tPredictor: A Micro-blog Based System for Teenagers' Stress Prediction

Jing Huang, Qi Li, Zhuonan Feng, Yiping Li, Ling Feng
Dept. of Computer Science and Technology
Centre for Computational Mental Healthcare Research, Institute of Data Science
Tsinghua University, Beijing 100084, China
{j-huang14,liqi13}@mails.tsinghua.edu.cn, fzn0302@163.com,
fengling@tsinghua.edu.cn

ABSTRACT

Too much stress is easy to do harm to the physical and psychological health of teenagers, because most teenagers neither have the ability to cope with the stress by themselves nor like seeking adults' help initiatively. Social media has demonstrated its feasibility in detecting teenagers' stress with the micro-blog having become a popular channel for teenagers' self-expression. In this demonstration, we present a system called *tPredictor*, which can predict teenagers' future stress based on the social media micro-blog. Two questions are to be resolved: (1) what will the stress level of the teenager(s) be in the next time unit? (2) how will the stress level of the teenager(s) change (increase, decrease, remain unchanged) in the next time unit? *tPredictor* tackles the above prediction questions, taking into account the influence of future stressful events on teenagers' emotion. Considering the similarity of stock price movement and the stress level change, we define the stress candlestick charts and the stress reversal signals for stress change trend prediction. *tPredictor* can predict the stress of both individuals and a group of teenagers, which provides a platform for teenagers themselves, their parents or some institutions such as schools to know teenagers' future stress for taking measures timely to prevent the serious consequences.

Keywords

Micro-blog, teenager, psychological pressure, prediction

1. INTRODUCTION

Nowadays, teenagers are inevitably suffering much stress from various aspects. A survey, conducted by the American Psychological Association in August 2013, showed that the teenagers were the most stressed-out age group in the U.S [1]. Due to the unsoundness of teenagers' psychological regulation mechanism, too much stress contributes to teenagers' bad consequences more easily such as sleeplessness, depression and even suicide. However, for most teenagers, they lack experience and can't realize the seriousness of their stress so that they seldom actively seek for adults' help. And for guardians, parents can't be around and focus on teenagers every minute and even abundant teachers are not sufficient to attend to each student. Therefore, it is very useful to find methods for teenagers to understand their own stress, for parents to timely know teenagers' stress situation when teenagers don't initiatively reveal their stress to them and for teachers to be able to get hold of all the students' stress status. Nowadays, micro-blog has become a major channel for teenagers' self-expression. Teenagers post so many tweets expressing their personal emotions everyday, which provides abundant available data to detect teenagers' stress and further predict their stress change. In the literature, many researchers have studied using micro-blog to analyze people's mental health. [8] found the difference between depressed and non-depressed people through analyzing their tweeting contents and behaviors . [9] proposed a depression detection model based on the sentiment analysis of the micro-blog. [10, 5] provided a machine learning method to detect teenagers' stress of study, affection, inter-personal and self-cognition. However, all the studies aimed to detect the existing psychological problems where the harm has done in fact. To prevent bad consequences, [3] investigated people's social media behaviors and built a statistical classifier to predict the depression. Our previous work focused on the teenagers group and predicted their future stress using a stress level time series detected from micro-blog. It integrated the stressful event to predict future stress level and defined the stress candlestick chart to predict future stress general change trend [7, 6].

In this demonstration, we present a system called *tPredictor* to predict teenagers' future stress value and change trend. It is a system implementation of our previous work [7, 6] which can visualize the prediction results and enable users to easily understand the stress prediction results. For the functionalities, we further extend the system to the group stress prediction. It aggregates all the prediction results of teenagers and presents some statistical results of the group stress prediction. It also picks up the teenagers needing help based on out predicted future stress status from a group of teenagers. *tPredictor* provides a straightforward way for users to obtain the most important information, as well as get hold of the overall stress situation and meanwhile know well each teenager's stress status. Therefore, our system can be applied to both institutional users, such as schools and

Series ISSN: 2367-2005

10.5441/002/edbt.2016.61

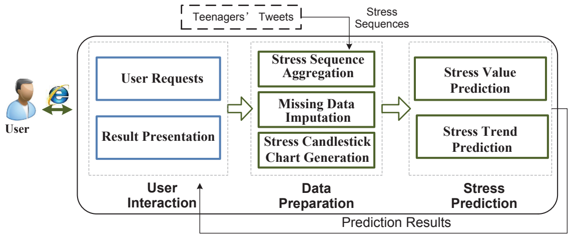
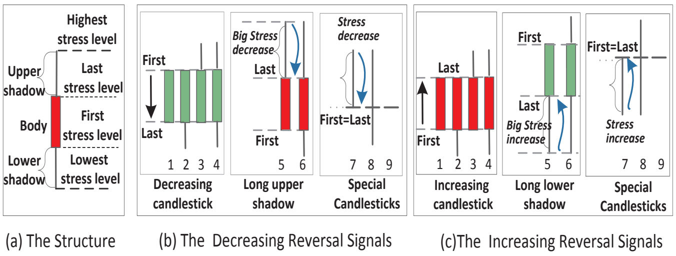


Figure 1: System Architecture



(a) The Structure (b) The Decreasing Reversal Signals (c)The Increasing Reversal Signals

Figure 2: The Stress Candlestick Charts

individual users such as teenagers themselves and parents.

2. SYSTEM ACHITECTURE

Fig.1 shows the system framework including the *User Interaction*, the *Data preparation* and the *Stress Prediction*.

2.1 User Interaction

User Requests receives users' requests including adding new teenagers to their concerning list, choosing teenagers to be predicted, and setting prediction parameters like the time granularity and the event information.

Result Presentation aims to visualize the prediction results for users' easy understanding. It analyzes the group prediction results and picks up the most important information to users. In details, it aggregates teenagers by different predicted stress values and change trends, and presents their distributions through lucid charts, which helps users get hold of the overall stress situation. Besides, it demonstrates each teenager's predicted stress together, which enables users to do the comparative analysis and quickly find the teenagers whose future stress will be severe.

2.2 Data Preparation

The *Data preparation* preprocesses the original stress sequences which detected from teenagers' micro-blog [10].

Stress Sequence Aggregation aggregates the tweets' stress sequence with different time granularity (day, month, etc.) to obtain six stress-related indexes. The stress sequence is the $Stress(T)$: $(t_1, l_1), (t_2, l_2)...(t_n, l_n)$, where t_i represents the time of i_{th} tweet, and the l_i represents the stress level of the i_{th} tweet. The six stress-related indexes include the max stress level L_{max}, the min stress level L_{min}, the average stress level L_{avg}, the sum stress level L_{sum}, the number of stress tweets L_{scount}, the total number of tweets L_{count}, and the proportion of stress tweets $L_{proportion}$ in the aggregated time interval.

Missing Data Imputation aims to solve the missing data problem which is because of the casualness of users' tweeting time and frequency. We apply the Gaussian Process Regression to do the data imputation. The adjacent data, including the stress value before and after the missing data, will be used to inference the missing value. The $(\Delta t, L)$, where Δt and L denote the time distance and the stress level of the adjacent time unit, is used as the input feature vector form. We use non-missing data as the training data and then get the estimated value of the missing data through the trained model. Our experiments showed that the Gaussian Process Regression performed better than other imputation methods including the nearest mean approach, the linear interpolation and exponential smoothing.

Stress Candlestick Chart Generation generates the stress candlestick chart SC and its corresponding feature vector SCF. The stress candlestick derives from the candlestick chart of stock price due to some similarities between stock price movement and stress level change. For instance, the stock price is influenced by the game between sellers and buyers and the related events of companies while the stress level is influenced by the personal self-regulation mechanism and the stressful events. The stress candlestick chart SC is defined as $(L_{first}, L_{last}, L_{high}, L_{low})$, namely the first, last, highest and lowest stress level in the time unit respectively. The stress candlestick chart feature vector SCF is defined as a five tuples $(shape, bodylen, upperlen, low-erlen, changeslope)$ which represents the shape, body length, uppershadow length, lowershadow length of the SC as Fig. 2 (a) shows and the stress change rate between two SCs.

2.3 Stress Prediction

2.3.1 Stress Value Prediction

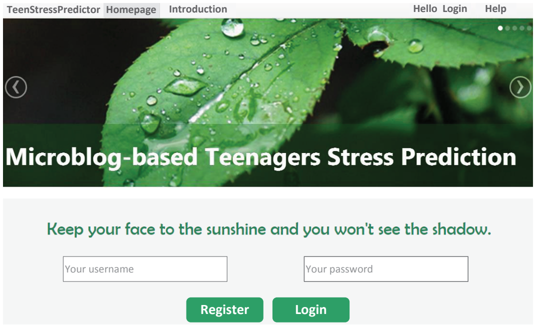
Stress Value Prediction aims to predict the max, min and average stress level of the next time unit. The three stress values can comprehensively represent the teenagers' stress status and are also the major concerns to their followers.

Feature-Aware Time Series Prediction. According to Granger Causality analysis, the L_{max}, L_{min} and L_{avg} are related to not only their past values but also the other stress-related indexes such as the L_{sum}, L_{scount}, L_{count} and $L_{proportion}$. Based on the confidence of 95%, we find that the L_{max} is correlated to L_{sum}, L_{min} is correlated to ($L_{proportion}$, L_{count}), and L_{avg} is correlated to (L_{sum}, $L_{proportion}$). We apply the seasonal Autoregressive Integrated Moving Average (SVARIMA)[2] approach to our problem. The key function of the stress value prediction is:

$$L_{n+1} = C + \sum_{i=0}^{k-1} A_i L_{n-i} + \sum_{i=0}^{k-1} B_i \overline{X_{n-i}} + \sum_{i=0}^{r-1} \theta_i \varepsilon_{n-i} + \varepsilon_{n+1}$$

L_{n+1} is the stress value we want to predict, where L is the past values of the corresponding predicted stress indexes (max, min and average). The $\overline{X}$ represents different features of the other correlated stress value sequences such as the (L_{sum}, $L_{proportion}$) to L_{avg}. The order k and r is determined by the Akaike's Information Criterion (AIC) and other parameters are estimated through damped least squares method. ε_i is the white noise error terms satisfying $E(\varepsilon_i) = 0$, and C is a constant.

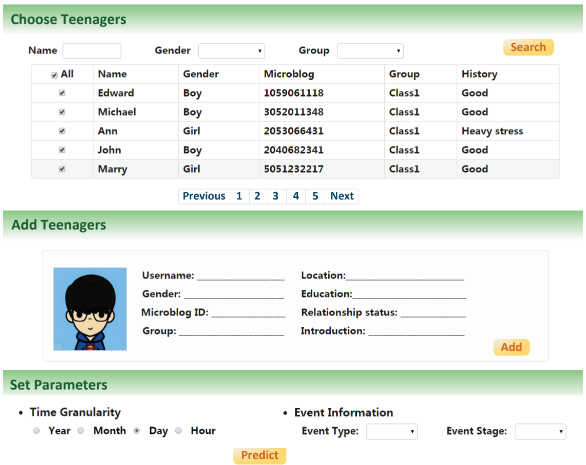
Stressful Event Influence. The SVARIMA model predicts stress values based on the historical stress change pattern. However, some happened or forthcoming stressful events may add extra influence which can't be neglected. Hence, we use the Moving Average Convergence/Divergence(MACD) to depict the extra stressful event influence. We first find



Microblog-based Teenagers Stress Prediction

Keep your face to the sunshine and you won't see the shadow.

Figure 3: The System Homepage



Choose Teenagers

Name: Gender: Group:

☑ All	Name	Gender	Microblog	Group	History
☑	Edward	Boy	1059061118	Class1	Good
☑	Michael	Boy	3052011348	Class1	Good
☑	Ann	Girl	2053066431	Class1	Heavy stress
☑	John	Boy	2040682341	Class1	Good
☑	Marry	Girl	5051232217	Class1	Good

Add Teenagers

Username: ______ Location: ______
Gender: ______ Education: ______
Microblog ID: ______ Relationship status: ______
Group: ______ Introduction: ______

Set Parameters

- Time Granularity
  ○ Year ○ Month ● Day ○ Hour
- Event Information
  Event Type: Event Stage:

Figure 4: User Requests

the stressful events and divide them into study and emotion event sets according to the users' tweets. Then we determine the start and end of the event influence cycle along with the stress value: 0. In the event influence cycle, we compute the MACD of the stress values, which can get a sequence representing the extra event influence. For the two event influence sequence sets, we use the Generalized Sequential Pattern(GSP) mining algorithm to mine the frequent sequence with 50% confidence to represent the extra influence of the two kinds of events. Finally, we divide equally the mined sequence into early, middle and later stage, and use the average MACD value of each stage as the adjustment value, which will be added to the original value predicted by SVARIMA model, if the predicted time is in the corresponding stage of different kinds of stressful events.

2.3.2 *Stress Trend Prediction*

Stress Trend Prediction aims to predict the future stress change trend including increase, decrease and remaining unchanged. We explore the candlestick chart to predict the stress trend. It outperforms the trend prediction method using the predicted stress values to subtract the last one, which might be a small fluctuation within the future trend.

Stress Reversal Signals. Fig. 2 (b) shows the decreasing reversal signals in a increasing process and Fig. 2 (c) shows the increasing reversal signals in a decreasing process. We take the decreasing reversal signals as examples for interpretation. For the 1-4 decreasing reversal signals, their last stress level is lower than the first stress level, which represents a decrease in the time unit of the stress candle stick and indicates a continual decrease trend in the future. For the 5-6 decreasing reversal signals, their max stress level is much higher than the last stress level. It indicates the heavy stress is gradually released through teenagers' stress regulation mechanism in the time unit of the stress candle stick, where the release mechanism is likely to remain effective in the future. For the 8-9 decreasing reversal signals, their first stress level is the same as the last stress level which means the power of stress release balances the stress accumulation. In the past increasing process, the power of stress accumulation is stronger than the stress release and now they are balanced. Therefore, the power of releasing stress may become stronger than accumulating stress next, which signifies the stress will decrease in the future.

Trend Decision Making. Let n be the current time, we trace back to the nearest local highest or lowest stress level and form a stress pattern $P_{current}$: $(SCF_{n-k+1}, ..., SCF_n)$ sequence, where k is the length of $P_{current}$.

[*Case 1*] If SCF_n is not the reversal signal, the stress change trend will continue.

[*Case 2*] If SCF_n is the reversal signal, we decide whether the stress change trend will reverse according to the past experience. Firstly, we define the distance $D(SCF_i, SCF_i^{'})$ between two SCFs to find similar past stress pattern P_{past} to the current stress pattern $P_{current}$.

$$D(SCF_i, SCF_j) = \sum_{k=1}^{5} w_k D(f_{ik}, f_{jk}), \quad \sum_{k=1}^{5} w_k = 1.$$

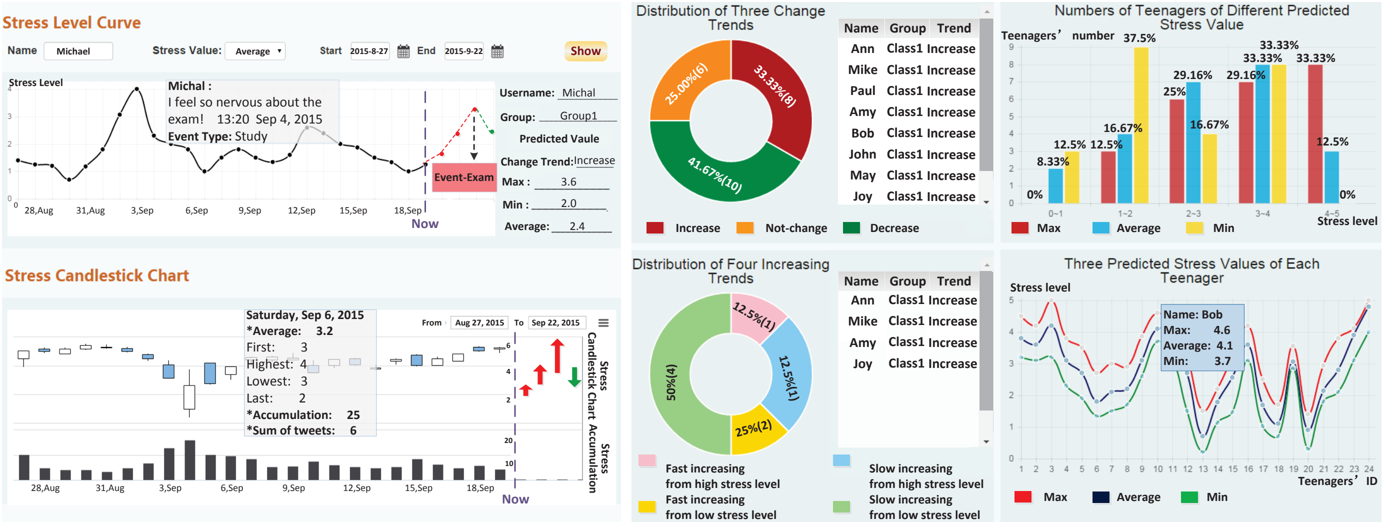
Here, f_{ik} and f_{jk} denote feature values of SCF_i and SCF_j, and w_k is the parameter determined by the analytic hierarchy process (AHP). For the nominal feature *Shape* of SCF,

$$D(f_{i1}, f_{j1}) = \begin{cases} 1, if\ f_{i1} \neq f_{j1} \\ 0, if\ f_{i1} = f_{j1}. \end{cases}.$$

For other four numeric features, $D(f_{ik}, f_{jk}) = |f_{ik}^{'} - f_{jk}^{'}|$, $j = 2, \cdots, 5$, where $f_{ik}^{'}$ and $f_{jk}^{'}$ are the normalized f_{ik} and f_{jk} between 0 and 1. We set a distance threshold to judge wether two SCFs match successfully. After matching the last SCF_{n-k+1}, we get a past pattern P_{past} with length of t. If the $matchrate = k/t$ is higher than a threshold, we consider the two patterns match successfully. Finally, we obtain a set of matched patterns $P_{matched}$ and observe the future trend after the reversal signals in $P_{matched}$, where the stress level time series have been segmented [4] to eliminate the influence of fluctuation to get the general change trend. We choose the majority change trends of $P_{matched}$ as the predicted result. If there is no matched sequence, the stress change trend will be predicted to reverse.

3. EVALUATION

We collect stress sequences of 91 teenagers obtained by the stress detection method whose precision is proved to be 82.6% [10], for the prediction results evaluation. For the stress value prediction, integrating stressful events and S-VARIMA model reduces the MAPE (Mean Absolute Percentage Error) of the predicted max, min and average stress level by 18%, 43%, and 3% separately, compared to the single SVARIMA model. For the stress trend prediction, the precision of our method achieves 83.79% which outperforms the time series (74.82%), MACD (63.59%) and KDJ (Stochastic Oscillator) (66.73%) based prediction approaches.



Name Michael Stress Value: Average Start 2015-8-27 End 2015-9-22

Michal :
I feel so nervous about the exam! 13:20 Sep 4, 2015
Event Type: Study

Username: Michal
Group: Group1
Change Trend: Increase
Max : 3.6
Min : 2.0
Average: 2.4

Saturday, Sep 6, 2015
*Average: 3.2
First: 3
Highest: 4
Lowest: 3
Last: 2
*Accumulation: 25
*Sum of tweets: 6

From Aug 27, 2015 To Sep 22, 2015

Name	Group	Trend
Ann	Class1	Increase
Mike	Class1	Increase
Paul	Class1	Increase
Amy	Class1	Increase
Bob	Class1	Increase
John	Class1	Increase
May	Class1	Increase
Joy	Class1	Increase

Name	Group	Trend
Ann	Class1	Increase
Mike	Class1	Increase
Amy	Class1	Increase
Joy	Class1	Increase

Name: Bob
Max: 4.6
Average: 4.1
Min: 3.7

(a) Individual Prediction Results

(b) Group Prediction Results

Figure 5: Prediction Results

4. SYSTEM DEMONSTRATION

During the demonstration, attendees can access *tPredictor* through the browser and experience our friendly interaction.

First, the user enters the home page of *tPredictor* and logins with his/her account as Fig. 3 shows.

After logining, the user enters the second page as Fig. 4 shows. In this page, the user can view the profile of the teenagers listed based on the user's search. The user can click teenagers in the list to see more detailed information such as the photo and self-introduction. Besides, a new teenager can be added to his/her teenagers list. For the prediction, the user chooses one or a group of teenagers to be predicted and sets the parameters including the time granularity and the stressful event information. Here the user set the time granularity to be day, event type to be study and the event stage to be early about the upcoming exam.

When the user chooses one teenager to be predicted, the prediction results are presented through two charts as Fig. 5 (a) shows. The stress curve exhibits the past and predicted stress values of the teenager, where the original tweets content and event information can also be presented in this chart when the user clicks the corresponding point. The stress candlestick chart is to demonstrate the future stress change trend of the teenager and the detailed stress-related indexes are also shown in this chart.

When the user chooses a group of teenagers to be predicted, he/she can see both individual prediction results of two charts in Fig. 5 (a) and the statistical group prediction results as Fig. 5 (b) shows. The two doughnut charts tell the user about the teenagers proportion of different future stress change trends and the corresponding teenagers will be listed directly when the user clicks one part of the doughnut chart. It helps the user easily know whose stress will increase. The histogram shows the corresponding number of teenagers whose future stress values are in different stress level range. Through the histogram, the user can understand the overall stress situation of the group in the future. For example, the stress situation of the group is severe when most teenagers' future stress values are between 2-5. The line chart shows the predicted stress value of each teenager of the group, through which the user can easily find teenagers who have much higher stress level

Acknowledgement

The work is supported by National Natural Science Foundation of China (61373022, 61532015, 71473146) and Chinese Major State Basic Research Development 973 Program (2015CB352301).

5. REFERENCES

[1] APA's 2013 Stress In America survey. http://www.apa.org/news/press/releases/stress/2013, 2013.

[2] G. Box and G. Jenkins. *Time series analysis:Forecasting and Control*. Holden-Day, San Francisco, 1970.

[3] M. De Choudhury, M. Gamon, S. Counts, and E. Horvitz. Predicting depression via social media. In *ICWSM*, 2013.

[4] J. Jiang, Z. Zhang, and H. Wang. A new segmentation algorithm to stock time series based on pip approach. In *Wireless Communications, Networking and Mobile Computing, 2007. WiCom 2007. International Conference on*, pages 5609–5612. IEEE, 2007.

[5] Q. Li, Y. Xue, J. Jia, and L. Feng. Helping teenagers relieve psychological pressures: A micro-blog based system. In *EDBT*, pages 660–663, 2014.

[6] Y. Li, Z. Feng, and L. Feng. Using candlestick charts to predict adolescent stress trend on micro-blog. *Procedia Computer Science*, 63:221–228, 2015.

[7] Y. Li, J. Huang, H. Wang, and L. Feng. Predicting teenager's future stress level from micro-blog. In *Proc. of CBMS*, 2015.

[8] M. Park, D. W. McDonald, and M. Cha. Perception differences between the depressed and non-depressed users in twitter. In *Proc. of ICWSM*, 2013.

[9] X. Wang, C. Zhang, Y. Ji, L. Sun, L. Wu, and Z. Bao. A depression detection model based on sentiment analysis in micro-blog social network. In *Trends and Applications in Knowledge Discovery and Data Mining*, pages 201–213. Springer, 2013.

[10] Y. Xue, Q. Li, L. Jin, L. Feng, D. A. Clifton, and G. D. Clifford. Detecting adolescent psychological pressures from micro-blog. In *Health Information Science*, pages 83–94. Springer, 2014.