DARE to Care: A Context-Aware Framework to Track Suicidal Ideation on Social Media

Bilel Moulahi, Jérôme Azé, Sandra Bringay

DARE to Care: A Context-Aware Framework to Track Suicidal Ideation on Social Media

Bilel Moulahi[1], Jérôme Azé[1], and Sandra Bringay[1,2]

[1] LIRMM, Université de Montpellier, CNRS, France,
bilel.moulahi@lirmm.fr
[2] AMIS, Université Paul Valéry Montpellier, France

Abstract. The abundance and growing usage of social media has given an unprecedented access to users' social accounts for studying people's thoughts and sentiments. In this work, we are interested in tracking individual's emotional states and more specifically suicidal ideation in microblogging services. We propose a probabilistic framework that models user's online activities as a sequence of psychological states over time and predicts the emotional states by incorporating the context history. Based on Conditional Random Fields, our model is able to provide comprehensive interpretations of the relationship between the risk factors and psychological states. We evaluated our approach within real case studies of Twitter' users that have demonstrated a serious change in their emotional states and online behaviour. Our experiments show that the model is able to identify suicidal ideation with high precision and good recall with substantial improvements on state-of-the-art methods.

Keywords: Social Media, Suicide, Emotional states, CRFs, Context

1 Introduction

In the last decade, social media platforms has been increasingly used by mental health professionals and clinicians with myriad purposes ranging from the detection and diagnosis of major depressive disorders to predicting flu epidemics and symptoms spread [4,5,8]. Recent studies have shown that people are more likely to seek support from informal resources in social media, rather than seeking formal treatment from professionals. In order to better understand the information being shared, a developing body of literature have examined references to depression and suicidal ideation in open social network such as Twitter and Facebook [2]. The current state of the art mainly focuses on assigning a polarity (positive, negative, neutral) describing the emotions conveyed by users. As such, most approaches are based on keyword matching mechanisms and static analysis of documents that ignore the whole user' behaviour. However, from a psychological perspective, suicidal ideation are perceived as a continuum of sequence of events and different mental states that may lead or not to real suicide. Given the stream nature of individual's online content in social media, this representation may be exploited to better study the mind of suicidal users over time by taking into account the whole context and online activity. To tackle this challenge,

we leverage the Conditional Random Fields to track suicidal ideation in social media. We enhance our model using risk factors derived from the psychological literature and features based on the document content as well as previous shared posts. We evaluated our approach on a manually annotated corpus of tweets, published by users that demonstrated serious signs of suicidal ideation. The collection of users is validated by a professional with expertise in mental health research to retain only users with real symptoms. Experimental results show that the predicted sequences of emotional states achieve good results when compared to conventional approaches.

In the remainder of the paper, we first summarize related work, then describe our framework for suicidal ideation detection. We present and discuss the results, then conclude with future work.

2 Related Work

According to the Centers for Disease Control and Prevention [7], more than 40,000 suicides were reported in the United States in 2012. Suicide is now the $10th$ leading cause of death in the country. With the advent of social media, at-risk individuals are using the Internet to post suicidal communications on emerging services such as Facebook, Twitter and Reddit. Recent work demonstrated that evaluating suicidal risk factors in social networks can be used to prevent suicide and detect suicidal ideation in its early stages [4,5,8]. Gunn and Lester [4] analysed the Twitter posts of a person who had recently died through suicide. They studied the posts sent in the twenty-four hours prior to her death, finding an increase in positive emotions and a change in focus from the self to others as the time of death approached. The authors used the Linguistic Inquiry and Word Count (LIWC) software[3] to identify parts of speech, emotional words and cognitive processes among other concepts. LIWC was also used in [5] as a sampling technique to identify "sad" Twitter posts that were subsequently classified using a machine learning into levels of distress on an ordinal scale, with around 64% accuracy in the best-case. In the same line of research, Sueki [8] used an online panel of young (early 20s) Twitter user to examine the association between suicide-related tweets and suicidal behaviour. The authors investigated the linguistic features of suicidal ideation and identified the most important markers of future suicide. For example, phrases such as "*want to commit suicide*" were found to be strongly associated with lifetime suicide attempts, while phrases that suggest suicidal intent, such as "want to die", were found to be less strongly associated. Despite the solid foundation, the current literature is missing potential key factors in the effort to track suicide related symptoms and ideation. Currently, few works analysed the evolution of an individual's online behaviour. Rather, the analysis is static and may take into consideration one post at a time while ignoring the whole context and the sequential nature of data streams.

[3] `https://liwc.wpengine.com`

3 A Context-Aware Framework to Track Suicidal Ideation on Social Media

In this section, we formulate the problem and describe our DARE[4] framework to suicidal ideation tracking, and then introduce the feature extraction process.

3.1 Problem Description

Let $P = < p_1, p_2, ..., p_N >$ be a continuous stream of user' posts that arrive in chronological order in a given time window W. The problem consists in predicting an output vector $Y = < y_0, y_1, ..., y_N >$ of emotional states given the observed sequence of posts P. The input of observations P represent the feature vectors, where each observation p_j contains various information about the post at time t_j, and each variable y_j is an emotional state inferred from p_j.

A natural way to represent the sequential emotional state changes is provided by the Markov network models. An interesting generalization of this framework is given by graphical models such as Conditional Random Fields (CRF) [6].

3.2 Conditional Random Fields Model

CRF is a type of undirected probabilistic graphical model that has been successfully applied in many text processing and computer vision problems [9]. One of the main strengths of this conditional model lies in its ability to encompass complex dependencies between the observations, in addition to the comprehensive interpretations of the relationship between the features it provides. In our context, this property is very important given that a transition from an emotional state to another is heavily dependent on the previous observed emotions.

Given a sequence of posts $P = < p_1, p_2, ..., p_N >$ and a sequence of hidden emotions $Y = < y_0, y_1, ..., y_N >$, CRF models the conditional probability as follows:

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

Where Z is a normalization factor (also called the partition function) to make $p(Y|P)$ a valid probability over all label sequences. Z is defined as the sum of exponential number of sequences:

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

The scalar w_k is the weight of feature f_i and w_k's are the parameters of the CRF model, and are learned by numerical optimizations techniques such as gradient based approaches. The feature functions $f_k(y_{i-1}, y_i, P, N)$ look at a

[4] DARE stands for conDitionAl Random fiElds

pair of adjacent (emotional) states y_{i-1}, y_i, the whole sequence of posts P, and the current position of the sequence i. Note that the use of CRFs allows us to define a large number of dependent or independent features without worrying about the complex statistical relationship between these features. The use of each feature depends on the weight w_k which acts as an activation factor of the feature. In our setting, we consider that a set of posts are included in the same sequence if they are published by the same user within the same time window $\mathcal{T}$.
For clarity, the online behavioural activity of a user is partitioned into sequences namely *sessions*. Each session $S_{\mathcal{T}}$ can be thought of as a sequence of observations (posts) spanning the same time period $\mathcal{T}$. During a single session, we assume that the behavioural activity of the user may be modelled by a sequence of emotional states inferred from the posts published during $\mathcal{T}$. The time interval between two sessions of activity is set to a boundary threshold θ to demarcate the users mental states at different granularity. This threshold parameter is typically set to one *day hours* (i.e., the time a user may potentially start with a new day). In the experiments, we tuned this parameter and also set it to $7h$ and $12h$.

3.3 Feature Extraction

Suicidal ideation expressed by users on social media depends upon the context of the posts. In order to train and test our model, we used the text of posts as the main marker for mental health. After cleaning the posts (ie., lower case, strip punctuation, remove special characters, mentions and URLs, etc), we derive three sets of features.

The first set of features includes *lexical characteristics* of the text. We use Parts of Speech (POS) to capture reference to self (first personal pronouns "I", "my", "me"), nouns, verbs and adverbs. In fact, literature in the field of sociolinguistics has shown that the use of first personal singular or plural in social media posts may reveal mental well-being [3]. Examples of Twitter posts may include: "*I just wanna be left alone, I'm at the end of my rope I don't know what to do at this point*" , etc. The second set of features includes *Psychological and emotional lexicon* features. This set refers to linguistic themes and terms that are commonly used by at-risk individuals in social media. Examples include reference to negative emotions, depression, self-harm, sadness, mental health and suicide[5] [3]. We enrich the lexicons by including other terms that refer to swear words. These features have been shown to carry important information in the context of sentiment analysis [3]. The last set of features are *Contextual features*, which are related to the posts observed during previous sessions. For a given post published at time t, we use information about the characteristics of posts observed at time $t-1$, $t-2$ and even $t+1$ ($t+1$ informations are only available in the training phase). For instance, we check whether we observed symptoms terms in previous posts at the same user session.

[5] `https://github.com/sbma44/begin_anew`

4 Experiments

4.1 Data Preparation and Evaluation Setting

Due to the absence of publicly available datasets for the evaluation of suicide detection methods in social media, we used the Twitter streaming API[6] to collect tweets containing references to themes such as self-mutilation and suicide. We exploit a list of key phrases generated from the American Psychological Association[7] (APA) list of suicide risk factors and keywords from the American Association of Suicidology[8] (AAS) list of warning signs. Within all the collected tweets, we only considered users that demonstrated serious suicide symptoms in their online behaviour. To ascertain the source of the tweets, the Twitter accounts are validated by a professional with expertise in mental health research, which resulted in 60 users. As an attempt to avoid over-fitting, we also included 60 Twitter accounts of normal users with the same keywords. Table 1 shows the statistics of dataset used in the experiments. We also show in Figure 1 the temporal distribution of the time creation of the tweets over the hours of the day.

Users	120
Tweets	29887
Sessions	8421
Avg. of #tweets per session	5
Max of #tweets per session	200
Min of #tweets per session	1

Table 1: Summary of statistics of data collected form Twitter. A session is a sequence of observations.

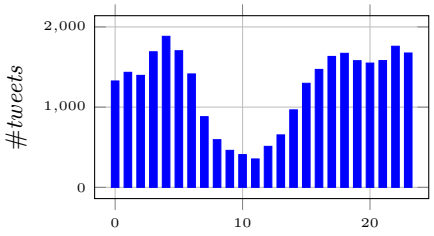

Fig. 1: Temporal distribution of all tweets over a 24-hour day (from midnight to 23H). Most of the tweets are published late at night.

The most challenging task of an experimental evaluation is the data labelling. To solve this problem, eight researchers and a mental health professional manually annotated and assessed a random subset of tweets to determine the class of tweets. We define three levels of mental states:: (i) *No distress*: in which the post discusses everyday occurrences such as work, going out, weekend activities, etc. (ii) *Minimal/Moderate distress*: which refers to post expressing distress and that could be considered common for most individuals (*i.e.*, exam, presentation for work, etc.), and (iii) *Severe distress*: that refers to posts that include mentions of self-harm, suicidal thoughts, apologies, feelings of worthlessness, self-hate, guilt, etc.

[6] `https://dev.twitter.com/streaming/overview`
[7] `http://www.apa.org/topics/suicide/`
[8] `http://www.suicidology.org`

Each tweet was reviewed by at least two annotators. The annotated data had a Cohen's kappa statistic of 69.1%, which is considered as a substantial agreement among the annotators. The weighted kappa statistic, which takes into account different levels of disagreement, was around 71.5%.

4.2 Results and Discussion

We present our results in two phases. We start by exploring the emotional states dynamics on Twitter, then we present and discuss the results of the predictions given by our Framework.

Analysis of the Emotional States Change In this section, we explore the users behavioural changes over time. We analyse the shifts between the emotional states. Table 2 shows the transitions between the 3 emotional states that we considered in our setting.

	No distress	**Minimal Distress**	**Severe Distress**
No distress	**0.54**	**0.37**	-1.36
Minimal Distress	0.16	0.01	-0.21
Severe Distress	-1.60	-0.0001	**1.41**

Table 2: Users emotion changes according to the three mental states. Bold values correspond to the most likely transitions. Each row may be interpreted by the transition from a state to another.

It is interesting to note from Table 2, that the most likely transitions between two different states led from the *No distress* to the *Minimal distress* state, with a lower probability value for the opposite transition. This is natural given that users shifting to a more risky emotional state are unlikely to return to the normal state. Interestingly enough, we also observe that users on the *No distress* and *Severe distress* states tend to remain in the same state with values of 0.54 and 1.41, respectively. The probability obtained for the *Severe distress* state is unexceptional and may be interpreted by the fact that this mental state is generally reached when the individual is focusing on suicide and usually ends by a suicide attempt [1], which also explains the very low probability value for the transition from *Severe distress* to *Minimal distress*.

Evaluation of the DARE Framework To examine the effectiveness of the proposed approach, we perform a comparative evaluation against a set of representative machine learning methods. We also compare our model against a configuration of CRFs that do not consider the sequences of observations (*i.e.*, each post is considered as a session). All the models are trained and tested using a 5-fold cross-validation approach. In the training phase, the models are built

using the default WEKA[9] parameter settings. To train our CRF, we exploit the Gradient descent using the L-BFGS method. The coefficients for L1 and L2 regularization are set by defaults to 0.1 and 0.1, respectively. The threshold parameter θ representing the time interval between two sessions of activity is set to 12 *hours* (7 and 24 hours led to the same results). We perform all experiments within a fair setting, using the same training and test data with a 5-fold cross validation.

Algorithm	**Average Precision**	**Avg. Recall**	**Avg. F1-score**
SVM	0.446	0.227	0.301
Naive Bayes	0.538	0.127	0.205
J48	0.451	0.127	0.198
Random Forest	0.500	0.127	0.202
Multilayer Perceptron	0.430	0.281	0.340
Bagging	0.400	0.290	0.336
Stacking	0.285	0.109	0.157
Vote	0.583	0.127	0.208
DARE without sequences	0.666	0.549	0.479
DARE approach	0.816	0.752	0.711

Table 3: Evaluation results of the DARE framework.

Table 3 reports the results by means of average Precision, Recall and F1-score measures. We start by noting that both configurations of our CRFs frameworks outperform the tested baselines in terms of F1-score. These gains are more expressive for the DARE approach, which confirm the benefits of modelling the user's activity as a sequence of dependent observations (context), which in turn, are modelled as a sequence of emotional states over time. The results achieved by the baselines are rather similar with a slight advantage of SVM, Neural Network and the Bagging algorithm. The low values may be related to the training phase in which we didn't considered feature selection, which may be considered as a disadvantage in text classification tasks. In addition, we considered three emotional states instead of two as used in most of sentiment analysis tasks. The analysis of results obtained by DARE for the individual emotional states are found to be better for the *No distress* (0.828) and *Severe distress* states (0.711) according to the F1-score. In fact, we found that the high rate of false negatives for *Minimal distress* state may be explained by the fact that: (i) the latter stands in two emotional states; and (ii) the self-transition for this state is very low (0.01) compared to the ingoing and outgoing transitions (see Table 2), especially from/to the *No distress* state.

[9] `www.cs.waikato.ac.nz/ml/weka/`

5 Conclusion and Future Work

In this paper, we proposed an approach for suicidal ideation tracking based on the Conditional Random Fields framework. Our model outshines the traditional machine learning methods in input features, flexibility and extensibility to other settings involving data streams.

This study has some limitations that can be explored in future work. We plan to further consider a more fine-grained classification to include emotional states such as anger, sadness, fear, etc., instead of considering only three level of distress states. The impact of the models parameters can also be investigated, eg., long-term user behaviour (θ parameter) and size of data. Another interesting perspective is to personalize the predictions by generating distinct users behaviour models based on their online activities. Additionally, we would also like to test the impact of sentiment-specific word embedding on our model.

References

1. Abby Adler, Ashley Bush, Frances K. Barg, Guy Weissinger, Aaron T. Beck, and Gregory K. Brown. A mixed methods approach to identify cognitive warning signs for suicide attempts. *Archives of Suicide Research*, 20(4):528–538, 2016.
2. Pete Burnap, Walter Colombo, and Jonathan Scourfield. Machine classification and analysis of suicide-related communication on twitter. In *Proceedings of the 26th ACM Conference on Hypertext & Social Media*, HT '15, pages 75–84, New York, NY, USA, 2015. ACM.
3. Munmun De Choudhury, Scott Counts, and Eric Horvitz. Predicting postpartum changes in emotion and behavior via social media. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI'13, pages 3267–3276, New York, NY, USA, 2013. ACM.
4. John F. Gunn and David Lester. Twitter postings and suicide: An analysis of the postings of a fatal suicide in the 24 hours prior to death. *Suicidologi*, 17(3):28–30, 2012.
5. Christopher Homan, Ravdeep Johar, Tong Liu, Megan Lytle, Vincent Silenzio, and Cecilia Ovesdotter Alm. Toward macro-insights for suicide prevention: Analyzing fine-grained distress at scale. In *Proceedings of the Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality*, ACL 2014, pages 107–117, Baltimore, MD, USA, 2014.
6. John D. Lafferty, Andrew McCallum, and Fernando C. N. Pereira. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Proceedings of the Eighteenth International Conference on Machine Learning*, ICML '01, pages 282–289, San Francisco, CA, USA, 2001.
7. Murphy SL, Kochanek KD, Xu J, and Arias E. Mortality in the united states. *NCHS Data Brief.*, (229):1–8, 2014.
8. Hajime Sueki. The association of suicide-related twitter use with suicidal behaviour: A cross-sectional study of young internet users in japan. *Journal of Affective Disorders*, 170:155 – 160, 2015.
9. Charles Sutton and Andrew McCallum. An introduction to conditional random fields. *Foundations and Trends in Machine Learning*, 4(4):267–373, 2012.