ACOUSTICS, CONTENT AND GEO-INFORMATION BASED SENTIMENT PREDICTION FROM LARGE-SCALE NETWORKED VOICE DATA

Zhu Ren[1,2], Jia Jia[1,2], Quan Guo[3], Kuo Zhang[4], Lianhong Cai[1,2]

[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
[4] Sogou Corporation, Beijing, China
bamboo.renzhu@gmail.com, jjia@tsinghua.edu.cn, guoquanscu@gmail.com,
zhangkuo@sogou-inc.com, clh-dcs@tsinghua.edu.cn

ABSTRACT

Sentiment analysis from large-scale networked data attracts increasing attention in recent years. Most previous works on sentiment prediction mainly focus on text or image data. However, voice is the most natural and direct way to express people's sentiments in real-time. With the rapid development of smart phone voice dialogue applications (e.g., Siri and Sogou Voice Assistant), the large-scale networked voice data can help us better quantitatively understand the sentimental world we live in. In this paper, we study the problem of sentiment prediction from large-scale networked voice data. In particular, we first investigate the data observations and underlying sentiment patterns in human-mobile voice communication. Then we propose a deep sparse neural network (DSNN) model to incorporate acoustic features, content information and geo-information to automatically predict sentiments. The effectiveness of the proposed model is verified by the experiments on a real dataset from Sogou Voice Assistant application.

Index Terms— Networked voice data, geo-social information, sentiment prediction, deep neural network

1. INTRODUCTION

Sentiments are playing important roles in daily life. What people feel may affect more or less every aspect of their lives, such as purchase decision making [1] or stock market prices prediction [2]. Previous researches have shown the success of using networked text or image data [3], [4] for sentiment prediction. However, voice is the fastest and the most natural way to communicate with each other [5]. It can express people's sentiments much more vividly and efficiently. Since the last decade, there has been tremendous research on speech emotion recognition, which can benefit lots of fields, e.g., improve the performance of speech recognition systems [6], or enhance the user-friendliness of voice communication applications. Nowadays, with an unprecedented increasing adoption of smart phone voice dialogue applications (e.g., *Siri*[1] and *Sogou Voice Assistant*[2]), people can share voice messages to their friends or make requests to the voice assistant easily. The traditional speech emotion recognition methods cannot deal well with the networked voice data, due to the large amount of speakers and the great diversity of their vocal traits. So analyze and predict the sentiments from large-scale networked voice data still remain a large problem.

Related Work: There have been considerable works in empirical analyses of sentiments based on text or image data from social networks. Some of these analyses are focused on specific events, such as the study about microbloggers' response to the death of Michael Jackson [7] or Flickr users' affect distribution around Thanksgiving [3]. While others further analyze broader social and economic trends, such as the relationship between Twitter mood and both stock market fluctuations [2] and consumer confidence and political opinion [8], [9]. However, owing to the difficulty of voice data acquisition, few have been done in studying sentiments from large-scale networked voice data. Ren et al. [10] have found that the classic machine learning method SVM cannot achieve a remarkable emotion prediction performance from large-scale networked voice dataset, where including too many speakers whose vocal characteristics and recording backgrounds have great distinctions. Recent research tries to apply deep network to large-scale image or speech data, which gains excellent results in classification tasks [11]. Srivastava et al. [12] propose a deep network model that fuses text and image data for classification and information retrieval tasks. These studies show capability of deep networks to learn features over large-scale and diverse data.

Our Approach: In this paper, employing a mobile voice assistant application (*Sogou Voice Assistant*) as the basic of our experiments, we systematically study the problem of

[1] http://www.apple.com/ios/siri/, an intelligent personal assistant and knowledge navigator works as an application for Apple's iOS.
[2] http://yy.sogou.com, a smart phone voice dialogue application developed by Sogou (one of China's internet service providers).

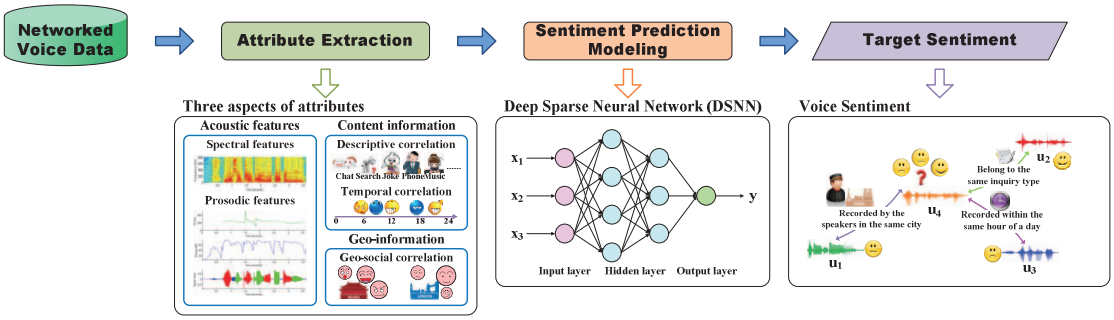

Fig. 1. The framework of our proposed method.

predicting sentiments from networked voice data. Fig. 1 illustrates the conceptual framework of the proposed sentiment prediction method. First, we investigate the data characteristics and underlying sentiment patterns in human-mobile voice communication, which turns out users' sentiments are associated with their personalized characters and environmental context. Hence at the second step, three aspects of features are extracted from all the utterances in the dataset: acoustic features (e.g., pitch, energy, MFCC), content information (e.g., descriptive correlation and temporal correlation) and geo-information (e.g., geo-social correlation). Finally, we propose a deep sparse neural network (DSNN) prediction model to incorporate all the features to automatically predict sentiments. The experimental results demonstrate that the proposed model can achieve better performance than alternative method (e.g., Naive Bayes, Artificial Neural Network). Discussions and analyses of the experimental results rationally verify the contribution of combining the acoustic features with correlation features to improve the performance.

The rest of this paper is organized as follows: Section 2 gives a series of analyses based on the networked voice data and presents our observations. Section 3 formally formulates the problem and describes the feature extraction procedure. Section 4 presents the proposed DSNN model for sentiment prediction. Section 5 carries out the experiments employed to analyze the feature contribution and evaluates the performance of the proposed model. We'd like to show an interesting case study in this section too. Finally, Section 6 summarizes this paper.

2. OBSERVATIONS

2.1. Data collection and observation

We collect a corpus of large-scale networked voice data from *Sogou Voice Assistant*. The raw dataset contains 6,891,298 utterances recorded in Chinese by 405,510 users during year 2013. Each utterance has some basic information (e.g. user ID, record time, GPS location and inquiry type) and its corresponding speech-to-text information provided by Sogou Corporation. We first select the active users who send more than 100 inquiries within 30 days. In fact, the inquiries sent by the selected 10,399 active users occupy 27.04% of the total amount. In order to reduce the impact of noise data, we use the voice data sent by active users for our data observation.

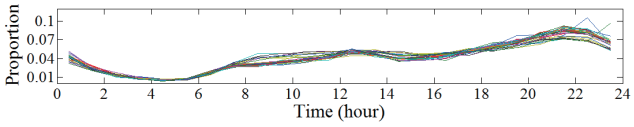

Fig.2. The proportion of inquiry amount in 24 hours of 30 days (each line represents one day).

Then we manage to know ***when*** the users are active to communicate with the voice assistant. Fig. 2 gives the variation trend of inquiry amount during 30 days. As we can see, the 30 lines are almost overlapping with each other, which indicates that the changes of inquiry amount are more relevant to hours than days. We also find some interesting phenomena: 1) the inquiry amount begins a rapid rise from 7:00, when most of the users have already gotten up; 2) it reaches a local peak between 12:00 and 13:00, during people's lunch break; 3) it starts another rise from 18:00 and runs up to the global peak at about 22:00, which because users have some leisure time for entertainments after work.

Furthermore, we take measures to know ***what*** kind of inquiries the users are active in. Fig. 3 shows the proportion of top 20 types with the most inquiry amount out of 70 types. We can see that *Chat* and *Search* are the most frequent inquiry types in human-mobile voice communication, which altogether occupy 56.94% of the total amount. As a matter of fact, users who are chatting with the voice assistant show more sentimental expressions than ones who just want to search for some useful information. We can take advantage of this fact to investigate users' sentiments more deeply.

Finally, we'd like to know ***where*** the most active users

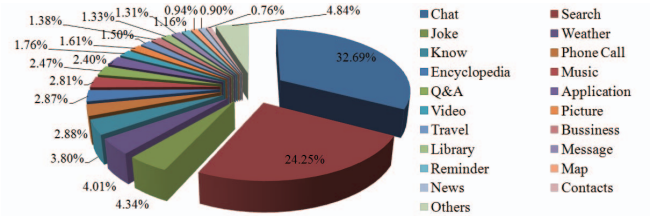

Fig. 3. The proportion of top 20 types with most inquiry amount.

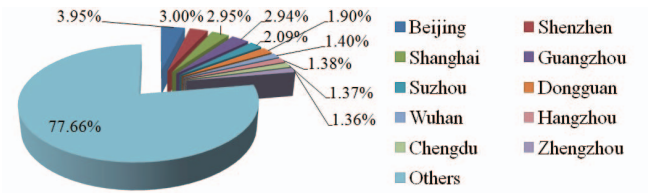

Fig. 4. The proportion of inquiry amount in the top 10 cities.

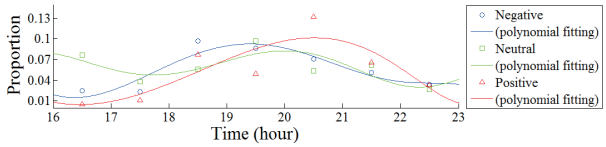

Fig. 5. Temporal correlation observation.

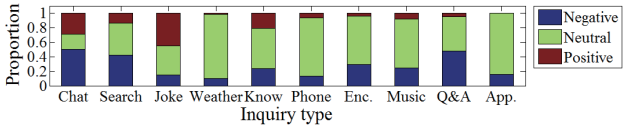

Fig. 6. Descriptive correlation observation.

are living in. We make use of the GPS information to locate a certain city, and count the inquiry amount in each city. The top 10 cities with the most inquiry amount out of 382 cities are listed in Fig. 4. As the political, economic and cultural center of China, Beijing apparently is the top one city with most active users. Due to the high-level of technology development, first-tier cities like Shenzhen, Shanghai and Guangzhou have lots of people using voice assistant as well. The rest of the cities are basically second-tier cities with large population and adequate economic level. It is remarkable to find that almost 22.34% of the inquiries are sent from users who live in these 10 cities, which makes us have plenty of data to further analyze the sentimental differences between cities later.

2.2. Sentiment pattern analysis

Do the users' sentiments have correlations with the ***when***, ***what*** and ***where*** factors? To further investigate the sentiment patterns, we manually label 2,000 utterances recorded by 20 active users respectively from the top 4 cities (100 utterances randomly chosen per user, 5 users randomly chosen per city). We invite three human labelers to annotate each utterance with one of three classes: *Positive*, *Neutral*, and *Negative*. When they have disagreement, they stop and discuss until they have final agreed views. The manually labeled results are regarded as the real sentiments for these utterances.

In the investigation, we focus on the following aspects:

- *Temporal correlation*: probability that a user's sentiment has correlation with the certain time of a day that he/she is experiencing.
- *Descriptive correlation*: probability that a user's sentiment is related to the type of his/her inquiry.
- *Geo-social correlation*: probability that a user's sentiment is associated with the social environment where he/she is living.

Observation on temporal correlation: The proportion of different sentiments' inquiry amount during 16:00 to 23:00 is shown in Fig. 5. It can be easily seen that the change trend of *Positive*, *Neutral*, and *Negative* are distinct from each other. For example, the users who are on their way home after work may get annoyed by the terrible traffic conditions around 17:00, which makes them more likely stay at "Negative" than "Positive". Nevertheless, they will

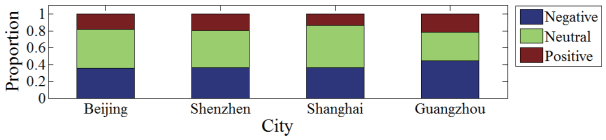

Fig. 7. Geo-social correlation observation.

feel relaxed and happy when they finally get home, which brings "Positive" feelings on most of them.

Observation on descriptive correlation: The distinction of the inquiry content may evoke different kinds of sentiments. Fig. 6 demonstrates the sentiment distributions of the top 10 inquiry types. It can be seen that when the users send some questions for answers (e.g. *Search*, *Encyclopedia*, *Q&A*), they are easy to turn into a bad mood if not getting useful information. However, when they are looking for fun (e.g. *Joke*), they more likely stay at "Positive". Users' sentiments are richer while they communicate with the voice assistant in a natural way (e.g. *Chat*). In contrast, they may tent to stay at "Neutral" when using the voice assistant to achieve certain functions of their cell phones (e.g. *Phone Call*, *Application*).

Observation on geo-social correlation: The Chinese often believe that human beings are shaped by the land around them. As we can see in Fig. 7, the sentiment distributions of different cities have their own traits. During April 2013, "Negative" is the most intense sentiment in Guangzhou, while "Neutral" is the dominant sentiment in other cities. We can also discover that there are more users who stay at "Negative" than "Positive" in all the four cities, which may be because the stressful living environment of first-tier cities makes people more prone to negative sentiment.

Based on the above analyses, we find that these three correlations are all related to users' sentiments, which makes us decide to incoporate acoustic features with temporal, descriptive and geo-social correlations to predict sentiments from large-scale networked voice data.

3. PROBLEM DEFINITION

3.1. Problem formulation

In this section, we present the problem formulation of sentiment prediction from large-scale networked voice data. The networked voice data can be represented as $G' = (V, E)$, where $V = \{u_1, \dots, u_N\}$ is the set of $|V| = \mathrm{N}$ utterances, $E \subset V \times V$ is the set of social correlations among utterances. Notation $e_{ij} \in E$ indicates u_i having a correlation with u_j (e.g. u_i and u_j recorded by the speakers in the same city or recorded within a same time interval). Given this, we can define the problem as follows:

Definition 1. **Sentiments**: The sentiment category of an utterance u_i is denoted as $y_i \in A$, where A is the space covered by three categories (*Positive*, *Neutral*, *Negative*).

Definition 2. **Partially labeled network**: The partially labeled network is denoted as $G = (V^L, V^U, E, \Gamma)$, where V^L and V^U are respectively the set of labeled and unlabeled

utterances with $V^L \cup V^U = V$; E is the correlations between utterances; Γ is an attribute matrix associated with utterances in V with each row corresponding to an utterance, each column representing an attribute and an element x_{ij} denoting the value of the j^{th} attribute of utterance u_i.

Problem. **Learning task**: Given a partially labeled network G, the target is to predict the sentiments of all the unlabeled utterances by learning a predictive function

$$f: G = (V^L, V^U, E, \Gamma) \rightarrow \boldsymbol{\Lambda} \quad (1)$$

where $\boldsymbol{\Lambda} = \{\boldsymbol{S_1}, \dots, \boldsymbol{S_M}\}$ is the set of predicted results, with each $\boldsymbol{S_m}$ belonging to one sentiment category in A; $s_m \in [0,1]$ is the probability score indicating whether the corresponding utterance $u \in V^U$ represents the sentiment m. In our experiments, we classify the utterance into the sentiment category with the maximum probability score.

3.2. Feature Extraction

Based on previous research about emotional speech analysis [13], [14], we extract 113 *Acoustic Features*:

- *Energy* (13): the energy envelop applied with 13 functionals (mean, std, max, min, range, quartile1/2/3, iqr1-2/2-3/1-3, skewness, kurtosis).
- *F0* (13): the fundamental frequency contour, which are extracted using a modified STRAIGHT procedure [15], applied with 13 functionals the same as Energy.
- *MFCC* (26): the mean and standard deviation of mel-frequency cepstral coefficients 1-13.
- *LFPC* (24): the mean and standard deviation of log frequency power coefficients 1-12, which are extracted using the method in [14] with α=1.4.
- *Spectral Centroid (SC)* (13): the spectral centroid contour applied with 13 functionals the same as Energy.
- *Spectral Roll-off (SR)* (13): the spectral roll-off contour applied with 13 functionals the same as Energy.
- *Syllable Duration (SD)* (11): the syllable duration sequence, which is extracted using the method in [16], applied with 11 functionals (mean, std, max, min, range, quartile1/2/3, iqr1-2/2-3/1-3).

All the spectral features (MFCC, LFPC, SC, SR) are extracted from voiced segments of the utterances with the 20ms frame length and 10ms frame shift. Each kind of feature is normalized first, hence the mean is zero and the standard deviation is one.

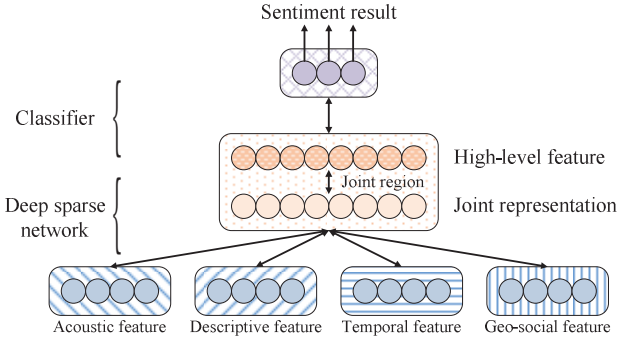

Fig. 8. The architecture of deep network for sentiment prediction with multi-modal features.

According to the data observations, we identify three kinds of correlation features to describe the relationships between utterances:

- *Temporal Correlation (TC)*: whether two utterances are recorded within the same hour of day. We formulate a N-dimensional (N=24 in our experiments) vector $TC = (t_1, \dots, t_N)$ for each utterance, where $t_i \in \{0,1\}$ indicates whether the utterance is recorded within the time interval $[i-1, i]$.
- *Descriptive Correlation (DC)*: whether two utterances belong to the same inquiry type. We use a N-dimensional (N=70 in our experiments) vector $DC = (d_1, \dots, d_N)$ to describe the descriptive feature of each utterance, where $d_i \in \{0,1\}$ is a binary variable indicating whether the utterance belongs to the rank i^{th} inquiry type.
- *Geo-social Correlation (GC)*: whether two utterances are recorded by the speakers in the same city. We represent the geo-social feature by a N-dimensional (N=21 in our experiments) vector $GC = (g_1, \dots, g_N)$, where $g_i \in \{0,1\} (i \in [1, \cdots, N-1])$ indicates whether the utterance is recorded by the speakers in the rank i^{th} city, and g_N indicates whether the utterance is recorded in other cities besides the top $N-1$ cities.

4. PROPOSED METHOD

4.1. Prediction model

In order to better handle the sentimental speech data and discover latent features, we propose a deep sparse neural network (DSNN) model for feature learning and classification. In this task, 113 acoustic features and three correlation features are extracted. The correlation features are abstract features rather than raw data. We employ early-fusion manner that combining input features at low level of network to learn a joint representation. High-level features are then learnt with a deep sparse network. Fig. 8 shows the overall architecture of our proposed model.

The model consists of 113 visible neurons in the acoustic region for the 113 extracted acoustic features. Descriptive features are assigned to 70 visible neurons each of which represents one inquiry type of the contents. There are extra 24 visible neurons for time interval of the utterance and also 21 neurons corresponding to the city it came from. Two layers of 400 and 200 hidden units are in the joint region. A softmax classifier is responsible for producing a probability result of the sentiment prediction.

We adopt Softplus activation function in each layer as proposed in [17], which is given by:

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

Rectifier networks are shown to have better performance by creating sparse representations with true zeros and preventing gradient vanishing effect. Softplus activation function is a smooth alternative to hard zero rectifier, which prevent hard zero to hurt back propagation optimization [18].

4.2. Model learning

To maximally utilize available data, we perform an unsupervised pre-training in auto-encoder scheme and then fine-tune the network with back-propagation optimization. In both phase we use batch update.

Pre-training can initialize lower layers in the whole network in an unsupervised manner. In pre-train phase, each two contiguous layers form an auto-encoder. Auto-encoder first represents inputs with higher layer neurons, then reconstruct the input with representation:

$$\tilde{x} = f\left(w^{(2)} f\left(w^{(1)} x + b^{(1)}\right) + b^{(2)}\right) \quad (3)$$

The goal of an auto-encoder with data set x and reconstruction $\tilde{x}$ is given by:

$$\min \frac{1}{m}\sum_{i=1}^{m}\frac{1}{2}\left\|\tilde{x}^{(i)} - x^{(i)}\right\|^2 + \frac{\lambda}{2}\sum_{l=1}^{2}\sum_{i=1}^{s_1}\sum_{j=1}^{s_2}\left(w_{ij}^{(l)}\right)^2 + \beta\sum_{j=1}^{s_2} KL(\rho||\rho_j) \quad (4)$$

where $w^{(1)}$ and $w^{(2)}$ are weight matrix of encoder and decoder, $b^{(1)}$ and $b^{(2)}$ are the bias. $f(\cdot)$ is the Softplus activation function. 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 given by

$$KL(\rho||\rho_j) = \rho\log\frac{\rho}{\rho_j} + (1-\rho)\log\frac{1-\rho}{1-\rho_j} \quad (5)$$

We calculate the gradient of the energy with respect to $w^{(1)}$, $w^{(2)}$, $b^{(1)}$ and $b^{(2)}$, and update them with average gradient over all the data.

Fine-tuning is in supervised manner with back-propagation optimization. The hypothesis of the network is defined by

$$a^{(4)} = \frac{e^{w_j^{(3)} a^{(3)}}}{\sum_{k=1}^{s_4} e^{w_k^{(3)} a^{(3)}}} \quad (6)$$

where $a^{(3)}$ is the activation of highest level feature neurons with feed forward network. The overall object of network is then given by

$$\min -\frac{1}{m}\sum_{i=1}^{m}\sum_{j=1}^{s_4} y_j^{(i)} \log a^{(4)} + \frac{\lambda_s}{2}\sum_{i=1}^{s_4}\sum_{j=1}^{s_3}\left(w_{ij}^{(3)}\right)^2 \quad (7)$$

where $y_j^{(i)}$ is the ground truth indicating whether example (i) belongs to class j by zero for false and one for true. By calculating the gradient all over the network with all the samples, we update the network in a batch with loop until it converges.

In our experiments, we set our parameters empirically to $\lambda = 0.15$, $\beta = 4$, $\rho = 0.25$ and $\lambda_s = 0.1$.

5. EXPERIMENT

5.1. Experimental setup

Datasets. We conduct our experiments on two datasets of different scales.

- *Dataset1*: We use the dataset described in subsection 2.2, which contains 2,000 utterances manually labeled with three sentiments: *Positive* (362), *Neutral* (875), *Negative* (763).
- *Dataset2*: We take advantage of the experimental dataset in [10], which contains 48,211 utterances consisted of six primary emotions (*Happy*, *Sad*, *Angry*, *Disgusted*, *Bored*, *Neutral*). Integrating four kinds of negative sentiments, we can construct a new dataset containing three categories: *Positive* (5721), *Neutral* (22740), *Negative* (19750).

Experiments. We design three kinds of experimental setups on *Deep Sparse Neural Network (DSNN)*, comparing with two baseline methods for the sentiment prediction task: *Naive Bayes (NB)* and *Artificial Neural Network (ANN)*.

- *Exp. 1*: We only use *Dataset1* for supervised learning with different methods separately.
- *Exp. 2*: We only use *Dataset2* for supervised learning with different methods separately.
- *Exp. 3*: We use the utterances and their labels in *Dataset1* for supervised training and testing, while combining the utterances without their labels in *Dataset2* for unsupervised training on DSNN.

Measures. In all experiments, we perform five-fold cross validation and quantitatively evaluated the sentiment prediction performance in terms of *Accuracy* and *F1-Measure* [10].

5.2. Results and Discussions

Performance Comparison. Table 1 shows the *Accuracy* and *F1-Measure* for NB, ANN and proposed DSNN. The *Accuracy* of the proposed DSNN model achieves 55.40%, while the NB and ANN model achieves 46.00% and 50.60% relatively in *Exp. 1*, which verifies that the more hidden layers in the neural network the greater the performance will be. DSNN also yields higher *F1-Measures* of *Positive* and *Negative* sentiments, which demonstrates that DSNN can

Table 1. Performance of sentiment prediction with different methods.

Classifier	F1-Measure (%)						
	Exp. 1			Exp. 2			Exp. 3
	NB	ANN	DSNN	NB	ANN	DSNN	DSNN
Positive	27.56	23.36	36.04	40.10	39.10	32.52	38.43
Neutral	50.47	58.83	64.55	43.29	64.41	69.20	64.54
Negative	49.86	50.78	53.46	55.80	52.14	59.62	55.40
Average	42.63	44.32	51.35	46.40	51.89	53.78	52.79
Accuracy (%)	46.00	50.60	55.40	49.26	56.91	61.85	56.50

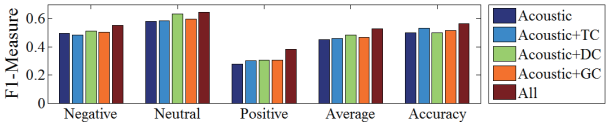

Fig. 9. Feature contribution analysis.

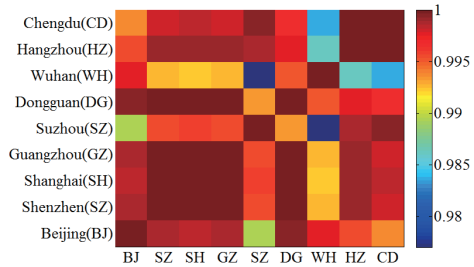

Fig. 10. The sentiment similarities between the top 10 cities.

better deal with the imbalance of data. The *Accuracies* of three methods in *Exp. 2* are all better than those in *Exp. 1*, which confirms that the large-scale data can benefit the performance of sentiment prediction. Comparing the results in *Exp. 1* with *Exp. 3*, we can see that the combination of unlabeled data for unsupervised learning slightly improves both *Accuracy* and *F1-Measure*, which shows the potential possibility of achieving better performance by DSNN model.

Feature Contribution Analysis. Fig. 9 illustrates the *Accuracy* and *F1-Measure* of each kind of correlation feature combining with acoustic features when conducting *Exp. 3* to predict sentiments. Comparing the *F1-Measures* of different correlation features, we find that *DC* makes the most improvement in all the sentiments, while *GC* benefits the prediction more than *TC*. However, *TC* achieves the best *Accuracy* in contrast with *DC* and *GC*. All the three correlation features make contributions to improving both *Accuracy* and *F1-Measure* of sentiment prediction, and their combination yields a better result.

Case Study. We also use one case study on *geo-social sentiment similarity* as the anecdotal evidence to further demonstrate the effectiveness of our method. We train a DSNN model on *Dataset2* and predict the sentiments of 213,997 utterances recorded by 1,216 active users in the top 10 cities. Then we can get a vector $P = (p_{pos}, p_{neu}, p_{neg})$ for each city, which indicates the proportion of *Positive*, *Neural* and *Negative* utterances in that city. Finally, we compute the correlation coefficients between the top 10 cities and draw the results in Fig. 10. The first-tier cities (Beijing, Shanghai, Shenzhen, Guangzhou) have a higher sentiment distribution similarity, since they have similar living environment that may evoke sentiments alike. Suzhou and Wuhan have the lowest sentiment similarity, reflecting the large gap between people's personalities in these two cities.

6. CONCLUSIONS

With the increasing adoption of smart phone voice dialogue applications, we can now use the large-scale networked voice data instead of text or image data to conduct sentiment prediction. Our main contributions are: 1) we investigate the data observations in human-mobile voice communication, and reveal several underlying sentiment patterns; 2) we define three kinds of correlation features based on the content and geo-information in our task, and combine them with acoustic features to achieve sentiment prediction from large-scale networked voice data; 3) we formulate the problem into a DSNN model, turning out good results.

Acknowledgement. This work is supported by the National Basic Research Program of China 2013CB329304, and partially supported by 2012CB316401. This work is also supported by the NSFC 61370023. We also thank Microsoft Research Asia-Tsinghua University Joint Laboratory for its support.

7. REFERENCES

[1] K. A. Goyal and A. Sadasivam, "A critical analysis of rational & emotional approaches in car selling," *In Proc. of IJBRM*, vol. 1, no. 2, pp. 59-63, 2010.

[2] J. Bollen, H. Mao, and X. Zeng, "Twitter mood predicts the stock market," *Journal of Computational Science*, vol. 2, no. 1, pp. 1-8, 2010.

[3] J. Jia, S. Wu, X. Wang, P. Hu, L. Cai, J. Tang, "Can We Understand van Gogh's Mood? Learning to Infer Affects from Images in Social Networks," *In Proc. of ACM Multimedia 2012*, pp. 857-860, Nara, Japan, 2012.

[4] Y. Yang, P. Cui, W. Zhu, and S. Yang, "User interest and social influence based emotion prediction for individuals," *In Proc. of ACM Multimedia 2013*, pp. 785–788, Barcelona, Catalunya, Spain, 2013.

[5] M. E. Ayadi, M. S. Kamel, F. Karray, "Survey on speech emotion recognition: Features, classification schemes, and databases," *Pattern Recognition*, vol. 44, pp. 572-587, Mar. 2011.

[6] J. Nicholson, K. Takahashi, R. Nakatsu, "Emotion recognition in speech using neural networks," *Neural Compute Application.*, vol. 9, pp. 290–296, 2000.

[7] E. Kim, S. Gilbert, M. Edwards and E. Graeff, "Detecting Sadness in 140 Characters: Sentiment Analysis of Mourning Michael Jackson on Twitter," Technical report, Web Ecology Project, Boston, 2009.

[8] B. O'Connor, R. Balasubramanyan, B. R. Routledge and N. A. Smith, "From tweets to polls: Linking text sentiment to public opinion time series," *In Proc. of AAAI 2010*, Washington, DC, May 2010.

[9] J. Bollen, H. Mao, A. Pepe, "Modeling Public Mood and Emotion: Twitter Sentiment and Socio-Economic Phenomena," *In Proc. AAAI 2011*, pp. 450-453, San Francisco, California, USA, 2011.

[10] Z. Ren, J. Jia, L. Cai, K. Zhang, J. Tang, "Learning to Infer Public Emotions from Large-scale Networked Voice Data," *In Proc. of MultiMedia Modeling 2014*, pp. 327-339, Dublin, Ireland, 2014.

[11] J. Ngiam, A. Khosla, M. Kim, J. Nam, H. Lee, and A. Ng, "Multimodal deep learning," *In Proc. of ICML-11*, pp. 689-696, 2011.

[12] N. Srivastava, and R. Salakhutdinov, "Multimodal learning with deep Boltzmann machines," *In Proc. of NIPS 25*, pp. 2231-2239, 2012.

[13] D. Cui, "Analysis and Conversion for Affective Speech," [doctoral dissertation], Tsinghua University, 2007.

[14] T. Nwe, S. Foo, L. De Silva, "Speech emotion recognition using hidden Markov models," *Speech Communication*, vol. 41, pp. 603–623, 2003.

[15] H. Kawahara, A. De Cheveigne, H. Banno, T. Takahashi, and T. Irino, "Nearly Defect-free F0 Trajectory Extraction for Expressive Speech Modifications based on STRAIGHT," *In Proc. of INTERSPEECH 2005*, pp. 537-540, Lisboa, 2005.

[16] D. Wang, and S. Narayanan, "An acoustic measure for word prominence in spontaneous speech," *In Proc. of TASLP*, vol. 15, no. 2, pp. 690-701, 2007.

[17] C. Dugas, Y. Bengio, F. Bélisle, C. Nadeau, and R. Garcia, "Incorporating second-order functional knowledge for better option pricing," *In Proc. of NIPS*, pp. 472-478, 2001.

[18] X. Glorot, A. Bordes, and Y. Bengio, "Deep Sparse Rectifier Networks," *In Proc. of AISTATS*, JMLR W&CP Volume, vol. 15, pp. 315-323. 2011.