In the field of data stream mining, this can be specifically seen as change point detection problem [21, 22]. However, unlike retrospective detection settings [23, 24], which focus on batch processing, here we are interested in the setting where the data arrives as a stream in real time.

To address this challenge, we chose an approach employing a martingale framework for change point detection [25]. This algorithm has been successfully applied to detecting changes in unlabeled data streams, video-shot change detection [26], and, more recently, in the detection of news events in social networks [27]. To the best of our knowledge, this is the first attempt to apply the martingale framework on a multi-dimensional data stream generated by Twitter users.

In the following section, we start by introducing and describing the proxy measurements for suicide warning signs that we use to assess the subject's level of suicide risk. As previously mentioned, these warning signs will be the input into the martingale framework.

Suicide warning signs in online behavior

To identify online behaviors that may reflect the mental state of a Twitter user, we established two groups of behavioral features: user-centric and post-centric features [11, 28]. User-centric features characterize the behavior of the user in the Twitter community, while post-centric features are characteristics that are extracted from the properties of a tweet. These features have been shown to successfully aid in determining the mental health of a user [6]. **Table 1** shows a detailed description of the features we selected.

The AAS identifies withdrawing from friends, family, or society as one of the warning signs of suicide. With the user-centric behavioral features, we aim to capture changes in a Twitter user's engagement with other users. The *friends* and *followers* features can quantify an individual's interaction with his or her online community, such as a sudden decrease in communication. On the other hand, they can also reflect an expansion of an individual's online community. This is relevant, as at-risk individuals have also been shown to increase their time online developing personal relationships [29]. It is important to note that we have chosen the terms *friends* and *followers* to represent the unidirectional relationships that are inherent on Twitter. We acknowledge that this term may not apply for certain user accounts such as celebrities and news outlets.

Additional features include *volume, replies*, *retweets*, and *links*, which were all identified by De Choudhury et al. [6] as markers for mental health. These measures can help to quantify the number of interactions a user has with their friends and followers for it could be the case that an individual's social network remains stable while their interactions increase or decrease. The final user-centric feature, *questions,* may also indicate a user's attempt to engage with others online.

Post-centric behavioral features are characteristics originating from the post itself. One important piece of information is the hour at which the tweet is published (*time* feature). Late-night activity can be an indication of unusual rhythms in sleep (insomnia and hypersomnia) [6] and can predict future episodes of depression. In addition to the time feature, we address the text of the post (*text score*), which holds the most vital information pertaining to an individual's current mood and mental health [30].

Table 1 Description of the behavioral features. (For the Text Score calculation and Distress Classifier mentioned at the bottom of the table, see the "Feature extraction for text scoring" sections of this paper.)

Feature	*Definition*
User-centric features	
Friends	The total amount of friends at the time of post. A friend is defined as another Twitter user that the author is following online (out-link).
Followers	The total amount of followers at the time of the post. A follower is another Twitter user that is following the author online (in-link).
Volume	The number of tweets per hour; retweets are included in this feature.
Replies	The number of tweets per day directed at another Twitter user. An author may post a tweet destined for another user by including "@" plus the user's screen name.
Retweets	The number of tweets per day that are retweeted. A retweet is defined as a tweet previously composed by another Twitter user and that is re-published or shared.
Links	The number of tweets per day that include a URL
Questions	The number of tweets per day posing a question.
Post-centric features	
Time	The hour at which the tweet was published.
Text score	Text score (based on NLP)/distress classifier

To classify the text of the post, we propose two different approaches. The first approach is a natural language processing (NLP) method that combines features generated from the text, based on an ensemble of lexicons. These lexicons are composed of linguistic themes commonly exhibited by at-risk individuals. The second approach, called the distress classifier, is based on machine learning. Although machine learning is commonly used to classify text, the supervised algorithms require annotated datasets, which may be costly in terms of time and potential annotator error. Additionally, traditional machine learning methods are difficult to apply in this context because of the nature of depression and distress in general. Two individuals suffering from depression may not express their symptoms in the same way, which translates to texts exhibiting the same level of depression or distress having vastly different content. This means it is difficult for the algorithm to find the concept mapping between the textual features and the level of depression/distress.

Feature extraction for text scoring: A natural language processing-based approach

To extract and compute features using NLP techniques, we start by creating a new symptom lexicon to identify the most discriminating terms commonly used by distressed or depressed individuals. Instead of manually translating questionnaires and generating synonyms [31], we decided to create the lexicon directly from a collection of tweets. For this purpose, we implemented the pointwise mutual information (PMI) [32] measure, which highlights the dependence between two random variables. This measure is formally defined as follows:

$$PMI(w,c) = \log\left(\frac{P(w,c)}{P(w)P(c)}\right),$$

where w is a word, c is a class, and $P(w,c)$ is the probability that word w occurs in class c. $P(w)$ is the frequency of the word across all classes, and $P(c)$ is the frequency of class c.

Our final PMI symptom lexicon is built using an annotated dataset of tweets belonging to two classes of users, distressed and everyday users. This dataset is discussed in more detail in the evaluation section. In the end, each word in the lexicon is assigned two scores, one reflecting the word's dependence on the class distressed and the other on the class every day. While a positive PMI score indicates that the word is a discriminating term for the distressed class, a negative score indicates that it is a discriminating term for the everyday class.

Note that we moved from the subject of depression (medical term) to the subject of distress, which is a more encompassing term. In fact, depression is characterized by one or more major depressive episodes, which translates to at least two weeks or more of depressive mood or loss of interest accompanied by at least four other symptoms of depression [33]. We hypothesize that the distressed class includes users that discuss suicidal ideation, depression, and self-harm.

To enrich the lexicon, and go beyond depression-only symptoms, we propose three other features: swear words, intensifier terms (very or extremely), and first-person pronouns (me, myself, and I). These features have been shown to carry important information in the context of sentiment analysis [28] and were similarly used by De Choudhury et al. [6].

To compute the final score reflecting the level of distress in a tweet, we aggregate the four features to obtain our *Suicide Prevention Assistant* (*SPA*) text score. The score is calculated using the following linear combination:

$$SPA = f_{\text{symptoms}} + f_{\text{swear}} + f_{\text{intensifiers}} + f_{\text{first\_pronouns}},$$

where f_{symptoms} represents the sum of PMI scores of every word in a tweet that appears in the symptom lexicon, and the f_{swear}, $f_{\text{intensifiers}}$, and $f_{\text{first\_pronouns}}$ components are the frequency of symptoms, intensifying adjectives, and the first-person pronouns, respectively, in a tweet. Finally, the *SPA* text score is normalized for each tweet by dividing by the number of total words. This helped us control for longer tweets that might have an inflated *SPA* text score because of more symptom words.

Feature extraction for text scoring: Classification with machine learning

In general, the challenge of categorizing tweets is traditionally addressed using text classifiers that rely on machine learning. Therefore, we considered it important to also test a classifying algorithm as a benchmark against our NLP approach detailed above. We took inspiration from [12, 34] and chose to categorize tweets according to different levels of distress. Again, although distress is not equivalent to suicide ideation and major depressive episodes, it is an important risk factor in suicide and one that is highly observable from micro-blog text [35].

In addition to the four features used in the previous section (f_{symptoms}, f_{swear}, $f_{\text{intensifiers}}$, and $f_{\text{first\_pronouns}}$), we split the text into n-grams, which are commonly used in text classification tasks and are popular as a base feature for sentiment analysis of tweets. The character limitation (140 maximum) of tweets lends itself to a choice of shorter n-grams, particularly uni-grams and bi-grams [36, 37].

Design of a martingale-based approach for emotion change detection

In the previous sections, we described the user-centric and post-centric behavioral features we extract from a user's online content. In this section, we present the approach we use to process these behavioral features and detect sudden emotional changes. If we consider the features as a series of multi-dimensional observations over time, the challenge of detecting an abrupt change in behavior resembles the classic problem of change point detection often seen in the field of data stream mining.

We implemented a martingale framework [25], which is an online real-time, non-parametric change point detection model. Let $\mathbf{X} = \{\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3 \ldots \mathbf{x}_n\}$ be a sequence of unlabeled m-dimensional data points with new data points, x_i, arriving in a sequence. The m dimensions correspond to the values of the user-centric and post-centric behavioral features identified in the previous section. When a new tweet is published, it is first characterized by this set of features. With this information, the tweet is then run through a hypothesis test to determine if its features represent a prominent change in the data stream. More formally, the test is as follows: H_0, if there is no change in the data stream (i.e., no marked emotional change), and H_1 otherwise [38].

The full martingale framework can be broken down into three steps. The first step is to calculate the strangeness measure, which quantifies for each specific user how much a tweet is different from previous ones. Next, a statistic is defined to rank the strangeness measures of the tweets. Finally, using this statistic, a family of martingales is defined in order to detect movements in the tweet stream and run the hypothesis test.

Unified strangeness measure

Given that the behavioral features we have chosen are represented as numerical attributes, we chose the Euclidean distance to measure the distance of each data point from the mean of the other data points. For a stream of unlabeled tweets, $\mathbf{X} = \{\mathbf{x_1}, \mathbf{x_2}, \mathbf{x}_3 \ldots \mathbf{x}_{n-1}\}$, where x_n is the most recent tweet in the data stream, the *unified strangeness measure* (*USM*) is defined as follows:

$$USM_i(X, x_n) = \sum_{k=1}^{m} \left| x_{ik} - \mu\left(\{x_{1k}, \ldots, x_{(n-1)k}\} \cup \{x_{nk}\}\right) \right|,$$

where $i : 1 \ldots n, m$ is the number of (user-centric and post-centric) behavioral features, x_{ik} is the value of the k-th feature for tweet x_i and μ is the *mean* of the previous values and the new observed value with respect the feature k. The USM for a tweet increases as it "moves away" from the representative of the entire stream of data. For continuous attributes, the representative value is the mean, while for categorical attributes, we use the mode as the representative.

In the second step of the martingale framework, we rank the USM of the new point with respect to the USM of the previously observed points using a statistic. This statistic is denoted as the $\hat{p}$-value for each instance x_i. Formally, the $\hat{p}$-value of x_i, for $i : 1 \ldots n$, can be calculated as follows:

$$\hat{p}_i(\{x_1, \ldots, x_n\}, \theta_i) = \frac{\#\{j : s_j > s_i\} + \theta_i \#\{j : s_j = s_i\}}{i},$$

where s_j is the USM for x_j, θ_i is a random value in [0,1] at instance i, and $\#\{\}$ is the number of elements satisfying the bracketed condition [25]. The random number θ_i ensures that the p-values are distributed uniformly in [0,1] provided that the input observations are generated by an exchangeable probability distribution in the input space.

When a change in the data stream occurs, this translates to high USM values for the observations following the change point. In turn, the computed $\hat{p}$-value become smaller, which means that the incoming data points are moving away from the representatives of the previously seen observations.

To decide whether there is an abrupt change in the user behavior or not (i.e., reject the null hypothesis H_0), a family of martingales is defined based on the derived $\hat{p}$-values. We use the *randomized power martingale* [27], defined as follows:

$$M_n^{(\varepsilon)} = \prod_{i=1}^{n} (\varepsilon \hat{p}^{\varepsilon-1}),$$

where the martingale value M_i measures the confidence of rejecting the null hypothesis of exchangeability, and ε is a value in [0,1] that controls the transitions between the current and the previous martingale. Suppose that $\{M_k : k \geq 0\}$ is a non-negative martingale. If $E(M_n) = E(M_0) = 1$, then from *Doob's Maximial Inequality* [25]:

$$\lambda P\left(\max_{0 \leq k \leq n} M_k \geq \lambda\right) \leq E(M_n)$$

and for any $\lambda > 0$ and $n \in N$, one has $P(\max_{k \leq n} M_k \geq \lambda) \leq \frac{1}{\lambda}$.

This inequality implies that it is unlikely for M_k to have a high value. Given this, the null hypothesis H_0 is rejected when the martingale value is greater than λ.

In order to detect the movements of the behavioral features towards or away from their representative value, we average two martingale sequences [38]. The first martingale is as defined above, and the second one is calculated using the complement of the $\hat{p}$-value at each observation (i.e., $1 - \hat{p}_i$). These two martingales sequences denoted M_1 and M_2 are averaged to generate the final martingale sequence that is monitored for values exceeding the threshold λ. One of the interesting properties of the martingale framework is the recursive property of the martingale sequence. Note that

$$M_n^{(\varepsilon)} = \varepsilon(\hat{p}^{\varepsilon-1}) M_{n-1}^{(\varepsilon)}.$$

In this way, it is not necessary to store previous $\hat{p}$-values. We only save the previous M_1 and M_2 values for the next hypothesis test upon the arrival of a new point.

Our motivation for applying the martingale framework is that it is a flexible approach to analyzing text streams. The algorithm can handle different mixes of features and can be implemented without the use of annotated datasets, which are common to text classification.

Again, to the best of our knowledge, this is the first time that the martingale framework is applied to a stream of Twitter data generated from a single user. De Choudhury et al. [6] used features generated from a user's Twitter history. However, the authors still employed a classifier to group users into two classes. Our approach is an improvement in that it considers an individual's behavior with respect to his or her own history and not with respect to other individuals.

Experimental evaluation

Data collection and annotation

Due to the absence of publicly available datasets for the evaluation of suicide detection in social media, we used the Twitter streaming API to collect tweets. To evaluate our methodology, we needed two datasets: a sufficiently large annotated set to create the PMI-scored symptom lexicon (cross-sectional) and another smaller set of selected Twitter users and their history (longitudinal) to test the martingale framework.

We used the Twitter streaming API to collect tweets containing key phrases generated from the APA's list of risk factors and AAS's list of warning signs related to suicide. We randomly investigated the authors of these tweets to identify 60 distressed users who frequently wrote about depression, suicide, or self-mutilation. We also randomly collected 60 "everyday" users. A professional with expertise in mental health research validated the selection of these distressed and non-distressed users. For each set of users, we collected at most the last 50 tweets to create a database of 5,446 tweets, of which 2,381 are from distressed users and 3,065 are from everyday users.

In order to test and compare our two methods for scoring the text of a tweet (NLP and machine learning), we randomly selected 250 tweets from the distressed users and 250 tweets from the non-distressed users for a total of 500 tweets. These tweets were removed from the set of 5,446 tweets, and the remaining ones were used to create the PMI-scored symptom lexicon. To create a ground truth set of tweets, eight researchers and a mental health professional manually annotated the 500 tweets using four classes from 0 to 3. The classes are defined as follows:

- 0: *No distress* – text discusses everyday occurrences such as work, going out, weekend activities, etc.
- 1: *Minimal distress* – text expresses distress that could be considered common for most individuals (i.e., exam, presentation for work, argument with friend, etc.)
- 2: *Moderate distress* – text expresses a level of distress that is above what an individual may experience in a normal week (i.e., insomnia, extreme crying, feeling alone, etc.)
- 3: *Severe distress* – text includes mentions of self-harm; suicidal thoughts; gratefulness without direction; apologies; and feelings of worthlessness, self-hate, guilt, or not being good enough, etc.

Each tweet was reviewed by at least two annotators, with a subset of 55 tweets being validated by the psychologist. The annotated data had a Cohen's kappa statistic of 69.1%, which is considered as a substantial agreement among the annotators [39]. The weighted kappa statistic, which takes into account different levels of disagreement, was around 71.5%. Finally, we used the Fleiss kappa to measure agreement for the 55 tweets with three annotators, which gives 78.3%. Overall, the three statistics show strong agreement between the annotators.

In addition to the cross-sectional dataset, we gathered Twitter history from a collection of users. This dataset,

which may be considered as longitudinal data, was collected from 10 unique real users who demonstrated a serious change in speech or online behavior in their Twitter accounts. Two initial validation cases, which ended with the individual committing suicide, were previously identified by [40]. We were able to identify a third fatal case and an additional seven other cases where the individuals demonstrated an abrupt change in behavior. These additional cases were manually chosen, and the abrupt change in behavioral was principally judged by a change in speech.

In order to analyze the evolution of the behavioral features in this longitudinal data, we first identified the change point status or the tweet where the individual showed a change in behavior (namely speech). Using this as a pivot point, we collected 1,000 tweets before and 100 tweets after the change point status, ending up with a dataset of 11,000 tweets. The stream of tweets from each of these users was then used to test the martingale method for change point detection.

Results and discussion

Evaluation of the NLP-based approach

To evaluate lexicon-based NLP approach, we used the cross-sectional set of 500 tweets and looked at the average, maximum, and minimum score given by each distress class. If performing well, the score should separate the tweets according to their annotated class. Ideally, there should be no overlap between the values in each class; that is, all tweets of class 2 should have an overall score higher than the largest score for a tweet of class 1. To visualize this, we plotted the score distribution for each class using a box-and-whisker plot (**Figure 1**). The horizontal black line in the box represents the median observation, the box covers the inter-quartiles range (where 50% of the observations lie), and the whiskers are the maximum and minimum SPA text score for each class.

Figure 1 clearly shows the significant differences in the distribution of classes 0, 1, and 2, which means that the SPA text score is capable of differentiating levels of distress. Class 2 and class 3 have largely overlapping distributions, but we consider this acceptable as both classes indicate a high level of distress.

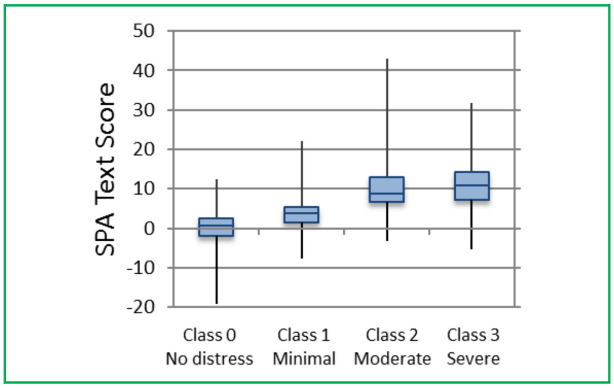

Figure 1

Distribution of SPA text scores by distress class.

Evaluation of the distress classifier

The cross-sectional set of 500 tweets was also used to test the distress classifier for text scoring. We chose to test eight different classification algorithms and compare their performance: multinomial Naïve Bayes, Sequential Minimal Optimization (SMO) with a poly kernel, C4.5 decision tree (J48), nearest neighbor classifier (IB1), multinomial logistic regression, rule induction (Jrip), Random Forest, and SMO with a Pearson VII universal kernel function (PUK).

Given that incorrectly classifying a tweet of class 3 is much more costly than incorrectly classifying a tweet of class 0 or 1, we use the recall for class 3 in addition to the other traditional measures to judge the performance of each classifier. The recall tells us the percentage of the class 3 observations that were correctly classified. If this percentage is low, then there is a risk that the classifier does not catch certain high-distress tweets.

Table 2 presents the best performing algorithms among the eight tested. The experiments were run with a 10-fold cross validation. The results show that the best performing model, in terms of the weighted precision, is SMO with a PUK kernel function, yielding a value of 66.4% with the features *n-grams*, *symptoms*, *pronouns*, and *swear*

Table 2 Best performing classifiers - four distress classes. The bold numbers highlight the best performing models in terms of each evaluation measure using the given behavioral features. The first test for each classifier only included n-grams as the features. We then added the other features, one by one (e.g., + Symptoms), and noted the performance of the classifier.

	SMO (PUK)				*Random Forest*			
Features	*Precision*	*Recall*	*F-measure*	*Class 3 Recall*	*Precision*	*Recall*	*F-measure*	*Class 3 Recall*
n-grams	0.655	0.413	0.315	0.029	0.483	0.468	0.434	**0.265**
+ Symptoms	0.625	**0.431**	**0.334**	**0.039**	0.540	0.553	0.506	0.245
+ Pronouns	0.614	0.427	0.329	0.029	0.541	0.551	0.507	0.245
+ Swear	**0.664**	0.429	0.333	**0.039**	0.554	0.555	**0.518**	0.245
+ Intensifiers	0.615	0.431	0.332	0.029	**0.571**	**0.559**	0.516	0.200

Table 3 Two-step classification results.

Step	*Classifier*	*Precision*	*Recall*	*F-measure*	*Highest Risk Recall (Class 3)*
1	RandomForest	0.676	0.676	0.676	0.341
2	SMO	0.829	0.817	0.823	0.791
1	SimpleLogistic	0.706	0.706	0.706	0.692
2	SMO	0.770	0.762	0.766	0.697
1	J48	0.71	0.71	0.711	0.72
2	SMO	0.807	0.795	0.793	0.692

components. However, we noticed that this promising preliminary result comes at the cost of a very low recall value for class 3.

This trade-off between precision and class 3 recall was seen across all eight classifiers that we chose to test. Noticing this trend, we decided to modify our classification task into a two-step process. In the first step, a tweet is classified as distressed or non-distressed, and the second step separates distressed tweets into two levels: class 2 or class 3.

For both steps, all eight classifiers were tested again. We experimented with three combinations of the best performing classifiers for the two-step process. We tested the Random Forest, Simple Logistic, and J48 classifying algorithms with the SMO algorithm as a second step classifier. The SMO algorithm performed the best when classifying tweets as class 2 or class 3. It had the highest recall for class 3; thus, we chose to keep the SMO as the second classifying algorithm. **Table 3** shows the results for each combination.

The experiments show that our two-step classification process continues to perform well on the testing set. Specifically, we were able to reach a distressed class recall of 72% for the first classification step, which is arguably the most vital. The combination of the J48 and SMO classifiers resulted in the least amount of class 2 and class 3 instances being incorrectly classified as non-distressed (2%). This fact, along with the weighted precision (71% and 80.7%)

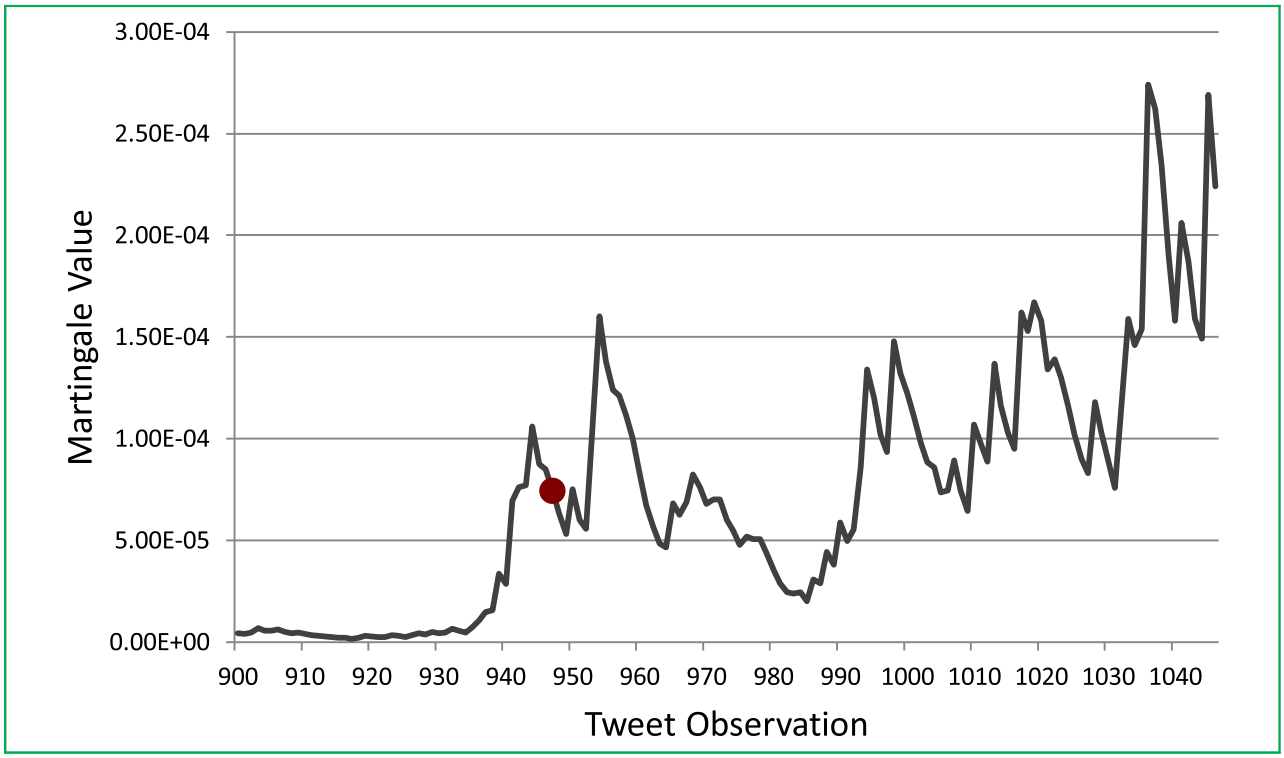

Figure 2

Stream of martingale values for tweets from user *Max*. The values are computed using only the feature SPA text score. The point represents the true abrupt emotion change point (#946: "Done with my life"). The martingale values may begin to increase prior to the change point, but with a sudden change in behavior, the values may only increase after the change point.

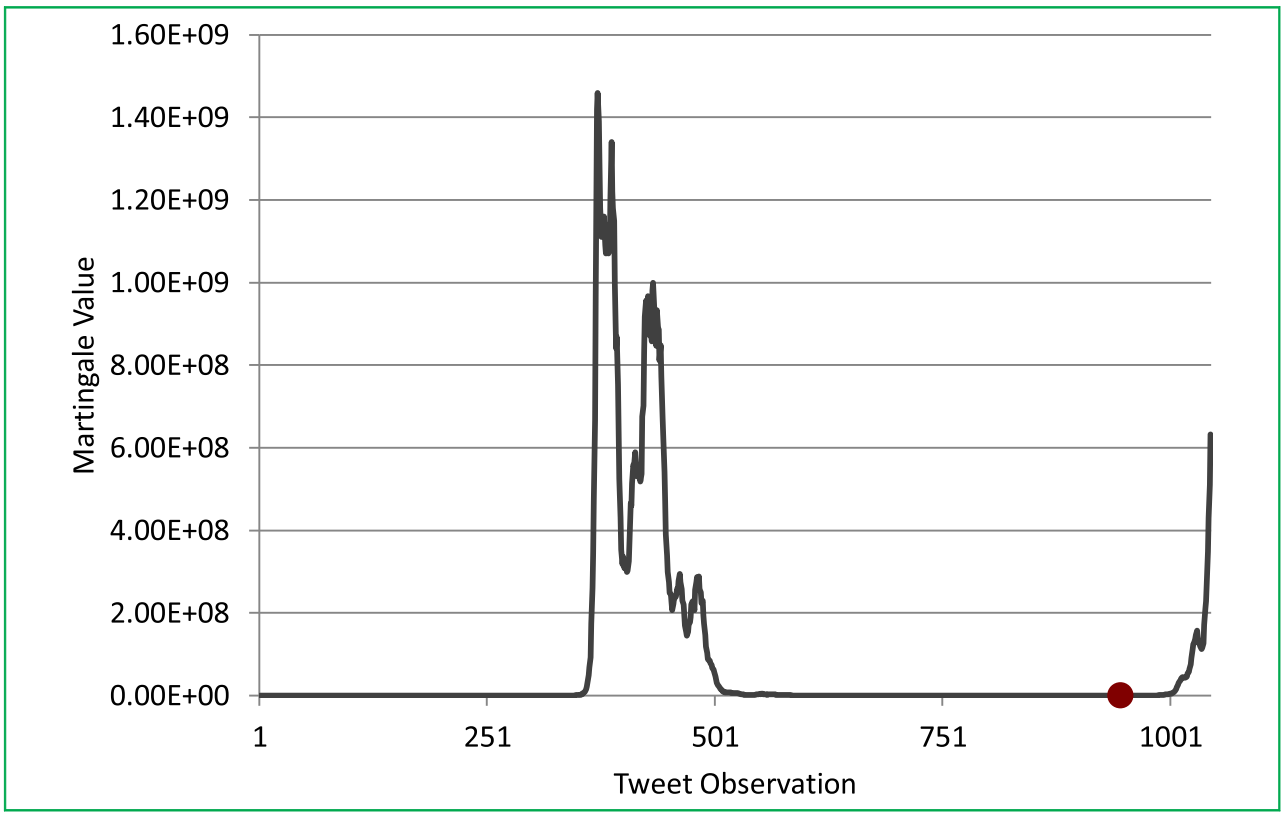

Figure 3

Martingale value distribution over the series of tweets for user *Max*. The values are computed using four dimensions (SPA text score, friends, volume, and retweets). The point represents the true abrupt emotion change point (#946).

and recall for the highest risk class (72% and 69.2%), motivated us to use the J48 and SMO classifiers for our two-step tweet classification model.

Evaluation of the martingale framework for emotion change detection

In this section, we present the results of the martingale framework on the set of longitudinal data. We use two of the Twitter users, Max and Frank, who presented marked change points in at least one behavioral feature. Their names have been changed to protect their privacy.

As a first step, we were interested in the effect of the different behavioral features on the martingale values. We examined the values with one dimension —the *SPA* text score— and four dimensions—the *SPA* text score, friends, volume, and retweets. We focused on the *SPA* text score because its values are continuous. We found that the discrete classes output from the text classifier were not sufficiently smooth and did not reflect the change points in online speech in side experiments.

Figure 2 shows the series of martingale values for user Max, from observation 900 onward, using the *SPA* text score dimension as a single feature. The graph of the last 145 values clearly shows an increase in the values just before the true change point (#946: "Done with my life"). We also notice another change point detected with a high score (#950: "Goodbye. . ."), just after the true change point.

As we add dimensions into the martingale framework for the user Max, the increase in the martingale values after the true change point becomes more apparent. **Figure 3** shows the distributions of the martingale values when they are computed using four dimensions—*SPA* text score, friends, volume, and retweets. We observe an interesting spike around observation (#372). This spike is due to a large number of tweets between Max and another user.

We performed the same analysis for our other validation user Frank. **Figure 4** shows the martingale values over the stream of tweets using one dimension, the *SPA* text score. Contrary to the first validation case, the one-dimensional martingale framework does not appear to increase after the indicated change point. In fact, the two peaks seen around observations 343 and 363 heavily affect the martingale values and they inhibit the framework from detecting a change point. Upon further investigation, we found that these two spikes are linked to negative *SPA* scores (positive

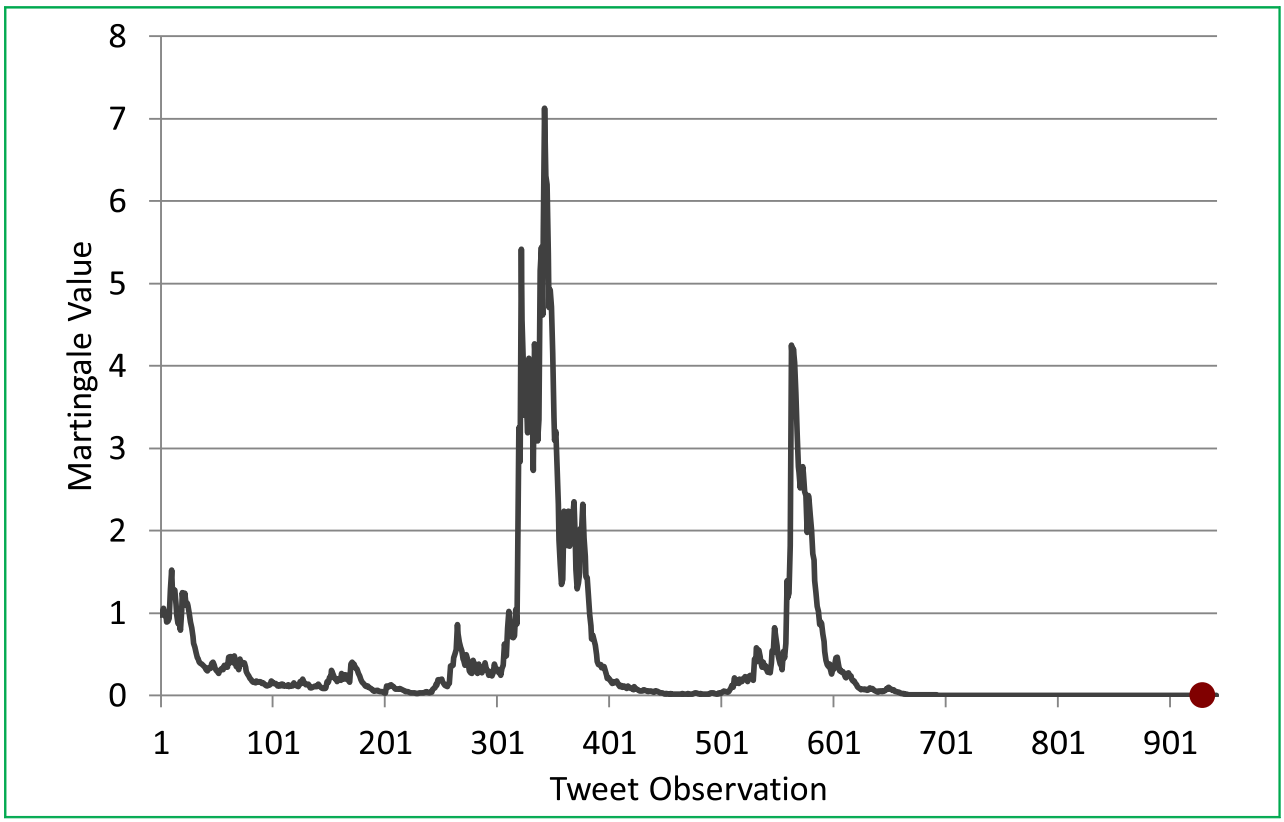

Figure 4

Martingale values distribution over the series of tweets for user *Frank*. The values are computed using only the feature SPA text score (one dimension). The point represents the true abrupt emotion change point (#930).

emotion) corresponding to birthday wishes that Frank received from other users.

These peaks highlight that our martingale framework will need to be adjusted to interpret negative *SPA* text scores as a positive behavioral change that does not require an alarm to be triggered. However, it does indicate that our methodology efficiently detects changes in online speech.

While Figures 2–4 show the evolution of the martingale values, the full martingale framework should detect the true change points. This occurs when the martingale value is greater than λ and the null hypothesis H_0. We ran several experiments to determine the appropriate λ threshold and the testing set size for each validation case. The testing set is the set of observations used to establish the representative value for each dimension.

As previously described, after a change point is detected, the testing set is reset to create new representative values. Our experiments showed that after such a reset, the original λ threshold may no longer be valid for the new martingale values. In other words, the threshold may be too high, and no change points are detected after the testing size is reached. Inversely, the threshold may be too low and multiple false change points are detected.

We present in **Tables 4** and **5** the best results of the emotion-change detection framework obtained for both users Max and Frank, respectively. The fact that the threshold λ does not necessarily remain relevant throughout the stream of tweets hinders the ability of the framework to detect a behavioral change. The true change point is detected for Max with a minimum delay of 14 observations, which is very satisfying. However, for Frank, the true change point is not detected.

Limitations of current study

While the presented methodology and experiments were carefully prepared, we are aware that there are certain limitations. Most notably, we present the results of the full methodology run on only two Twitter users. In order to best summarize the entire methodology from feature generation to change detection in a stream of data, we have decided to present the results of two specific users. One avenue for future research might include further testing on more user timelines. If possible, the testing would be done with a cohort of users identifying as suffering from mental distress.

Furthermore, the parameter setting of the martingale framework could be improved upon. This was one of the

Table 4 Change point detection results, for Max. The bold-font numbers show the best results in terms of true change points detected, number of false alarms, and delay, by tuning the number of dimensions, the parameter λ, and the testing size.

Dimensions	λ	*Testing Size*	*Change Points Detected*	*False Alarms*	*Delay*
1	0.006	200	211, 411, 611, 811, 1011	5	-
1	0.2	50	50, 100, 183, 233, 283, 333, 930, 980	8	-
4	**0.5**	**200**	**200, 400, 600, 960**	**3**	**14**
9	0.1	200	200, 400, 600, 964	3	18

Table 5 Change point detection results, for Frank.

Dimensions	λ	*Testing size*	*Change Points Detected*	*False Alarms*	*Delay*
1	0.5	200	262	1	-
1	0.1	200	201	1	-
5	20	100	100, 200, 300, 627, 727, 827, 927	7	-
9	500	50	61, 135, 535, 607, 657, 707	6	-

major challenges when implementing the framework. Additionally, the combination of behavioral features fed into the martingale framework could be another area of exploration. The features used to determine the mental health of a user may differ between individuals. Given this, it may even be interesting to monitor each behavioral feature individually with the martingale framework.

Conclusion

In this paper, we designed and evaluated a novel approach to monitor the mental health of a user on Twitter. Building off existing research, we worked to translate and quantify suicide warning signs in an online context (user-centric and post-centric behavioral features). In particular, we focused on detecting distress-related and suicide-related content and developed two approaches to score a tweet: an NLP-based approach and a more traditional machine learning text classifier.

To detect changes in emotional well-being, we considered a Twitter user's activity as a stream of observations and applied a martingale framework to detect change points within that stream. Our experiments show that our NLP text-scoring approach successfully separates out tweets exhibiting distress-related content and acts as a powerful input into the martingale framework. While the martingale values "react" to changes in online speech, the change point detection method needs improvement. We were able to detect the true change point for one validation case, but the approach needs to be more robust with respect to parameter setting and positive changes in speech.

For future research, we plan to further explore the impact of martingale parameters on the change detection effectiveness. We also hope to expand the approach to include image processing and other social media outlets in order to assess the effectiveness in other settings. Another interesting perspective is to consider more fine-grained emotion classes such as anger, sadness, fear, etc., instead of considering four levels of distress. However, overall, we believe our initial work presents an innovative approach to detecting suicide-related content in a text stream setting.

References

1. "Preventing suicide: A global imperative," World Health Organization, Geneva, Switzerland, 2014. [Online]. Available: http://www.who.int/mental_health/suicide-prevention/world_report_2014/en
2. "Mental health action plan 2013–2020," World Health Organization, Geneva, Switzerland, 2013. [Online]. Available: http://www.who.int/mental_health/publications/action_plan/en
3. American Foundation for Suicide Prevention (AFSP). [Online]. Available: https://afsp.org
4. P. Corrigan, "How stigma interferes with mental health care," *Amer. Psychologist*, vol. 59, no. 7, pp. 614–525, 2004.
5. D. Rickwood, F. P. Deane, C. J. Wilson, et al., "Young people's help seeking for mental health problems," *Aust. e-J. Adv. Mental Health*, vol. 4, no. 3, pp. 218–251, 2005.
6. M. De Choudhury, M. Gamon, S. Counts, et al., "Predicting depression via social media," in *Proc. 7th Int. AAAI Conf. Weblogs Social Media*, Boston, MA, USA, 2013, pp. 128–137.
7. M. Moreno, L. Jelenchick, K. Egan, et al., "Feeling bad on Facebook: depression disclosures by college students on a social networking site," *Depression Anxiety*, vol. 28, no. 6, pp. 447–455, 2011.
8. J. F. Gunn and D. Lester, "Twitter postings and suicide: An analysis of the postings of a fatal suicide in the 24 hours prior to death," *Suicidologi*, vol. 17, no. 3, pp. 28–30, 2012.
9. V. Kailasam and E. Samuels, "Can social media help mental health practitioners prevent suicides?" *Current Psychiatry*, vol. 14, no. 2, pp. 37–51, 2015.
10. M. De Choudhury, E. Kiciman, M. Dredze, et al., "Discovering shifts to suicidal ideation from mental health content in social media," in *Proc. 2016 CHI Conf. Human Factors Comput. Syst.*, San Jose, CA, USA, 2016, pp. 2098–2110.
11. M. De Choudhury, S. Counts, and E. Horvitz, "Predicting postpartum changes in emotion and behavior via social media," in *Proc. SIGCHI Conf. Human Factors Comput. Syst.*, New York, NY, USA, 2013, pp. 3267–3276.
12. M. T. Lehrman, C. O. Alm, and R. A. Proano, "Detecting distressed and non-distressed affect states in short forum texts," in *Proc. 2012 Workshop Lang. Social Media*, Montreal, QC, Canada, 2012, pp. 9–18.
13. American Psychiatric Association, "Practice guideline for the assessment and treatment of patients with suicidal behaviors," *Amer. J. Psychiatry*, vol. 160, no. 11, pp. 1–60, 2003.

14. M. D. Rudd, A. L. Berman, T. E. Joiner, et al., "Warning signs for suicide: Theory, research, and clinical applications," *Suicide Life-Threatening Behav.*, vol. 36, no. 3, pp. 255–262, 2006.
15. J. Jashinsky, S. H. Burton, C. L. Hanson, et al., "Tracking suicide risk factors through Twitter in the US," *J. Crisis Intervention Suicide Prevention*, vol. 35, no. 1, pp. 51–59, 2014.
16. S. S. Rude, E. M. Gortner, and J. W. Pennebaker, "Language use of depressed and depression-vulnerable college students," *Cogn. Emotion*, vol. 18, no. 8, pp. 1121–1133, 2004.
17. G. Coppersmith, et al.,"Exploratory analysis of social media prior to a suicide attempt," in *Proc. 3rd Workshop Comput. Linguistics Clinical Psychol.*, 2016, pp. 106–117.
18. B. O'Dea, S. Wan, P. J. Batterham, et al., "Detecting suicidality on twitter," *Internet Interventions*, vol. 2, no. 2, pp. 183–188, 2015.
19. S. R. Braithwaite, C. Giraud-Carrier, J. West, et al., "Validating machine learning algorithms for twitter data against established measures of suicidality," *JMIR Mental Health*, vol. 3, no. 2, 2015, Art. no. e21.
20. J. Robinson, G. Cox, E. Bailey, et al., "Social media and suicide prevention: A systematic review," *Early Intervention Psychiatry*, vol. 10, no. 2, pp. 103–121, 2016.
21. R. P. Adams and D. J. MacKay, "Bayesian online change point detection," Tech. Rep., 2007. [Online]. Available: https://arxiv.org/abs/0710.3742
22. R. Garnett, M. A. Osborne, S. Reece, et al., "Sequential Bayesian prediction in the presence of change points and faults," *Comput. J.*, vol. 53, no. 9, pp. 1430–1446, 2010.
23. M. Basseville and I. V. Nikiforov, *Detection of Abrupt Changes: Theory and Application*. Upper Saddle River, NJ, USA: Prentice-Hall, 1993.
24. S. Liu, M. Yamada, N. Collier, et al., "Change-point detection in time-series data by relative density-ratio estimation," *Neural Netw.*, vol. 43, pp. 72–83, 2013.
25. V. Vovk, I. Nouretdinov, and A. Gammerman, "Testing exchangeability on-line," in *Proc. 20th Int. Conf. Mach. Learn.*, Washington, DC, USA, 2003, pp. 768–775.
26. S. Ho and H. Wechsler, "A martingale framework for detecting changes in data streams by testing exchangeability," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 32, no. 12, pp. 2113–2127, Dec. 2012.
27. S. Babakhani, N. Mozaffari, and A. Hamzeh, "A martingale approach to detect peak of news in social network," Tech. Rep., 2014. [Online]. Available: https://arxiv.org/abs/1409.2002
28. M. De Choudhury, S. Counts, and E. Horvitz, "Social media as a measurement tool of depression in populations," in *Proc. 5th Annu. ACM Web Sci. Conf.*, New York, NY, USA, 2013, pp. 47–56.
29. K. M. Harris, J. P. McLean, and J. Sheffield, "Suicidal and online: How do online behaviors inform us of this high-risk population?" *Death Studies*, vol. 38, no. 6, pp 387–394, 2014.
30. G. Coppersmith, M. Dredze, and C. Harman, "Quantifying mental health signals in Twitter," in *Proc. Workshop Comput. Linguistics Clin. Psychol., Linguistic Signal Clin. Reality*, Baltimore, MD, USA, 2014, pp. 51–60.
31. C. Karmen, R. C. Hsiung, and T. Wetter, "Screening internet forum participants for depression symptoms by assembling and enhancing multiple NLP methods," *Comput. Methods Programs Biomed.*, vol. 120, no. 1, pp. 27–36, 2015.
32. R. Krestel and S. Siersdorfer, "Generating contextualized sentiment lexica based on latent topics and user ratings," in *Proc. 24th ACM Conf. Hypertext Social Media*, Paris, France, 2013, pp. 129–138.
33. American Psychiatric Association, *Diagnostic and Statistical Manual of Mental Disorders (DSM–5)*. Arlington, VA, USA: Amer. Psychiatric, 2013.
34. F. Benamara, C. Cesarano, A. Picariello, et al., "Sentiment analysis: Adjectives and adverbs are better than adjectives alone," in *Proc. Int. Conf. Weblogs Social Media*, Boulder, CO, USA, 2007, pp. 203–206.
35. C. Homan, R. Johar, T. Liu, et al., "Toward macro-insights for suicide prevention: Analyzing fine-grained distress at scale," in *Proc. Workshop Comput. Linguistics Clin. Psychol., Linguistic Signal Clin. Reality*, Baltimore, MD, USA, 2014, pp. 107–117.
36. G. B. Gil, A. B. de Jesus, and J. M. M. Lopez, "Combining machine learning techniques and natural language processing to infer emotions using Spanish Twitter corpus," in *Proc. Highlights Practical Appl. Agents Multi-Agent Syst.*, Salamanca, Spain, 2013, pp. 149–157.
37. A. Pak and P. Paroubek, "Twitter as a corpus for sentiment analysis and opinion mining," in *Proc. 7th Int. Conf. Lang. Resources Eval.*, Valletta, Malta, 2010, pp. 1320–1326.
38. N. Mozafari, S. Hashemi, and A. Hamzeh, "A precise statistical approach for concept change detection in unlabeled data streams," *Comput. Math. Appl.*, vol. 62, no. 4, pp. 1655–1669, 2011.
39. A. J. Viera and J. M. Garrett, "Understanding inter observer agreement: the Kappa statistic," *Family Med.*, vol. 37, no. 5, pp. 360–363, 2005.
40. A. Abboute, Y. Boudjeriou, G. Entringer, et al., "Mining twitter for suicide prevention," in *Proc. 19th Int. Conf. Appl. Natural Lang. Inf. Syst.*, Montpellier, France, 2014, pp. 250–253.

Received February 25, 2017; accepted for publication March 26, 2017

Mia Johnson Vioulès *AXA Global Direct, 92150 Suresnes, France (miakay.johnson@gmail.com).* Ms. Johnson Vioulès currently works in the insurance industry for AXA as a Data Scientist. She earned master's degrees in applied mathematics and business intelligence from Sorbonne-Panthéon Paris 1 and École Centrale Paris, where she completed a research internship with the Montpellier Laboratory of Informatics, Robotics, and Microelectronics. Her research focused on sentiment analysis and data stream mining for application to suicide detection in social media.

Bilel Moulahi *LIRMM UM CNRS, UMR 5506, 34095 Montpellier, France (bilel.moulahi@lirmm.fr).* Dr. Moulahi is a Postdoctoral Researcher in the Montpellier Laboratory of Informatics, Robotics and Microelectronics. He earned his Ph.D. degree in computer science from the University of Toulouse, France, where he worked on information retrieval and multi-criteria relevance ranking. His current research primarily focuses on suicide detection and prevention in social media.

Jérôme Azé *LIRMM UM CNRS, UMR 5506, 34095 Montpellier, France (jerome.aze@lirmm.fr).* Dr. Azé is a Professor in the Montpellier Laboratory of Informatics, Robotics, and Microelectronics. Previously, he was Assistant Professor and member of the bioinformatic team at the University of Paris-Sud. Currently, he is the head of the Multimedia and Internet Department at the University of Technology, Béziers, France. His research mainly includes data mining and its applications to e-health, sentiment analysis, and suicide detection in social networks.

Sandra Bringay *LIRMM UM CNRS, UMR 5506, 34095 Montpellier, France (sandra.bringay@lirmm.fr).* Dr. Bringay is a Professor in the Montpellier Laboratory of Informatics, Robotics, and Microelectronics and is a member of the toxico-vigilance working group managed by the National Institute of Health Monitoring. She is leading several projects on biomedical research and medical information. Her current research primarily focuses on health data mining, sequential pattern mining, and sentiment analysis.