PSYCHOLOGICAL STRESS DETECTION FROM CROSS-MEDIA MICROBLOG DATA USING DEEP SPARSE NEURAL NETWORK

Huijie Lin[1,2], *Jia Jia*[1,2], *Quan Guo*[3], *Yuanyuan Xue*[1], *Jie Huang*[1], *Lianhong Cai*[1,2], *Ling Feng*[1]

[1] Department of Computer Science and Technology, Tsinghua University, Beijing, China
[2] TNList and Key Laboratory of Pervasive Computing, Ministry of Education
[3] Machine Intelligence Laboratory, College of Computer Science, Sichuan University, Chengdu, China
linhuijie@gmail.com, jjia@tsinghua.edu.cn, guoquanscu@gmail.com, xue-yy12@mails.tsinghua.edu.cn,
qazwsxed_3372@qq.com, clh-dcs@tsinghua.edu.cn, fengling@tsinghua.edu.cn

ABSTRACT

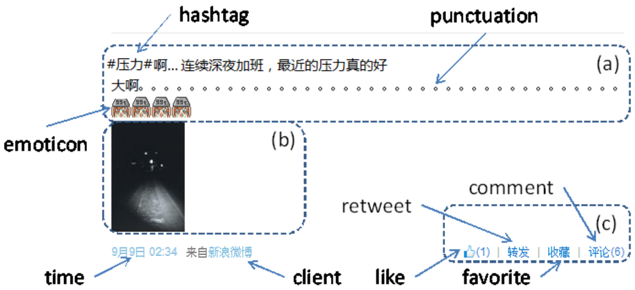
Long-term stress may lead to many severe physical and mental problems. Traditional psychological stress detection usually relies on the active individual participation, which makes the detection labor-consuming, time-costing and hysteretic. With the rapid development of social networks, people become more and more willing to share moods via microblog platforms. In this paper, we propose an automatic stress detection method from cross-media microblog data. We construct a three-level framework to formulate the problem. We first obtain a set of low-level features from the tweets. Then we define and extract middle-level representations based on psychological and art theories: linguistic attributes from tweets' texts, visual attributes from tweets' images, and social attributes from tweets' comments, retweets and favorites. Finally, a Deep Sparse Neural Network is designed to learn the stress categories incorporating the cross-media attributes. Experiment results show that the proposed method is effective and efficient on detecting psychological stress from microblog data.

Index Terms— Stress detection, cross-media, deep learning, microblog

1. INTRODUCTION

1.1. Motivation

People's psychological stress is severely rising. Nowadays, many people feel increasingly stressed under the rapid pace of modern life. According to a worldwide survey conducted by the Regus Business Tracker in 2009[1], over half of the business population (53.8%) have experienced an appreciable rise in stress over the last two years. Though stress can be a positive aspect in our daily life, excessive amount of stress can be rather harmful to physical and mental health. Long-term stress may lead to many severe physical and mental problems, such as clinical depressions, insomnia and even suicide. According to China's Center for Disease Control and Prevention[2], suicide has become the top cause of death among Chinese youth, and excessive stress is considered to be a major factor of suicide. All these reveal that the rapid increase of stress has become a great challenge to human health and life quality.



#压力#啊... 连续深夜加班，最近的压力真的好大啊。。。。。。。。。。。。。。。。。。。。。。。。。。。。。。。。。。。。。。

9月9日 02:34 来自新浪微博

(1) | 转发 | 收藏 | 评论(6)

Fig. 1. A tweet instance with dash line marked out multimedia areas. (a) text area; (b) image area; (c) user interaction area. The hashtag (stress label) word given by the user himself is shown between two "#" marks.

Thus, there is significant importance to detect stress before it turns into severe problems. While traditional psychological stress detection is mainly based on face-to-face interviews conducted by psychologists, it is usually labor-consuming, time-costing and hysteretic.

Microblog has become a popular platform for people to express themselves. Nowadays, with the rapid development of social networks, people are more and more willing to use microblog to express their moods. As reported by Sina weibo (the largest microblog platform in China), the number of weibo users has reached 500 million[3]. And according to the report given by Chinese Academy of Social Sciences [1], self-expression is the first main usage of microblog (74.3%).

[1] http://www.regus.com.au/images/Regus%20Business%20Tracker%20Report%20-%20April%202010_tcm78-21557.pdf

[2] The Week News, http://theweek.com/article/index/ 252199/the-rise-of-youth-suicide-in-china

[3] Xinhuanet News, http://news.xinhuanet.com/ newmedia/2013-02/21/c_124369896.htm

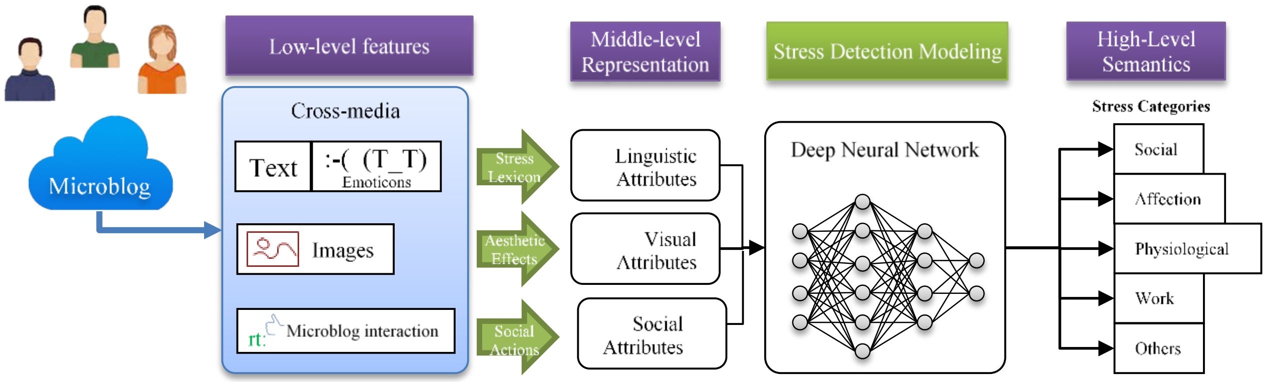


Fig. 2. The framework of stress detection from cross-media microblog data.

This makes the detection of users' psychological stress through their tweets on microblog feasible.

1.2. Related work

Due to the challenges of traditional stress detection, many efforts have been devoted to detecting psychological stress more efficiently and timely recent years. Most of them [2, 3] focus on detecting stress in real-time via body-worn sensors and mobile devices. The equipment is used to gather physiological parameters and daily activity data. Andrew Raij .etc. [2] designed a mobile phone based system, named mStress, using 6 extra wireless sensors to collect physiological data to support real-time detection of stress. Hong Lu. etc. [3] proposed StressSense for unobtrusively recognizing stress from smartphone-recorded conversational voice data. However, such applications rely on additional sensors or devices to collect one's real-life data. It makes stress detection invasive to normal life, and can't be used widely in more people.

There are also some researchers turning to non-invasive ways to automatically detect psychological stress from social networks. [4] and [5] presented methods to detect psychological stress from forum posts and microblog tweets respectively. These methods are mainly based on text data in the social networks, while other equally important content, like images and social actions are ignored.

1.3. Our work

In this paper, we propose a novel method of detecting psychological stress using cross-media microblog data. Fig.1 shows a typical example of stress expression by microblog tweet. We can indicate that the user feels stressed from his/her work. Besides the main text, the long series of punctuation marks and emoticons, the gloomy image and comments from users are also indicators of such stress in this tweet. Based on the analyses of the main composition of a microblog tweet, we first obtain the low-level features from each component of the tweet: texts, images, and social actions like comments, retweets and favorites. Then according to related social psychological and art theories, we define and extract middle-level representations in three aspects: linguistic attributes from tweet's texts, visual attributes from tweet's images and social attributes from social actions. Finally, to maximize the contribution of the cross-media attributes, we design a Deep Sparse Neural Network to learn the stress detection model. We conduct experiments on large-scale Sina weibo data. Using real stress labels given by users themselves as the ground-truth, the results indicate the effectiveness and efficiency of the proposed method.

The main contributions of our work are:

1） We propose a novel method of detecting psychological stress from microblog utilizing cross-media microblog data;

2） We construct a three-level framework to formulate the problem, and propose a middle-level representation according to psychological and art theories, which can narrow the gap between low-level cross-media features and high-level stress semantics;

3） We design a Deep Neural Sparse Network based classification model to solve the problem of sparse in cross-media data.

2. FRAMEWORK OVERVIEW

An overview of our three-level stress detection framework is shown in Fig.2. Tweets from the microblog usually consist of the following parts: 1) text content with no more than 140 words; 2) images uploaded together with the text message; 3) social interactions with comments, favorites and retweets within friends. These three parts together make tweets of microblog to be actually cross-media data.

To incorporate different features from the cross-media microblog data to enhance stress detection, we obtain a set of low-level features from the microblog tweets at the first stage, such as key words from texts, colors from images, and the numbers of social actions from microblog interactions. Then we define and extract stress-related middle-level representations from the low-level features of tweets, namely linguistic attributes from tweets' texts, visual attributes from tweets' images, and social attributes from tweets' comments, retweets and favorites. The definition and extraction of the middle-level representations is based on psychological principles and art theories in [6, 7]. Finally, a deep neural network is implemented to learn the stress detection model from the extracted middle-level representations.

According to the psychological lexicon LIWC2007 [8], in this paper we define 5 different psychological stress categories as our output (affection, work, social, physiological, and others). The "others" categories represents having stress caused by other reasons. Given a tweet ***T***, our task is to detect whether ***T*** reflect its author with psychological stress and which category the stress belongs to.

3. MIDDLE-LEVEL REPRESENTATION DEFINITION

To narrow the semantic gap between low-level features and high-level stress semantics, we define and extract several middle-level representations from tweets' low level features based on previous psychological principles and art theories in [6, 7]. The definitions are as follows:

1) Linguistic Attributes:

Three stress-related linguistic lexicons are constructed for stress detection, including stress-category lexicon, negative emotion lexicon and negation lexicon. The stress-category lexicon is built based on the "Simplified Chinese Language Inquiry and Word Count Dictionary" [9]. It contains 1307 words, which are categorized into four typical types, i.e., physiological, work-related, social and affection-related stress. The negative emotion lexicon is composed of negative emotion words such as "sad, distressing" to indicate the bad emotion under some stress. The negation lexicon is proposed here to judge whether a sentence have the negative meaning.

Based on the stress related lexicons, we define the text content related features as the middle-level linguistic attributes:

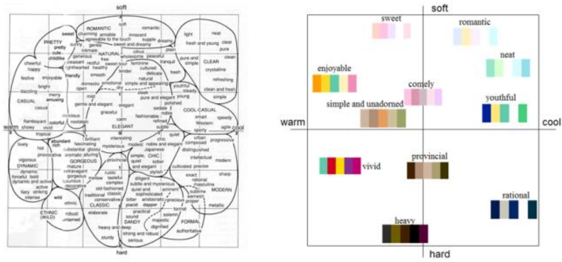
- **Linguistic Association between Stress Category and Negative Emotion Words** (1 dimension). To find out associations among the stress-related words, we apply a graph-based Chinese parser to generate a word-association tree from the tweet-text-content. Each node of the tree denotes a word token, and each edge between two nodes denotes a word association. If there exists a path between a stress-category-related word node and a negative emotion word node and no negation words in between, a stress in the corresponding category is detected. $L_{category}$ is denoted as the number of edges between the stress-category word and negative word in the path. We use $L_{category}$ to determine the stress existence in a tweet.
- **Number of Negative Emotion Words** (1 dimension). The more negative emotion words are used in tweets, the more possibility of stress can be detected. We calculate the number of negative emotion words to improve the accuracy of the stress detection.
- **Positive and Negative Emoticons** (2 dimensions). Emoticons are often used to express users' emotional inclination in tweets because of its iconicity. We use this attribute to distinguish users' emotional leaning in a tweet. Sina weibo platform provide 129 emoticons which contain positive and negative emoticons. We give each of them a weight ranged from -4 to 4 to denote its strength of positive and negative sentiment. The higher the weight assigned, the higher possibility of stress can be determined.
- **Punctuation Marks and Associated Emotion Words** (4 dimensions**).** Punctuation plays a vital, though subtle, role in showing a character's emotion, .i.e. exclamation mark is often used to strengthen mood, while question mark is often used to imply confused mood. 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.

Thus, we get 8-dimensional vector to denote the linguistic attributes from the tweets' content.

2) Visual Attributes:

Based on previous work on affective image classification [6] and color psychology theories [10], 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 [10]. Fig. 3. demonstrates how the colors are correlated to emotions.



(a) (b)

Fig. 3. (a) Emotion categories in two-dimensional (war-cool, soft-hard) art space; (b) Examples of five-color theme and their corresponding emotion categories in the same space.

- **Saturation** (2 dimensions): the mean value of saturation and its contrast. It describes the colorfulness and the differences of an image. Psychological experiments in [11] find out that people under stress and anxiety prefer lower saturation than normal states, revealing the correlation between stress and saturation of images.

- **Brightness** (2 dimensions): the mean value of brightness and its contrast. It illustrates the perception elicited by the luminance and the related differences of an image (e.g., low brightness makes people feel negative, while high brightness elicits mainly positive emotional associations).
- **Warm or cool color** (1 dimension): ratio of cool colors with hue ([0-360]) in the HSV space between 30 and 110. It is defined based on human feelings elicited by colors. Warm colors like red, orange and yellow often evoke feelings of happiness, optimism and energy, Cool colors like green, blue, and purple are usually calming and soothing but can also express sadness sometimes.
- **Clear or dull color** (1 dimension): ratio of colors with brightness ([0-1]) and saturation less than 0.6. It depicts the clear or dull feeling of an image, determined by both brightness and saturation. An image with low brightness and saturation usually feels dull, otherwise clear.

To extract the five-color theme feature, we use the method described in [6]. Saturation, brightness and warm or cool color features of an image can be easily calculated in HSV (Hue, Saturation and Value) space, which is the cylindrical-coordinate representations of RGBs. The clear or dull color feature can be derived from saturation and brightness.

Thus, based on the psychological studies and color theories, we finally get a 21 dimensional middle-level representation from the low-level RGBs of an image.

3) Social Attributes:

Besides the text content and image content of a tweet, some additional features like comments, retweeting and favorites can also imply one's stress state to some degree. We define a tweet's **social attention degree** based on these additional features into social attributes.

An apparently stressful tweet may attract more attention from friends. The number of comments, retweets and favorites reveals how much attention a tweet attracts. To find out the changes of attention degree of one's tweet, we first calculate the mean M_i and variance V_i of the number of one's tweets' comments, retweets and favorites respectively. Then with the number of a tweet's comments, retweets and favorites N_i, we define the variation characteristic of social interaction VC_i as:

$$VC_i = (N_i - M_i)^2/V_i \tag{1}$$

Thus, we get a 3-dimensional vector to represent the social attributes of a tweet.

4. MODEL AND LEARNING

4.1. Architecture

In order to handle microblog data from multiple modalities in a cross-modal setting, we propose to use a deep sparse neural network. In our stress detection task, representations for features from each single modal are extracted using proposed method in previous section. A joint representation layer follows after the single modal representation layer to fuse information from different modalities. We further stack another auto-encoder to extract high-level representation for microblog data. On the top of the hierarchy, we employ 5 binary classifiers that each indicates one category of stress. The overall architecture of the network is shown in Fig. 4.

We assigned 8 visible neurons for content representation, 18 visible neurons for image representation and 5 visible neurons for social representation corresponding to the representations extracted from content, image and social features. The model also contains 100 hidden neurons for joint representation and another 100 for high-level representation. The 5 classifiers work independently.

We adopt Softplus activation function which is given by:

$$\text{Softplus}(x)=\log(1+e^{x}). \tag{2}$$

Softplus activation function is a smooth alternative to hard zero rectifier, which prevent hard zero to hurt back propagation optimization while creating an easy-to-train sparse representations [12].

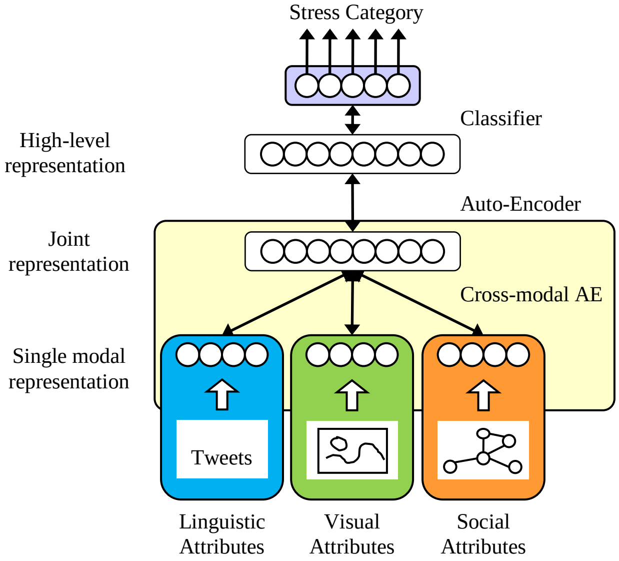


Fig. 4. The architecture of deep network for cross-media stress detection

4.2. Cross-media Auto-Encoder

Though microblog is an open platform, we may not have access to complete features of tweets all the time due to the fact like user did not provide any image for a tweet by privacy issues. This is a typical cross-media setting. It has been reported that deep neural networks can be used to learn cross-media or share representations in speech video [13] or images with text tag [14].

We propose a Cross-media Auto-Encoder (CAE) to learn the joint representation using Denoising Auto-Encoder (DAE) style learning [15]. Fig.5 shows a sketch of a CAE. With training data containing all three modality information, we manually disable image representation and/or social representation when training the auto-encoder, while requiring it to reconstruct all three. In this part of training, thought the missing part of data did not contribute as input, they enforce the network to reconstruct them even if they are missing.

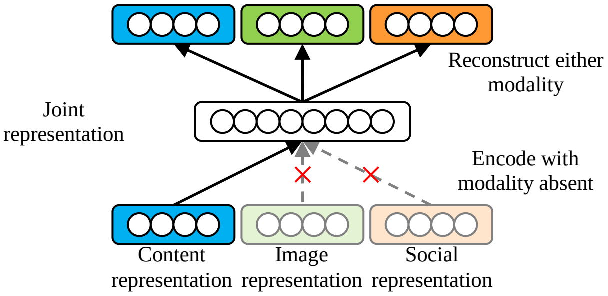


Fig. 5. Cross-modal Auto-Encoder: encode with incomplete data and reconstruct all modalities to learning cross-modal representation in joint layer

4.3. Model learning

Training a CAE is quite different from standard auto-encoders. There is not explicit objective for standard auto-encoders to learn feature across modalities and they tend to learn feature within each single modality with few correlation connection between multiple modalities [13]. We train the CAE with a cropped set of data $\{\bar{x}_c, \bar{x}_{\mathrm{i}}, \bar{x}_s\}$ that input from one or two modalities are absent while require it to reconstruct data from three.

$$\begin{cases} a^{(2)} = f\left(w_c^{(1)}\bar{x}_c + w_i^{(1)}\bar{x}_i + w_s^{(1)}\bar{x}_s + b^{(1)}\right) \\ \quad (\tilde{x}_c, \tilde{x}_i, \tilde{x}_s) = f\left(w^{(2)}a^{(2)} + b^{(2)}\right) \end{cases}. \tag{3}$$

where $a^{(2)}$ is the joint representation, $w_c^{(1)}$, $w_i^{(1)}$, $w_s^{(1)}$, $w^{(2)}$ are the connection weights of encoders and decoder for different modalities, $f(\cdot)$ is the Softplus activation function and $\{\tilde{x}_c, \tilde{x}_i, \tilde{x}_s\}$ are the reconstruction results. In practice, to fully utilize available data, we augment the training dataset with four cropping setting: 1) content + image + social, 2) content + image, 3) content + social and 4) content only, with input from other modality being set to 0. For all these four sets of input we require our model to reconstruct data from content, image and social and evaluate with the energy function given by

$$\min \frac{1}{4m} \sum_{mo\in\{c,i,s\}} \sum_{i=1}^{m} \frac{1}{2} \left\|\tilde{x}_{\mathrm{mo}}^{(i)} - x_{mo}^{(i)}\right\|^2 + \frac{\lambda}{2} \sum_{l=1}^{2} \sum_{i=1}^{s_1} \sum_{j=1}^{s_2} \left(w_{ij}^{(l)}\right)^2 + \beta \sum_{j=1}^{s_2} KL(\rho||\rho_j). \tag{4}$$

λ and β are weight decay and sparse penalty while ρ is the sparse parameter. $KL(\rho||\rho_j)$ is the Kullback-Leibler (KL) divergence to enforce sparsity in joint representation.

The high-level representation is then trained with outputting joint representation with all four cropping settings and fine-tuning is in supervised manner with a Softmax classifier.

5. EVALUATION

5.1. Experiment setup

Dataset. We crawl about 600 million tweets from the biggest Chinese microblog site, Sina weibo, from 2009.10 to 2012.10. Many of these tweets contain a variety of user-labeled hashtags, with keywords surrounded by “#”. These hashtags are usually used to express users’ opinion on some topics or there psychological states. We use these user-labeled hashtags as our ground-truth, and collect the tweets with hashtags only belonging to one of our five stress categories (affection, work, social, physiological, and others) by Simplified Chinese Language Inquiry and Word Count Dictionary. We also crawl tweets with no stressed hashtags. We finally get 57785 hashtag labeled tweets as ground-truth in total. The details of the dataset are shown in Table 1.

To test the reliability of the ground-truth labeling using hashtags, we randomly select and manually label 500 tweets from each category. Each tweet is manually labeled by 3 persons, and the hashtag label is treated as correct only if at least 2 persons agree to label the tweet into the same category as the hashtag label. The experiment results show that 95% of hashtag labels are correct, indicating that the hashtag labeled data is quite reliable.

Table 1. Tweet dataset for stress detection

Categories	Number of Tweets	Number of Images
affection	3634	681
Work	3966	868
social	5747	1160
Physiological	13973	2017
others	14543	3186
None(not stressed)	14931	6014
Total	57785	13926

Experiments. We design 3 experimental settings to demonstrate the performance of our model dealing with cross-media data.

- Detecting stress using only tweets’ text information, which means using only linguistic attributes in testing model.
- Detecting stress using visual attributes in additional to linguistic attributes in testing model.
- Detecting stress using tweets’ cross-media information, which means using all the proposed attributes in testing model.

For each of the setting, we compare our model (Deep Sparse Neural Network with CAE) with Support Vector Machine (SVM), shallow Artificial Neural Network (ANN) and DNN with Stacked Auto-Encoder (SAE). For both SAE and CAE we set the parameter of model to $\lambda = 0.001$, $\beta = 4$ and $\rho = 0.25$. For all experiments, we involve a 5-

fold cross validation to have all data tested and measure the performance by accuracy.

5.3. Results and Discussions

Table 1. Performance of stress detection (using only linguistic attributes or cross-media attributes).

F1-measure (%)	**Only Linguistic Attributes**			**Content + visual**			
	SVM	**ANN**	**SAE**	**SVM**	**ANN**	**SAE**	**CAE**
Affection	0	0	0	0	20.34	35.02	33.97
Work	0	4.52	1.48	56.09	53.68	53.70	54.39
Social	42.44	39.53	43.25	51.58	44.61	48.24	49.80
Physiological	33.18	34.97	39.11	70.08	68.97	70.24	71.18
Other	46.83	47.63	49.02	81.29	78.73	80.70	81.53
None	76.05	75.48	75.92	99.20	99.61	99.89	99.91
Average	33.08	33.69	34.80	59.71	60.99	64.63	65.13
Accuracy	59.74	58.77	59.40	81.73	80.87	81.77	82.29

Experimental results using only text content and using visual feature with different model have been reported in Table 1. It is clearly shown that using visual feature boosts evaluation result overwhelmingly. Different classification model get similar result under same experiment setup while our proposed method, CAE based deep neural network, holds a slim lead on total accuracy and average F1-measure. SVM achieve good accuracy but is subject to data bias issue that fails to work well on categories with few samples.

Table 2. Performance of stress detection with three attributes.

F1-measure (%)	**Content + visual**			**All**		
	SVM	**SAE**	**CAE**	**SVM**	**SAE**	**CAE**
Physiological	78.69	77.46	77.80	79.12	82.30	83.67
Other	58.81	67.39	66.37	64.59	72.49	74.87
None	98.65	99.81	99.84	96.22	99.73	99.81
Average	78.72	81.56	81.34	79.98	84.84	86.12
Accuracy	86.23	87.29	87.27	85.97	89.68	90.55

Table 2 demonstrates results of detecting stress using all proposed attributes. Adding social attribute, comparing to combination of linguistic attributes and visual attributes, detection result can also be improved the by using deep neural network models. Our proposed CAE model achieves also better result among listed methods.

6. CONCLUSION

In this paper, we present a three-level framework for stress detection from cross-media microblog data. By combining a Deep Sparse Neural Network to incorporate different features from cross-media microblog data, the framework is quite feasible and efficient for stress detection. Using this framework, the proposed method can help to automatically detect psychological stress from social networks. In our future work, we plan to investigate the social correlations in psychological stress to further improve the detection performance.

7. ACKNOWLEGEMENT

This work is supported by the National Basic Research Program (973 Program) of China 2012CB316401, and partially supported by 2011CB302201. This work is also partially supported by the National Natural, and Science Foundation of China 61370023 and 61373022. We would also like to thank Microsoft Research Asia-Tsinghua University Joint Laboratory for its support.

8. REFERENCES

[1] “Social psychology blue book,” *Chinese Academy of Social Sciences*, 2011.

[2] A. Raij, et al., "mStress: Supporting Continuous Collection of Objective and Subjective Measures of Psychosocial Stress on Mobile Devices," *In Proc. of ACM Wireless Health 2010, San Diego,* 2010.

[3] H. Lu, et al., "StressSense: Detecting Stress in Unconstrained Acoustic Environments Using Smartphones," *In Proc. of UbiComp 2012, Pittsburgh,* 2012.

[4] S. Saleem, et al., "Automatic Detection of Psychological Distress Indicators and Severity Assessment from Online Forum Posts," *In Proc. of COLING 2012, Mumbai,* pp. 2375-2388, 2012.

[5] Y. Xue, et al., "Towards a micro-blog platform for sensing and easing adolescent psychological pressures," *In Proc. of UbiComp 2013, Zurich,* 2013.

[6] X. Wang, J. Jia, J. Yin and L. Cai, "Interpretable aesthetic features for affective image classification," *In Proc. of ICIP 2013, Melbourne,* pp. 3230-3234, 2013

[7] Richard S. Lazarus, “Stress and emotion: A new synthesis,” *Springer Publishing Company*, 2006.

[8] J. W. Pennebaker, R. J. Booth and M. E. Francis, "LIWC2007: Linguistic inquiry and word count," *liwc. Net*, 2007.

[9] R. Gao, B. Hao, H. Li, Y. Gao and T. Zhu, “Developing Simplified Chinese Psychological Linguistic Analysis Dictionary for Microblog,” *In Proc. of BHI 2013,* Maebashi, 2013.

[10] X. Wang, J. Jia, H. Liao and L. Cai, “Affective image colorization,” *Journal of Computer Science and Technology,* vol. 27, no. 6, pp. 1119-1128, 2012

[11] S. R. Ireland, Y. M. Warren and L. G. Herringer, "Anxiety and color saturation preference." *Perceptual and Motor Skills,* vol. 75, pp. 545-546, 1992

[12] X. Glorot, A. Bordes and Y. Bengio, "Deep Sparse Rectifier Networks," *In Proc. of AI Statistics 2011, Lauderdale,* pp. 315-323, 2011.

[13] J. Ngiam, et al., "Multimodal deep learning." *In Proc. of ICML 2011, Bellevue,* pp. 689-696, 2011.

[14] N. Srivastava and R. Salakhutdinov, "Multimodal learning with deep Boltzmann machines." *In Proc. of NIPS 2012, Lake Tahoe,* pp. 2231-2239, 2012.

[15] P. Vincent, H. Larochelle, Y. Bengio and P. A. Manzagol, "Extracting and composing robust features with denoising autoencoders." *In Proc. of ICML 2008, Helsinki,* pp. 1096-1103, 2008.