USER-LEVEL PSYCHOLOGICAL STRESS DETECTION FROM SOCIAL MEDIA USING DEEP NEURAL NETWORK

1st Author
1st author's affiliation
1st line of address
2nd line of address
Telephone number, incl. country code
1st author's E-mail address

2nd Author
2nd author's affiliation
1st line of address
2nd line of address
Telephone number, incl. country code
2nd E-mail

3rd Author
3rd author's affiliation
1st line of address
2nd line of address
Telephone number, incl. country code
3rd E-mail

ABSTRACT

It is of significant importance to detect and manage stress before it turns into severe problems. However, existing stress detection methods usually rely on psychological scales or physiological devices, making the detection complicated and costly. In this paper, we explore to automatically detect individuals' psychological stress via social media. Employing real online micro-blog data, we first investigate the correlations between users' stress and their tweeting content, social engagement and behavior patterns. Then we define two types of stress-related attributes: 1) low-level content attributes from a single tweet, including text, images and social interactions; 2) user-scope statistical attributes through their weekly micro-blog postings, leveraging information of tweeting time, tweeting types and linguistic styles. To combine content attributes with statistical attributes, we further design a convolutional neural network (CNN) with cross autoencoders to generate user-scope content attributes from low-level content attributes. Finally, we propose a deep neural network (DNN) model to incorporate the two types of user-scope attributes to detect users' psychological stress. We test the trained model on four different datasets from major micro-blog platforms including Sina Weibo, Tencent Weibo and Twitter. Experimental results show that the proposed model is effective and efficient on detecting psychological stress from micro-blog data. We believe our model would be useful in developing stress detection tools for mental health agencies and individuals.

Keywords

Stress detection, convolutional neural network, cross autoencoders, deep learning, micro-blog, social media

1. INTRODUCTION

1.1 Motivation

Psychological stress is the root cause to many health problems and mental diseases. Chronic stress increases the risk of developing health problems such as insomnia, obesity, heart diseases, cancer etc. [1]. Many studies have revealed a link between stress and mental diseases like anxiety disorders, depression etc. [2]. Stress has been a threat to human health for a long time. Time magazine's June 6, 1983 cover story called stress "The Epidemic of the Eighties" and referred to it as our leading health problem [1]. Meanwhile, stress has been progressively worsened and spread recent years. With the rapid development of modern society, many people feel increasingly stressed under the rapid pace of life. Numerous surveys have confirmed that adult Americans are feeling under much more stress than a decade or two ago. A 1996 Prevention magazine survey found that almost 75% feel they have "great stress" one day a week and with more than 30% indicating they feel this way more than twice a week, which is 55% compared to the same survey conducted in 1983[2]. In a word, the rapid increase of stress has become a great challenge to human health and life quality.

Psychological stress detection remains a large problem at the present stage. Detecting and managing stress before it turns into severe problems is of significant importance. Recent decades, many efforts have been devoted to stress detection by researchers from diverse areas. They have developed many methods to measure psychological stress, including psychological questionnaire based interviews [3, 4] and physiological signal based measures [5, 6]. However, these methods have their limitations in many aspects. Psychological questionnaires often contain a range of questions designed by psychologists. People are usually unwilling to do these questionnaires unless they have to. Physiological methods usually require professional devices to measure users' physiological and biochemical properties and need specialists to analyze the acquired data. Thus, it is very important and useful to find a way to detect user's stress state reliably, automatically and non-invasively.

With the fast development of social networks, people are widely using social media platforms to share their thoughts and feelings. A statistic report from statisticbrain.com (http://www.statisticbrain.com/twitter-statistics/) shows that by 2014.1.1, the total number of active registered users on Twitter has reached more than 645 million, with an average 58 million tweets posted per day. As for Sina weibo (the largest micro-blog platform in China), the number of weibo users has reached more than 600 million[3]. People post tweets containing text and images on micro-blog platforms to share opinions, express emotions, record daily routines and communicate with friends. We can obtain linguistic and visual content that may indicate stress related symptoms. This makes the detection of users' psychological stress through their tweets and posting patterns from micro-blog feasible.

Conference'10, Month 1–2, 2010, City, State, Country.

[1] http://www.stress.org/americas-1-health-problem/

[2] http://www.anxietycentre.com/stress.shtml

[3] http://www.comsoc.org/blog?page=3

1.2 Related Work

Existing methods for stress detection. Many efforts have been devoted to developing convenient tools for individual stress detection recent years. Researchers are trying to leverage pervasive devices like personal computers and mobile phones for routine stress detection. Hong L. etc. [7] proposed StressSense to unobtrusively recognize stress from human voice using smartphones. Paredes, P. etc. [8] investigated the initial lab evidence of the use of a computer mouse in the detection of stress. However, such applications rely on collecting one's real-life data, which is easy to trigger antipathy. It makes stress detection invasive to normal life, and can't be used widely in more people.

Researches on using social media for healthcare. With the rapid spread of social networks, researches on using social media data for physical and mental healthcare are also increasingly growing. Sadilek et al. [9] leverage Tweeter postings to identify the spread of flu symptoms. Paul M.J. etc. [10] apply the Ailment Topic Aspect Model to over 1.5 million health related tweets and discover correlations between behavioral risk factors and aliments. Munmun etc. [11] leverage behavioral cues indicated from Twitter postings to predict depression before it is reported. These studies show the feasibility of harnessing social media data for developing healthcare tools. However, they mainly leverage the textual content in the social media data, while other equally important content, like images and social behavior are ignored.

Deep learning approaches for cross-media data modeling. Micro-blog data is typical cross-media data. Items may come from diverse sources and modalities. It is difficult to handle the heterogeneous cross-media data. Recent years, extensive researches on deep learning show superior ability of deep neural networks (DNN) in learning features from large scale unlabeled data [12-14]. [15, 16] further extend the deep models for multimodal learning. [17] design a cross-media learning method based on DNN, and leverage the model for detecting psychological states and corresponding categories from a single tweet. However, stress is a continuous state compared to instant emotions, indicating that the stressed stated can last for several days in psychology [3]. It remains a challenge to make use of aggregated cross-media data for user-level modeling.

1.3 Our Work

In this paper, we explore the potential to use social media to detect psychological stress for individuals. Micro-blog is one of the most popular social media that can be publicly accessed. People can post text with no more than 140 words, upload images or have social interactions with others. Employing real online micro-blog data, we first investigate the correlations between users' stress and their tweeting content, behavior patterns and social engagement. Then we define two types of stress-related attributes: 1) low-level content attributes from a single tweet, including text, images and social interactions like comments, retweets and favorites; 2) user-scope statistical attributes through their weekly micro-blog postings, leveraging information of tweeting time, tweeting types, linguistic styles, and social engagement with friends indicated from the @-mentions and @-replies, etc. To combine low-level content attributes with user-scope statistical attributes, we further design a convolutional neural network (CNN) with cross autoencoders to learn the latent high-level attributes on cross-modal units [17][18]. Finally, we propose a deep neural network (DNN) model to incorporate the two types of user-scope attributes to detect users' psychological stress. The experimental results on four datasets from different Micro-blog platforms indicate the effectiveness and efficiency of the proposed method.

Table 1. The details of the observation dataset (DB1).

Tweets' label	Number of tweets	Number of users	Number of weeks	Number of tweets per week
Non-Stressed	253638	12230	17861	14.2
Stressed	239038	11074	19136	12.5
Summation	492676	23304	36997	13.3

We have to face several challenges in this work. And the corresponding contributions are:

1） **Challenge 1**: Micro-blog platforms contain massive data. It is infeasible to manually label the data. How to find effective methods to automatically label the ground truth remains a challenge.

Our solution: Inspired by previous research [19], we have built a stressed-twitter-posting database using the "I feel stressed" sentence pattern as the ground-truth label for detecting stress from micro-blog data. With a small set of psychological stress scale score labeled dataset as test, it is proved that our ground truth labeling method is reliable;

2） **Challenge 2**: Attributes in a tweet come with multiple modalities and the components are often incomplete, which is a typical problem in cross-media. Numbers of tweets in a certain period of time also differ from person to person and from week to week. Traditional models have limited abilities to extract modality-invariant attributes from such data.

Our solution: We design a convolutional neural network with cross autoencoders to aggregate low-level content attributes and generate modality-invariant user-scope attributes which support user-level stress detection;

3） **Challenge 3**: Modeling stress in user-level is more difficult than in discrete tweet-level, since both the overview and detailed attributes should be concerned about.

Our solution: We propose a stress detection model based on DNN to incorporate content attributes and statistical attributes together. The DNN model along with CNN forms a unified integral deep network which can extract attributes from single tweets and detect user-level continuous psychological stress.

2. DATA OBSERVATION

2.1 Observation dataset

We first crawl 350 million tweets data via Sina Weibo's streaming APIs from 2009.10 to 2012.10. Then we collect tweets containing sentence patterns like "I feel stressed this week" and "I feel stressed so much this week" as the weekly stressed state label, and tweets containing "I feel relaxed" and "I feel non-stressed" as the non-stressed state label. The "I feel" pattern has been proved to be effective as ground truth data labels in emotion analysis in [19]. In this way, we collect over 19000 weeks of users' tweets that are labeled as stressed, and over 17000 weeks of non-stressed users' tweets. There are 492,676 tweets from 23304 users in total. We take this dataset for observation and further experiments, which is represented by DB1 in this paper. The details of the dataset are shown in Table 1.

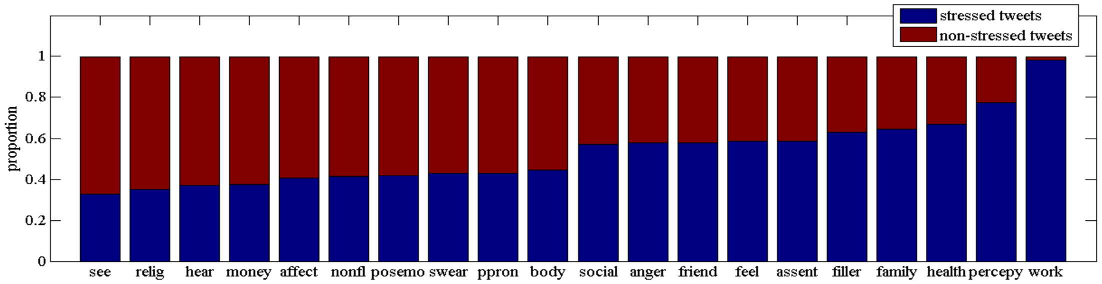

Figure 1. The proportion of top 12 most frequently occurred word categories from non-stressed and stressed weeks of tweets

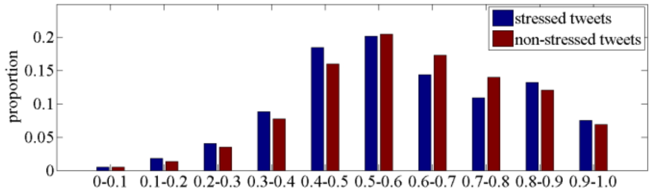

Figure 2(a). Brightness distribution from tweets' image

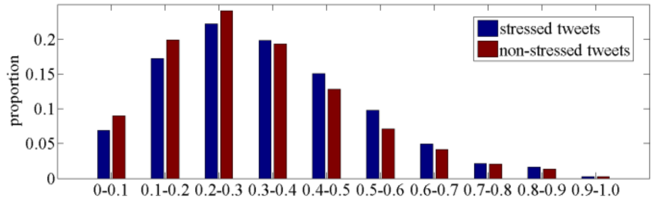

Figure 2(b). Saturation distribution from tweets' image

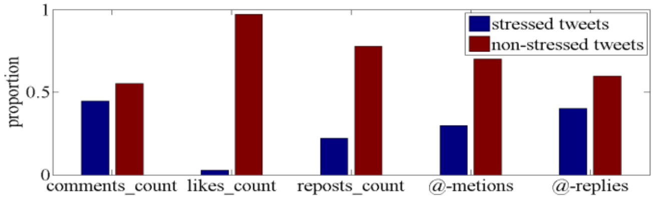

Figure 3. Social engagement observation

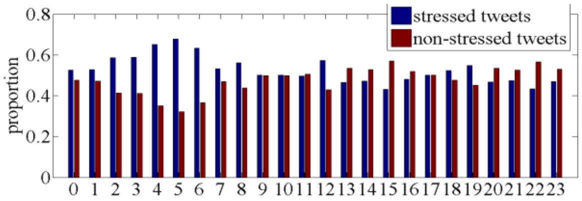

Figure 4. Tweeting time observation

2.2 Observation and analysis

We first conduct a series of analyses on the DB1 and present some patterns related to individuals' psychological stress reflected by tweets. In the analysis, we randomly pick 1000 weeks of stressed and non-stressed tweets from the DB1 and focus on the following aspects:

- *Content correlation*: the difference of stressed and non-stressed tweets in tweets' content, including text and images;
- *Social engagement correlation*: the difference between stressed and non-stressed weekly tweets on users' social interactions with friends via @-mentions, @-replies and tweets' comments, retweets and likes;
- *Behavioral correlation*: the difference of stressed and non-stressed tweeting behavior in tweeting frequency, tweeting types and tweeting time.

2.2.1 Observations on content correlation

Tweets on micro-blog mainly consist of text and images. We leverage a widely used psychological dictionary LIWC [20] to measure the most frequently occurred words in stressed and non-stressed tweets text content. The results are shown in Figure 1.[4] From the figure, we observe that there is evident difference in text content between the stressed and non-stressed tweets. For the non-stressed tweets, there are more words categories from negative emotions, social, friends and family etc. While for the stressed tweets, there exist more word categories from positive emotions, work, health and anxiety etc.

As for the image content of tweets, we consider brightness and saturation as observed visual features. The results are shown in Figure 2(a) and Figure 2(b).

From Figure 2(a), we can observe that the presence of images with low brightness (<0.3) from stressed class is obviously higher than that from non-stressed class, indicating that stressed users are more likely to post images with lower brightness.

As for the saturation distribution in Figure 2(b), we observe that the saturation of non-stressed users' images are more likely to be lower (<0.5), while the stressed class is more likely to be in the higher range (>0.5).

2.2.2 Observations on social engagement correlation

Micro-blog is an important platform for users to share information and interact with friends. The social interactions on micro-blog usually consist of @-mentions, @-replies, retweets, comments and likes etc. We analyze the correlation between social interactions and users' stress states.

Figure 3 shows the social interaction patterns from tweets of users in stressed and non-stressed states. The patterns are measured as the proportion of the numbers of comments, likes, retweets, @-mentions and @-replies in users' weekly tweets.

From the figure, we observe that for the non-stressed class, users' tweets get more comments, likes and retweets from friends, indicating that people are generally more likely to interact with the followed users when they are at a non-stressed state. Meanwhile, compared to non-stressed weeks, the stressed weeks have less @-mentions and @-replies of friends. This also proves that stressed users are less social active than non-stressed users.

[4] http://www.liwc.net/descriptiontable1.php gives the full name of the word categories.

2.2.3 Observations on Behavioral Correlation

As revealed by psychology theories [1], there are many common symptoms may be related to stress, including insomnia, social withdrawal .etc. These symptoms can also be reflected by tweeting behavior changes on micro-blog. We observe tweeting time distributions to measure users' tweeting behavior.

Figure 4 shows the results of tweeting time distribution of users from the two classes. Tweeting time distribution is measured in tweet postings in hours of a day. From the result, we observe that there are more stressed postings during 0 to 6 in the morning, revealing that stressed users are more likely to be insomnia.

Summary To very briefly summarize, we have the following intuitions which will be further leveraged and incorporated in our method design:

- The different content of a single tweet including text, image and social interactions are all related to different one's stress state at some point.
- One's stress state can be related to the social engagement with friends in weekly unit.
- One's stress state can also be related to the tweeting behavior on micro-blog.

3. ATTRIBUTES DEFINITION

The micro-blog data is a typical type of cross-media data, containing text, emoticons, images and social interactions. Besides, the patterns of micro-blog usage behavior in a period such as one week unit also contain useful information for stress detection. To leverage both content information contained in single cross-media micro-blog tweet and the micro-blog usage behavior in weekly tweets, guided by psychological theories, we define two sets of attributes to measure the differences of the stressed and non-stressed users on micro-blog: 1) *content attributes* from the content of a single tweet; 2) *statistical attributes* from the users' behavior of weekly tweet postings.

3.1 Content Attributes

The content of a tweet from micro-blog usually consists of text, image and social interaction. We define linguistic, visual and social attributes from each part of a tweet respectively as follows:

1) Linguistic Attributes:

As users usually express their emotions using tweets, we measure the emotions in a single tweet using linguistic attributes. To describe the linguistic attributes, we leverage a psychological dictionary named "Language Inquiry and Word Count Dictionary" [20]. The simplified Chinese LIWC dictionary [21] is developed by Chinese psychologists and linguists, based on the psycholinguistic dictionary LIWC (http://www.liwc.net), which has been proved to be effective on determining affect in Twitter. It is composed of almost 4500 words and categorized into over 60 categories [20].

Based on the dictionary, we define the text content related features as the tweet's linguistic attributes:

- **Positive and Negative Emotion Words** (2 dimension). Measured by the number of positive and negative emotion words in the tweet's text, indicating how positive or negative emotions are expressed in the tweet.
- **Positive and Negative Emoticons** (2 dimensions). Measured by the number of positive and negative emoticons. Emoticons are widely used in micro-blog platforms to express users' emotional states. We manually categorize the 129 emoticons provided by Sina Weibo platform into positive and negative categories.
- **Punctuation Marks and Associated Emotion Words** (4 dimensions). We use this attribute to signify the intensity of emotion in a tweet, either positive or negative according to the associated emotional words. Four typical punctuation marks (exclamation mark, question mark, dot mark and the Chinese full stop mark "。") are considered.
- **Degree Adverbs and Associated Emotion Words** (2 dimensions). Degree adverbs are also used to express the degree of emotions. For example, "I feel a little bit sad" and "I feel terribly sad" express different level of negative feelings. We use a number range of 1-3 to represent neural, moderate and severe degrees of positive expression and the minus to represent the negative ones.

Thus, we get 10-dimensional vector to denote the linguistic attributes from a tweet's text content.

2) Visual Attributes:

Based on previous work on affective image classification [22] and color psychology theories [23], we combine the following features as the visual middle-level representation:

- Five-color theme (15 dimensions): a combination of five dominant in the HSV color space, representing the main color distribution of an image. It has been revealed to have important impact on human emotions according to psychology and art theories [22].
- Saturation (2 dimensions): the mean value of saturation and its contrast.
- Brightness (2 dimensions): the mean value of brightness and its contrast.
- Warm or cool color (1 dimension): ratio of cool colors with hue ([0-360]) in the HSV space between 30 and 110.
- Clear or dull color (1 dimension): ratio of colors with brightness ([0-1]) and saturation less than 0.6.

Thus, based on the psychological studies and color theories, we finally get a 21 dimensional attributes from the tweet's image content.

3) Social Attributes:

Besides the text content and image content of a tweet, some additional features like comments, retweets and likes indicate the tweet's social attention from one's friends. They can also imply one's stress state to some degree. We use the number of comments, retweets and likes of a tweet to measure the tweet's social attention degree into social attributes. Thus, we get a 3-dimensional vector to represent the social attributes of a tweet.

3.2 Statistical Attributes

Statistical attributes are summarized from users' tweets in a specific sampling period. We use one week as the sampling period in this paper. On one hand, psychological stress often results from cumulative events or mental states; on the other hand, users may express their chronic stress in a series of tweets rather than one. Appropriately designed statistical attributes can provide a macro-scope of a user's stress states, and avoid noise or missing data. We define statistical attributes from three aspects to measure the differences between stressed and non-stressed states based on users' weekly tweet postings. The details of the statistical attributes are described as follows:

1) Social Engagement:

We consider 3 measures to characterize the social engagement from users' weekly tweet postings: the @-mentions, @-replies and the retweets from a user's friend. These three behaviors are the most commonly used ways to interact with friends on micro-blog platforms. Unlike the social attributes in a single tweet, the social engagement attributes are measured in numbers of @-mentions and @-replies in weekly tweet postings, indicating one's social interaction activeness with friends.

2) Behavioral Attributes:

We define a set of behavioral measures for users, including tweeting time and tweeting types, based on the weekly tweet postings. These measures are described as follows:

- **Tweeting time:**

Tweeting time can indicate users' daily routines at some point. We consider two measures that derive from the tweeting time information of tweets: tweeting frequency and tweeting time distribution. Tweeting frequency is measured in the average number of tweets posted in a day, while tweeting time distribution is measured in numbers of tweets posted in hours with a 24 dimensional vector.

- **Tweeting Type:**

Users usually post tweets on micro-blog with diverse motivations, making the tweets to be presented in different types. We categorize users' tweets into mainly four types: 1) image tweets (tweets containing images) 2) original tweets (tweets that are originally posted by tweets' users) 3) information query tweets 4) information sharing tweets (tweets that contain outside hyperlinks). We use a 4-dimensional vector of the numbers of tweets in the above 4 types respectively to represent the tweeting type attribute.

3) Linguistic Style:

We introduce measures to characterize linguistic styles in users' weekly tweet postings using the psychological dictionary LIWC [20]. LIWC categorizes frequently-used words into more than 60 categories. We adapt 10 categories from LIWC that are related to daily life, social events, e.g.: personal pronouns, home, work, money, religion, death, health, ingestion, friends and family. We extract words from users' weekly tweet postings and use a 10 dimensional vector of numbers of words in the 10 categories to represent the linguistic style attribute. Different from the linguistic attributes of a single tweet which mainly measures the emotions, the linguistic style can measure one's linguistic behavior in aggregated tweets.

4. MODEL AND LEARNING

4.1 Architecture

As described in section 3, we define low-level content attributes from each single tweet in tweet-scope, and statistical attributes from aggregated tweets in user-scope. These two sets of attributes cannot be combined directly since their mathematical descriptions are not in the same domain. So we need to generate latent user-scope content attributes from low-level content attributes at first. After that, both of the two user-scope attribute sets can be finally fed into a classifier for user-level stress detection.

In the following sections, we will address our solution through the following two key components: 1) First we design a **convolutional neural network with cross autoencoders** to generate user-scope content attributes from low-level content attributes, thus the tweet-scope content attributes can be combined with the user-scope statistical attributes; 2) We propose a deep neural network model to incorporate the two types of user-scope attributes for user-level psychological stress detection.

Table 2. List of notations that are used in the following sections.

.x	An input pattern to the network
y	Hidden activation of units in an AE
$\hat{x}$	Reconstruction of input pattern x from an AE
$W^{(l)}$	Connection weight matrix of layer l in a network
$b^{(l)}$	Activation bias of units of layer l in a network
$g^{(l)}$	Activation function of units of layer l in a network
$S, \dot{S}$	Sigmoid function and its derivate
θ	The set of parameters in a network
J	Performance function of a network
λ	Weight decay penalty
Ω	Set of all modalities in the problem domain
K	Number of modalities in the problem domain
M	Subset of modalities, whose elements are actually presented to the network
x_i	An input pattern with modality i
x_s	An input pattern with modalities in set s
f_i	Modality specified mapping of modality i
w_i	Weight matrix of modality i
$\hat{x}_i$	Reconstruction of input pattern with modality i from CAE
$\hat{x}_s$	Reconstruction of input pattern with modalities in set s from CAE
P	Subset of modalities, whose elements are available in dataset

4.2 Cross autoencoders

Rather than summarizing the user's state alone, we further incorporate the detail attributes with multiple modalities of every tweets by utilizing a recently proposed cross-media model, namely the Cross Autoencoders (CAE) [17].

An autoencoder is a basic unit in deep neural networks for learning distinctive attributes from data [12-14]. It is a shallow network containing one hidden layer and paired input/output layers. The network is trained to reconstruct input pattern from activation of the hidden layer, which is actually stimulated by the input itself. The reconstruction can be formulated by

$$\begin{cases} y = g^{(1)}\left(W^{(1)}x + b^{(1)}\right) \\ \hat{x} = g^{(2)}\left(W^{(2)}y + b^{(2)}\right) \end{cases} \quad (1)$$

where x is the input pattern and y is the activation of hidden units. $\hat{x}$ denotes the reconstruction result from the hidden units. $W^{(1)}$ and $W^{(2)}$ are the connection weights while $b^{(1)}$ and $b^{(2)}$ are bias to the postsynaptic units. $g^{(1)}$ and $g^{(2)}$ are activation functions of the units, where sigmoid functions are rational choice in many scenario. In this work, we use the following sigmoid function as activation function of all neurons:

$$S(x) = \frac{1}{1+e^{-x}}, \tag{2}$$

and its derivative is then given by $\dot{S}(x) = S(x)\big(1 - S(x)\big)$.

To train the autoencoder to reconstruct input pattern and learn distinctive attributes on the hidden layer, we minimize the following performance function by updating the parameter set $\theta = \{W^{(1)}, W^{(2)}, b^{(1)}, b^{(2)}\}$ with gradient descent

$$J(\theta; x) = \frac{1}{2}\|\hat{x} - x\|^2 + \frac{\lambda}{2}\sum_{i,j,l}\left(W_{ij}^{(l)}\right)^2, \tag{3}$$

where the second term is a regularization to prevent model over fitting which is known as weight decay.

In order to learn attributes robust to partial corruption of the input pattern, Vincent et al. proposed a noticeable extension to the autoencoders, namely the denoising Autoencoder (dA) [24]. The autoencoders are trained to reconstruct a repaired input from corrupted ones, what is done by first corrupting original input pattern to get the corrupted version with stochastic sampling. Part of the components is randomly chosen and set to 0 while others remain unchanged. The trained model is robust to small irrelevant changes in input.

Researchers further extended the model for multimodal learning [15, 16] and cross-modal learning [17]. Lin et al. proposed an autoencoder structure to leverage data with different modalities from different sources for utilizing superior attribute learning power of deep networks on cross-media data [17], namely the CAE. Similar to dA, CAE holds back part of components of training data and require the model to reconstruct the held back part. Rather than select randomly like dA do, CAE orderly hold back each combination of modalities. The feedforward pass of CAE can be formulated by

$$\begin{cases} f_i(s) = w_i s,\ i \in M \\ y = g^{(1)}\left(\sum_{i\in M} f_i(x_i) + b^{(1)}\right) \\ \hat{x}_\Omega = g^{(2)}\left(W^{(2)}y + b^{(2)}\right) \end{cases} \tag{4}$$

where f_i is the modality specified mapping and x_i is data of modality $i \in M$. Denoting $\Omega = \{1,2,\dots,K\}$ as the set of all modalities, $M \subseteq \Omega$ is the set of modalities that data are available with. $W^{(1)} = [w_1, w_2, \dots, w_K]$ is the encoder layer weight matrix while $b^{(1)}$ is the bias. $g^{(1)}$ is the activation function of hidden neurons and y is activities of them. The last equation represents the decoder part. $W^{(2)}$ is the weight matrix and $b^{(2)}$ is the bias. $g^{(2)}$ is the activation function of reconstruction neurons. $\hat{x}_\Omega = [\hat{x}_1; \hat{x}_2; \dots; \hat{x}_K]$ is the reconstruction for all modalities in the problem domain.

CAE can be trained with standard gradient descent algorithms, but with a special designed data set. Data available with all modalities are used for training. The network is fed with data which contain combination of modalities. Error feedback is calculated on all modalities and updates the network by back-propagation. The new performance function can be formulated by

$$J(\theta; x) = \frac{1}{2}\sum_{i\in\Omega}\|\hat{x}_i(x_M) - x_i\|^2 + \frac{\lambda}{2}\sum_{i,j,l}\left(W_{ij}^{(l)}\right)^2, \tag{5}$$

where we explicitly denote $\hat{x}_i$ as a deterministic function of x_M.

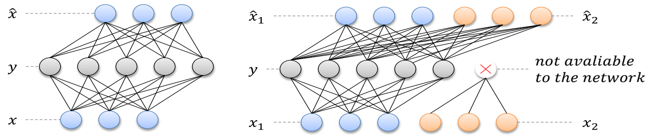

Figure 5. Comparison of standard autoencoder and CAE.

Figure 5 demonstrates a comparison on structure of standard autoencoder and CAE. The key idea here is, as data with M is presented, data with all modalities Ω are required to reconstruct.

One limitation of this training method is that, in a real world problem, data with all modalities are rare. To make use of more data available with a couple of (not all) modalities, we further employ the Extended Feature Learning (EFL) phase. In this phase, error feedback is calculated for available modalities. Thus the performance function is given by

$$J(\theta; x) = \frac{1}{2}\sum_{i\in P}\|\hat{x}_i(x_P) - x_i\|^2 + \frac{\lambda}{2}\sum_{i,j,l}\left(W_{ij}^{(l)}\right)^2, \tag{6}$$

where $P \subset \Omega$ is the set of available modalities.

It has to be noticed that EFL should be carried out after initial training of CAE that general correlation between modalities are learned. Otherwise the network may tend to learn trivial attributes for each modality as mentioned in [15].

Micro-blog data is typical cross-media data. It consists of text body, emoticons, attached images, replies and retweets from other users etc. Meanwhile, these parts do not necessarily exist for any tweet item. Using CAE, we can model tweets to capture relationship of information with different modalities and learn modality-invariant attributes.

4.3 Convolutional neural networks with CAE

The attributes of tweets from a user's weekly tweet postings in timeline form a time-series. To further model a user as a subject of series of tweets, we apply Convolutional Neural Networks (CNN) [18, 25]. CNNs have large learning capacity while having much fewer connections and parameters to learn comparing to similar size standard network layer. They focus on learning stationary local attributes for series like images (pixel series), speeches and other time-series. We can learn user-scope content attributes from a series of single tweet in time-series to describe one's stress state in a week.

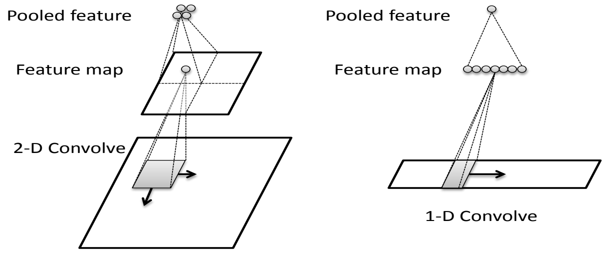

Figure 6. CNN units for 2-D convolution and 1-D convolution.

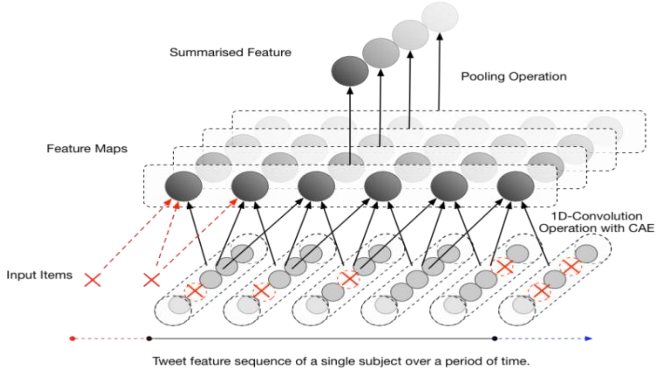

Figure 7. CNN with CAE units.

Table 3. The details of the collected tweet dataset DB2-DB4 from different micro-blog platforms.

Platform	Stress label	Number of tweets	Number of users	Number of weeks	Tweets per week
DB2:Sina Weibo (2010.2-2011.9)	stressed	1459	98	98	14.9
	non-stressed	1845	112	112	16.5
	summation	3304	210	210	15.7
DB3:Tencent Weibo (2011.11-2013.3)	stressed	138570	7845	8974	15.4
	non-stressed	172585	8239	9976	17.3
	summation	311155	16084	18950	16.4
DB4:Twitter (2009.6-2009.12)	stressed	54748	4905	6081	9.0
	non-stressed	75357	4018	6545	11.5
	summation	130105	8923	12626	10.3

After this step, all attributes of tweets in time-series are formed to a 1-Dimensional series. Figure 6 demonstrates a 2-D unit (left) which is often used in image processing and a 1-D unit (right) which is used in our model.

We propose to use CAE rather than standard units in CNN for tweet-series. 1) On one hand, CAE units works directly on multi-modality attributes of tweets because CAE can learn modality-invariant attributes of tweets directly. Consequently the generated attribute maps are also modality-invariant and the rest of the network can work in modality-invariant domain. 2) On the other hand, individuals may have fewer tweets than the patch size of convolutional units. We call it "missing tweet" phenomenon. In such case, we cannot apply attribute extraction with standard units, while we can handle such users with CAE by treating their tweet-series as incomplete patches.

Figure 7 presents a detailed view of CNN with CAE units. Each cylinder is a tweet instance with multiple modalities. Red circles with cross are attributes with missing modalities. The cylinders form a 1-D sequence along the time-line in a week. The leftmost red crosses are 'missing tweet' instances when the patch is applied to the first instance of the week. In our case, three continuous instances make a patch. CAE units are listed in the attribute maps. They connect to a patch of instance. CAE units take patches with missing modalities as well as 'missing' instance, and generate the modality-invariant attribute maps.

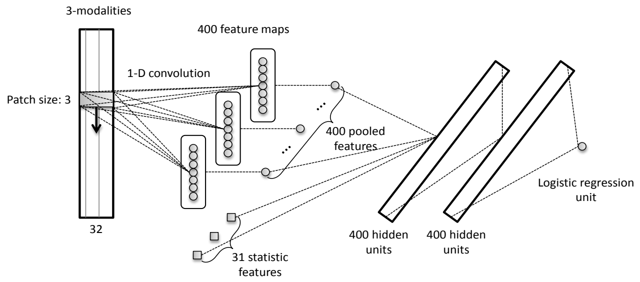

Figure 8. Overall architecture of the proposed model with DNN classifier.

Pooling is another important step to summarize attribute maps into fewer attribute instances. Though different users have different number of tweets in different weeks, the period of time that the tweets are sampled are the same. We simply pool each attribute map into one pooled attribute. There are two commonly used pooling operations: max-pooling and mean-pooling. When max pooling is used, the pooled attribute unit is assigned with the maximally activation among all units in the attribute map. When mean-pooling is applied, the mean of activations of all units in the attribute map is assigned to the pooled attribute unit. Since we pool over the period of time rather than a certain number of tweets, we consider using mean-over-instances (MOI) and mean-over-time (MOT). Mean-over-instances is simply the average value of activations of the units while mean-over-time can be calculated by summing up the activations since they are sampled in same length of time. We test all the three pooling methods in our experiments in Section 5.

4.4 Classifiers

Through the key steps in the above section, we get two types of user-scope attributes: content attributes learnt by CNN with CAE units, and the manually defined statistical attributes (Section 3.2). We can measure one's stress using these user-scope attributes now. Determining the stress states of a user can be formulated as a typical binary classification problem. In general, any state-of-the-art classifier can be utilized to predict one's stress states with the extracted attributes. In our work, we focus on using a deep neural network (DNN) over the extracted attributes for classification [14]. Deep neural networks have shown superior ability in a variety of classification tasks. The deep architecture of the network can further learn higher-level semantically related attributes from the user-scope attributes. Since both the content attributes learnt by CNN with CAE units and the manually defined statistical attributes are modality-invariant, we use standard fully-connected layers of stacked autoencoders and classify with a final logistic regression unit. A 4-layer architecture is used in this paper. Figure 8 demonstrates an overall architecture of our proposed model with DNN classifier.

5. EXPERIMENTS

5.1 Experimental setup

Dataset. We perform our experiments on four datasets DB1-DB4 collected from three different micro-blog platforms: Sina Weibo, Tencent Weibo[5], and Twitter. DB1 from Sina Weibo has the most number of tweets and users which has been described in Section 2, Table 1. The details of the other 3 datasets are shown in Table 3. The Tencent Weibo (DB3) and Twitter (DB4) are labeled using the sentence pattern method described in Section 2. Especially, to avoid the noise in data ground truth, we establish a small scale dataset DB2 from Sina Weibo. DB2 is collected from the users that have shared the score of a psychological stress scale[6] with 50 items via Sina Weibo. If the resulted score is over 80, then the test subject is claimed to be stressed. We crawl the shared scores and the corresponding users' information and weeks' tweets. In this way, for DB2 we finally get 98 weeks of stressed tweets (scale score > 80) and 112 weeks of non-stressed tweets (scale score < 80) as a small but reliable ground truth data to further validate the reliability of the sentence pattern based ground truth labeling method.

In the following experiments, we first train and test our model on the large-scale Sina Weibo dataset DB1. Then we further test our model on the other 3 datasets to show effectiveness of the proposed model on different data sources or different ground truth labeling methods. For all of our analyses, we use 5-fold cross validation, over 10 randomized experimental runs.

[5] http://t.qq.com another popular micro-blog platform in China.

[6] http://types.yuzeli.com/survey/pstr50

Table 4. Comparison of results using different classification models.

Pooling Method	Classifier	SVM			RF			NB			DNN		
	Feature	statistic	content	both	statistic	content	both	statistic	content	both	statistic	content	both
Max	Accuracy		0.7117	**0.7562**		0.7340	**0.7675**		0.6714	0.6727		0.7186	0.7635
	F1-score		0.7921	**0.8312**		0.8109	**0.8341**		0.7493	0.7504		0.8049	0.8284
MOI	Accuracy	0.6987	0.6569	0.7433	0.7329	0.7327	0.7630	0.6432	**0.7099**	0.6813	0.7146	0.7201	0.7640
	F1-score	0.7999	0.7929	0.8224	0.8124	0.8071	0.8292	0.7496	**0.7896**	0.7567	0.7965	0.8040	0.8285
MOT	Accuracy		0.7165	0.7457		0.7186	0.7556		0.6480	0.6484		0.7227	**0.7857**
	F1-score		0.8062	0.8261		0.7970	0.8255		0.7701	0.7700		0.8072	**0.8443**

Comparison Methods. We compare the following classification methods for user-level psychological stress detection:

- **Naive Bayes (NB)** is a simple probabilistic classifier based on Bayes' theorem that calculates the posterior probability by calculating prior probability of attributes. The classifier assigns sample with the largest calculated posterior [26].
- **Support Vector Machine (SVM)** is a popular and binary classifier that is proved to be effective on a huge category of classification problems. It tries to find a hyperplane that divides training samples into their classes with maximum margin [27]. In our problem we use SVM with RBF kernel which can handle most nonlinear binary classifications better.
- **Random Forest (RF)** is an ensemble learning method for decision trees by building a set of decision trees with random subsets of attributes and bagging them for classification results [28].
- **Deep Neural Network (DNN)**. The proposed model in this paper. We use a 4-layer DNN with a softmax classifier for the detection task. We also evaluate the influence of using different size of networks.

Measures. For a fully investigation of proposed methods, we consider the following aspects:

- **Performance**. To evaluate the detection performance of our method, we evaluate the results with Accuracy and F1-score. By dividing user samples as stressed (positive) and non-stressed (negative) ones, detection results of testing data can be categorized into the following classes:
    - **True Positive (TP)**: stressed user sample correctly detected (true) as stressed (positive).
    - **False Negative (FN)**: stressed user sample incorrectly determined (false) as non-stressed (negative).
    - **False Positive (FP)**: non-stressed user sample incorrectly detected (false) as stressed (positive).
    - **True Negative (TN)**: non-stressed user sample correctly determined (true) as non-stressed (negative).

  Accuracy is the proportion of correct prediction or true results among testing samples. More formally it is given by

  $$Accuracy = \frac{TP+TN}{TP+FN+FP+TN}. \qquad (7)$$

  F1-score, on the other hand, considers both the precision $p = \frac{TP}{TP+FP}$ and recall $r = \frac{TP}{TP+FN}$ of the result, which is given by

  $$F_1 = 2 \cdot \frac{p \cdot r}{p+r} = \frac{2TP}{2TP+FP+FN}. \qquad (8)$$

- **Efficiency**. We evaluate efficiency of the methods by comparing the CPU time of training each model. All experiments are conducted on a server running a Windows 7, with Intel(R) Core(TM) i7-3930K CPU @ 3.20GHz (12 CPUs) and 32 GB RAM. For DNN, we add up both pre-training time and fine-tuning for a fair comparison.

5.2 Detection Performance

To evaluate the effectiveness of the proposed model, we first perform a fully test against large-scale DB1 from Sina Weibo. We consider working with statistical attributes and content attributes extracted by proposed CNN with CAE from cross-modal tweets data of a week respectively, and then using both of them together. For the pooling method, we also test the all three methods: max pooling, mean-over-instance (MOI) pooling and mean-over-time (MOT) pooling. For comprehensive comparisons, we test SVM, RF, NB as well as the proposed DNN as classifiers in this experiment. For this experiment, a 4-layer DNN is used.

Table 4 demonstrates the results of extensive experiments. Regarding different classifiers, SVM gets an accuracy of 75.62% and F1-score 0.8341 using both attributes together and max pooling. RF gets similar results where the accuracy is 76.75% and F1-score is 0.8341. NB does not work well with statistical attributes. It gets its best result working with content based attribute alone using MOI pooling. The proposed DNN classifier reaches the overall best performance with an accuracy of 78.57% and F1-score of 0.8443. Classification using two types of attributes together with MOT pooling outperforms all the baselines. It achieves a ~3% improvement over SVM and ~2% improvement over RF. When it works with the single type of attribute or other pooling methods it also get competitive results.

5.3 Factor Contribution Analysis

Impact of content and statistical attributes: Table 4 also reveals the impact of two types of attributes. With solely statistical or content attribute, all classifiers get fair results around accuracy of 70%. While both types of attributes are used, there is a growth of about 5%. Trend of F1-score is similar that using both types of attributes provides a better result. These results show the effectiveness of combining both classes of attributes, which also prove that the proposed model is reliable for user-level stress detection.

Impact of pooling methods: Comparison results using max pooling, MOI pooling and MOT pooling are also shown in Table 4. We can see that MOT pooling gets an obvious better result working with DNN. When SVM or RF is considered, all three methods get similar results and max pooling is fractionally ahead in all three pooling methods. In summary, MOT is a better choice for high performance detection.

Impact of different modalities in content attributes: Tweets content come with multiple modalities. To evaluate the contribution of each data modality, we conduct experiments with different combination of attributes. Since text is the necessary part of a tweet, we test using solely text attributes, using combination of text and visual attributes, using combination of text and social attributes, as well as using all attributes.

As shown in Table 5, we report predict performance of using content attributes (composed with only the named attributes in Table 5) alone as well as combining with statistical attributes. Using just text attribute gains rather high performance. Simply combining visual or social attributes even reduces the result, especially the social attributes. This trend is even more obvious when both types of attributes (content and statistical) are used. Nevertheless, using all attributes together outperforms using only text attributes. Highest detection performance is observed when using all attribute and working with both types of attributes.

Table 5. Comparison of results using different attributes.

		Text	Text + visual	Text + Social	All
content	Accuracy	0.7147	0.7187	0.7090	**0.7227**
	F1-score	0.8031	0.8054	0.7993	**0.8072**
both*	Accuracy	0.7613	0.7610	0.7228	**0.7849**
	F1-score	0.8294	0.8265	0.8011	**0.8443**

*both content and statistical attributes

Impact of scale of data. Model learning of the proposed CNN attributes extraction model with CAE is a key link of the whole framework. The model is trained in unsupervised scheme and takes advantage of large-scale unlabeled data. DNN classifier model also utilizes large-scale training data. We investigate the impact of data scale on training the network.

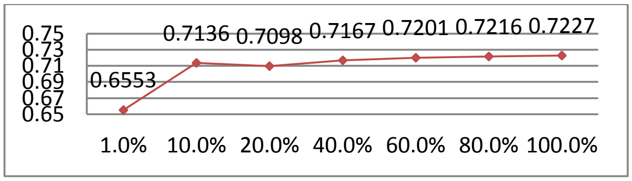

Figure 9. Influence of data scale in training the DNN model, measured in accuracy.

We measure the overall quality by final detection performance. In order to focus the discussion on neural network model, we evaluate with all attributes and only use content attributes. Figure 9 shows the trend of detection performance with different proportion of training data. It is clear that with too small scale of training data, i.e. 1% of all training data, DNN fails to achieve meaningful performance. When it comes on the order of 10%, the model can get significant result around 71%. When more data are involved, it get rise and eventually reach 72.27%. The amplitude is small but it continues to grow. We can see the advantage of using larger training set from the result.

Impact of size of network. Size of network is a critical issue in setting up DNN model. Shallow networks result in trivial model that cannot catch any underlying correlation in data, whereas too deep networks lead to over-complex model which is difficult to tune and may suffer from problems like over-fitting. To choose an appropriate DNN model for classification, we test DNN with different number of layers.

Table 6 summarizes the experiment results. It is clear that 2-layer is not enough for the model to get a satisfactory result. 3-layer model improve significantly while 4-layer model reaches the peak. 5-layer model does not get better result. This is mainly due to the network is too large that it cannot be tuned to a good local minimum with available data and within a feasible training time.

Table 6. Comparison of results using different number of layers in DNN.

	Accuracy	F1-score
DNN-2	0.6843	0.7926
DNN-3	0.7816	0.8423
DNN-4	**0.7857**	**0.8443**
DNN-5	0.7762	0.8386

5.4 Model Efficiency

Table 7. Comparison of results using more classifiers.

	SVM	RF	NB	DNN
CPU time (s)	3.59× 10^4	4.59× 10^3	3.10	1.93× 10^4

For the classification models aforementioned, we also consider their efficiency performance. Though the training of model can be done offline, efficiency is still a considerable factor for evaluating an algorithm. For DNN model, we sum up both pre-training phase and finetuning phase. Table 7 lists the CPU time of each model to train with all labeled data. SVM takes around 10 hours to finish all the training which is the longest. Random Forest is much faster with one and more hour, which is a good alternative since it gets competitive result. Training NB is fast but the detection performance is poor. Training DNN takes around 5 hours which is still reasonable while it get the best detection performance results.

5.5 Results on Other Datasets

We further evaluate our model on other datasets DB2-DB4 to show that our model is a universal model. For this part of experiments, we use statistical attributes together with content attributes using MOT pooling, and with 4-layer DNN model.

DB2 from Sina Weibo with PSTR label. We use a matured model trained with large scale Sina Weibo dataset, and then test it against another set of subject independently sampled from Sina Weibo. For the test set, we collect weekly tweets from the users that have shared the score of a psychological stress scale with 50 items via Sina Weibo. Detection result shows that the test accuracy is 74.13% and f1-score is 0.7778, which approves that the overall model is consistent and the sentence pattern based ground truth labeling method is reliable.

DB3 from Tencent Weibo. We test on data collected from another major Chinese Micro-blog platform. For this test, we use the attribute extractor trained with large scale Sina Weibo dataset and only finetune the network with Twitter dataset in 5-fold. The accuracy is 76.78% and f1-score is 0.7915 which demonstrate the capability of the proposed model.

DB4 from Twitter. We also test against the twitter dataset. We still use the attribute extractor trained with large scale Sina Weibo dataset and only finetune the network with Twitter dataset in 5-fold. The accuracy is 67.43% and f1-score is 0.7224. One reason for this modest result is that users in Twitter dataset and Sina Weibo dataset come from different language and culture background. Another factor could be that the scale of this dataset is rather small. Subjects in the Twitter dataset are on the order of 10% of large scale Sina Weibo dataset. We look into the collected data and find that, by coincidence, all tweets in this dataset have no social activity. We suggest this is also a cause of the unsatisfactory result.

6. CONCLUSION

In this paper, we present a user-level psychological stress detection from users' weekly micro-blog data. First we use the sentence patterns like "I feel stressed" to collect the ground truth labeled micro-blog data in week unit. Then we define a set of low-level content attributes from single tweet's text, images and social interactions. We also present a variety of statistical attributes like behavioral attributes, social engagement and linguistic style attributes from users' weekly tweet postings. A convolutional neural network with cross autoencoders is designed to aggregate weekly low-level content attributes and generate user-scope attributes. Finally we propose a deep neural network model to further learn higher-level attributes in user-scope and predict users' stress. In our proposed method, the user-scope attribute extractor and classification model forms a uniform deep architecture which bridges the gap between each single tweet and user's psychological stress state. We test the model on four different datasets from major micro-blog platforms with different scales and ground truth labeling methods, and deeply discuss the influence of model parameters on experimental results. The results show that the proposed model is effective and efficient on detecting psychological stress from micro-blog data.

In our future work, we plan to investigate using the social media data to develop tools for large-scale public stress states tracking. We are also interested in developing such tools in personal mobile devices to put the research output into practice.

7. REFERENCES

[1] Kasl, S. V. (1984). Stress and health. Annual review of public health, 5(1), 319-341.

[2] Hammen, C. (2005). Stress and depression. Annu. Rev. Clin. Psychol., 1, 293-319.

[3] Cohen, S., Kamarck, T., & Mermelstein, R. (1983). A global measure of perceived stress. Journal of health and social behavior, 385-396.

[4] Kanner, A. D., Coyne, J. C., Schaefer, C., & Lazarus, R. S. (1981). Comparison of two modes of stress measurement: Daily hassles and uplifts versus major life events. Journal of behavioral medicine, 4(1), 1-39.

[5] Hetz, C., Martinon, F., Rodriguez, D., & Glimcher, L. H. (2011). The unfolded protein response: integrating stress signals through the stress sensor IRE1α. Physiological Reviews, 91(4), 1219-1243.

[6] Healey, J. A., & Picard, R. W. (2005). Detecting stress during real-world driving tasks using physiological sensors. Intelligent Transportation Systems, IEEE Transactions on, 6(2), 156-166.

[7] Lu, H., Frauendorfer, D., Rabbi, M., Mast, M. S., Chittaranjan, G. T., Campbell, A. T., ... & Choudhury, T. (2012, September). StressSense: Detecting stress in unconstrained acoustic environments using smartphones. In Proceedings of the 2012 ACM Conference on Ubiquitous Computing (pp. 351-360). ACM.

[8] Paredes, P., Sun, D., & Canny, J. (2013, May). Sensor-less sensing for affective computing and stress management technology. In Pervasive Computing Technologies for Healthcare (PervasiveHealth), 2013 7th International Conference on (pp. 459-463). IEEE.

[9] Sadilek, A., Kautz, H. A., & Silenzio, V. (2012, June). Modeling Spread of Disease from Social Interactions. In ICWSM.

[10] Paul, M. J., & Dredze, M. (2011, July). You are what you Tweet: Analyzing Twitter for public health. In ICWSM.

[11] De Choudhury, M., Gamon, M., Counts, S., & Horvitz, E. (2013, July). Predicting depression via social media. In AAAI Conference on Weblogs and Social Media.

[12] Hinton, G. E., & Salakhutdinov, R. R. (2006). Reducing the dimensionality of data with neural networks. *Science, 313*(5786), 504-507..

[13] Hinton, G. E. (2007). Learning multiple layers of representation. *Trends in cognitive sciences, 11*(10), 428-434.

[14] Bengio, Y. (2009). Learning deep architectures for AI. *Foundations and trends® in Machine Learning*, 2(1), 1-127.

[15] Ngiam, J., Khosla, A., Kim, M., Nam, J., Lee, H., & Ng, A. Y. (2011). Multimodal deep learning. In *Proceedings of the 28th International Conference on Machine Learning (ICML-11)* (pp. 689-696) .

[16] Srivastava, N., & Salakhutdinov, R. (2012). Multimodal Learning with Deep Boltzmann Machines. In *NIPS* (pp. 2231-2239).

[17] H. Lin, J. Jia, Q. Guo, Y. Xue, J. Huang, L. Cai, L. Feng. Psychological Stress Detection from Cross-media Microblog Data Using Deep Sparse Neural Network. In proceedings of IEEE International Conference on Multimedia & Expo, 2014.

[18] [Alex2012] Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). ImageNet Classification with Deep Convolutional Neural Networks. In *NIPS* (Vol. 1, No. 2, p. 4).

[19] Kamvar, S. D., & Harris, J. (2011, February). We feel fine and searching the emotional web. In Proceedings of the fourth ACM international conference on Web search and data mining (pp. 117-126). ACM.

[20] Tausczik, Y. R., & Pennebaker, J. W. (2010). The psychological meaning of words: LIWC and computerized text analysis methods. Journal of Language and Social Psychology, 29(1), 24-54.

[21] Rui Gao, Bibo Hao, He Li, Yusong Gao, Tingshao Zhu. Developing Simplified Chinese Psychological Linguistic Analysis Dictionary for Microblog. 2013 International Conference on Brain & Health Informatics (BHI'13). Oct. 2013. Maebashi, Japan.

[22] X. Wang, J. Jia, J. Yin, and L. Cai. Interpretable aesthetic features for a_ective image classi_cation. In ICIP, pages 3230-3234, Melbourne, Australia, 2013.

[23] Wang, X. H., Jia, J., Liao, H. Y., & Cai, L. H. (2012). Affective image colorization. Journal of Computer Science and Technology, 27(6), 1119-1128.

[24] Vincent, P., Larochelle, H., Bengio, Y., & Manzagol, P. A. (2008, July). Extracting and composing robust features with denoising autoencoders. In *Proceedings of the 25th international conference on Machine learning* (pp. 1096-1103). ACM.

[25] LeCun, Y., & Bengio, Y. (1995). Convolutional networks for images, speech, and time series. *The handbook of brain theory and neural networks, 3361*.

[26] Ng, A. Y., & Jordan, M. I. (2002). On discriminative vs. generative classifiers: A comparison of logistic regression and naive bayes. *Advances in neural information processing systems*, 2, 841-848.

[27] Chang, C. C., & Lin, C. J. (2011). LIBSVM: a library for support vector machines. *ACM Transactions on Intelligent Systems and Technology (TIST), 2*(3), 27.

[28] Svetnik, V., Liaw, A., Tong, C., Culberson, J. C., Sheridan, R. P., & Feuston, B. P. (2003). Random forest: a classification and regression tool for compound classification and QSAR modeling. *Journal of chemical information and computer sciences, 43*(6), 1947-1958.