Emotion Inferring from Large-scale Internet Voice Data: A Multimodal Deep Learning Approach

Suping Zhou
Department of Computer Science
and Technology
Beijing National Research Center
for Information Science and Technology
Tsinghua University
Beijing, China
supingzhou@qq.com

Jia Jia*
Department of Computer Science
and Technology
Key Laboratory of Pervasive Computing,
Ministry of Education
Tsinghua University
Beijing, China
jjia@tsinghua.edu.cn

Yanfeng Wang
Sogou Corporation
Beijing, China
wangyanfeng@sogou-inc.com

Wei Chen
Sogou Corporation
Beijing, China
chenweibj8871@sogou-inc.com

Fanbo Meng
Sogou Corporation
Beijing, China
mengfanbosi0935@sogou-inc.com

Ya Li
Institute of Automation
National Laboratory of
Pattern Recognition
Chinese Academy of Sciences
Beijing, China
yli@nlpr.ia.ac.cn

Jianhua Tao
Institute of Automation
National Laboratory of
Pattern Recognition
Chinese Academy of Sciences
Beijing, China
jhtao@nlpr.ia.ac.cn

***Abstract*—Voice Dialogue Applications(*VDAs*) increase popularity nowadays. As the same sentence expressed with different emotion may convey different meanings, inferring emotion from users' queries can help give a more humanized response for VDAs. However, the large-scale Internet voice data involving a tremendous amount of users, bring in a great diversity of users' dialects and expression preferences. Therefore, the traditional speech emotion recognition methods mainly targeting at acted corpora cannot handle the massive and diverse data effectively. In this paper, we propose a semi-supervised Emotion-oriented Bimodal Deep Autoencoder (EBDA) to infer emotion from large-scale Internet voice data. Specifically, as the previous research mainly focuses on acoustic features only, we utilize EBDA to fully integrate both acoustic and textual features. Meanwhile, to employ large-scale unlabeled data to enhance the classification performance, we adopt a semi-supervised strategy. The experimental results on 6 emotion categories based on a dataset collected from Sogou Voice Assistant [1] containing 7.5 million utterances outperform several alternative baselines (+10.18% in terms of F1 on average). Finally, we show some interesting case studies to further demonstrate the practicability of our model.**

***Index Terms*—Emotion, Internet Voice Data, Bimodal Deep Autoencoder**

I. Introduction

The increasing popularity of Voice Dialogue Applications(VDAs), such as Siri [2], brings great convenience to our daily life. As we all know, the same words said in different emotion can convey quite different messages. If we can infer emotion from these large-scale Internet voice data of users' queries in VADs, it would assist to understand the true meaning of users as well as provide more humanized responses.

However, fulfilling the task is not a trivial issue. As for speech emotion recognition, many previous works on feature selection and learning methods have been done. [1] propose to use phase-based features to build up such an emotion recognition system. And [2] introduce a Boosted-GMM algorithm which boost the emotion recognition rates effectively and significantly. However, these works are primarily focused on acoustic features only. Besides, done on corpora data (IEMOCAP database [3], etc.), these works have limited and easily labeled benchmark data. Although latest work [4], [5], [6] focus on proposing solutions to infer emotion from large-scale Internet voice data, there still remain two challenges unsolved in the specific situation of VDAs: 1) Beside speech information, the speech-to-text information is also provided by VDAs. Can we integrate multiple modalities(speech and text) to help enhance the performance on inferring emotion? 2) Unlike the traditional speech emotion recognition methods based on acted labeled data, the tremendous amounts of VDA users bring in a great diversity of users' dialects and expression preferences. Besides, due to the massive scale of our dataset, manually labeling the emotion for every utterance is not practical. Therefore, how to utilize those large-scale unlabeled data to increase the emotion inferring accuracy?

In this paper, employing a real-world voice dataset from Sogou Voice Assistant containing 7.5 million utterances assigned with its corresponding speech-to-text information (provided by [1]), we study the problem of emotion inferring for large-

*Corresponding author: J. Jia (jjia@mail.tsinghua.edu.cn)

[1] http://yy.sogou.com

[2] http://www.apple.com/ios/siri/

978-1-5386-5311-1/18/$31.00 ©2018 IEEE

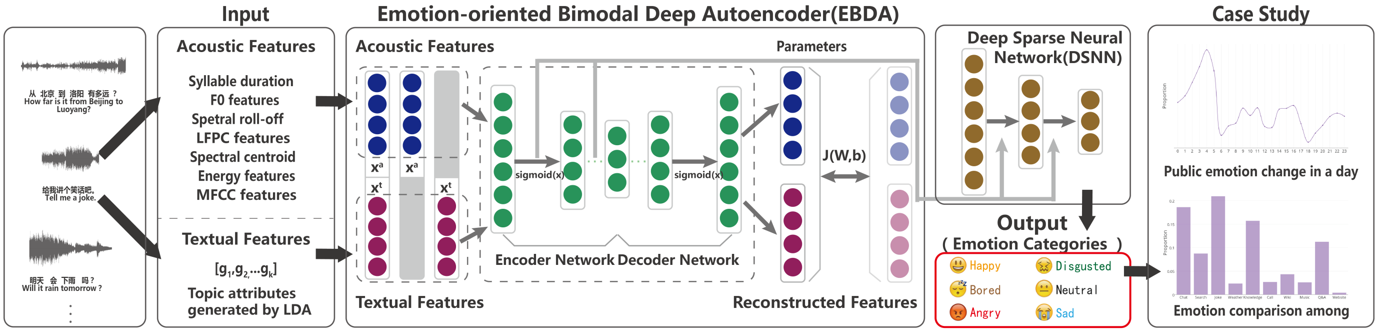


Fig. 1. The workflow of our framework.

scale Internet voice data. Specifically, we firstly propose a semi-supervised Emotion-oriented Bimodal Deep Autoencoder (EBDA) solution to infer emotion from large-scale Internet voice data integrating both acoustic and textual features. We adopt Latent Dirichlet Allocation (LDA) [7] which is widely used in the text-based sentiment analysis and achieves good performance to get the text features, and the feature selection algorithm used in [5] to extract 113 acoustic features(e.g. energy, f0, MFCC, LFPC). Next, we manually label 3000 utterances into six emotion categories [5], namely disgust, happiness, anger, sadness, boredom and neutral. Those labeled utterances as well as 7.5 million unlabeled data are employed in our semi-supervised model to make our model flexible to diverse users. Then, we use the parameters we learn in EBDA as a more comprehensive pattern to initial the classifier Deep Sparse Neural Network(DSNN). Benefiting from involving the large-scale unlabeled data, the training process can cover more abundant linguistic phenomenon. These, to some extent, can help solve the problem of user diversity and no training data for specific user in the real-world VDAs. The experimental results on six emotion categories based on our dataset outperform several alternative baselines (+10.18%in terms of F1 on average). We also discover that the unlabeled data used in EBDA enhances the performance for +2.81% in terms of F1 on average. To further demonstrate the practicability of our model, we conduct some interesting case studies. The illustration of our work is shown in Figure 1.

II. Problem Formulation

Given a set of utterances V, we divide it into two sets V^L (labeled data) and V^U (unlabeled data). For each utterance $v \in V$, we denote $v = \{x^a, x^t\}$. x^a represents the acoustic features of each utterance, which is a N_a dimensional vector. x^t represents the textual features of each utterance, which is a N_t dimensional vector. In addition, X^a is defined as a $|V| * N_a$ feature matrix with each element x^a_{ij} denoting the jth acoustic feature of v_i. The definition of X^t is similar to X^a.

Definition. **Emotion.** Previous research [5] discovers that in human-mobile interaction, the emotion categories are different from theories about emotion related to facial expressions [8]. According to their findings, we adopt {*Happiness*, *Sadness*, *Anger*, *Disgust*, *Boredom*, *Neutral*} as the emotion space and denote it as E_S, where $S = 6$.

Problem. **Learning task.** Given utterances set V, we aim to infer the emotion for every utterance $v \in V$:

$$f : (V^L, V^U, X^a, X^t) \Rightarrow E_S \tag{1}$$

III. Methods

In order to incorporate both acoustic and textual information of the utterances, we propose a multimodal deep learning [9] based model, Emotion-oriented Bimodal Deep Autoencoder (EBDA), to fuse the two modalities better. EBDA utilizes the large-scale unlabeled data for feature learning, which covers the diversity of various utterances. The pretrained parameters in EBDA are used as the initial parameters of Deep Sparse Neural Network (DSNN), which employs the labeled data for classification to make better inferring on emotion categories. The structure of EBDA and DSNN is shown in Figure 2.

A. Emotion-oriented Bimodal Deep Autoencoder

Although the traditional Deep Autoencoder (DA) is an approach for feature learning, it cannot make use of the internal correlation between acoustic and textual features. Thus we propose an Emotion-oriented Bimodal Deep Autoencoder (EBDA) for feature learning. Regarding acoustic and textual information as two modalities of utterances, we can train EBDA to fuse the two modalities into a shared representation.

Given an utterance $v_i \in V$, the initial input vector $x_i = \{x^a_i, x^t_i\}$ represents the extracted feature vector. We use a multilayer neural network to rebuild x_i into $\hat{x}_i = \{\hat{x}^a_i, \hat{x}^t_i\}$, where $\hat{x}^a_i$ and $\hat{x}^t_i$ are optimized to be similar to the initial input vector x^a_i and x^t_i specifically. The hidden layers of the EBDA contain encoder network and decoder network illustrated as the green circles of Figure 2. The relationship between two adjacent layers depends on model parameters. After training, we determine the final parameters as the output of this step.

In order to capture the internal correlation between acoustic and textual information, we influence the training process of EBDA by preprocessing the dataset. Concretely, tripling the original dataset, we get $X_1 = \{X^a, X^t\}$, $X_2 = \{X^a, X^t\}$, and $X_3 = \{X^a, X^t\}$. Then we set the textual features of X_2 and the acoustic features of X_3 to zero. Now we get a new dataset $X = \{X_1, X_2', X_3'\}$, where $X_2' = \{X^a, 0\}$,

Algorithm 1 Emotion-oriented Bimodal Deep Autoencoder

```
Require: $X = \{X^a, X^t\}$, a preprocessed feature matrix.
Ensure: Final parameter $W$, the parameter after the training process.
Initialize model parameters $\theta^{(l)}, \alpha, \lambda_1, \lambda_2$
repeat
    $W^{(l)} = W^{(l)} - \alpha \frac{\delta}{\delta W^{(l)}} J(W, b)$
    $b^{(l)} = b^{(l)} - \alpha \frac{\delta}{\delta b^{(l)}} J(W, b)$
until convergence (Gradient Descent)
return $W$
```

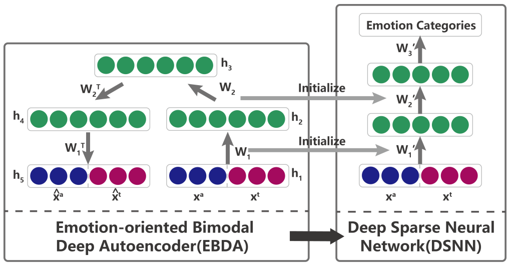
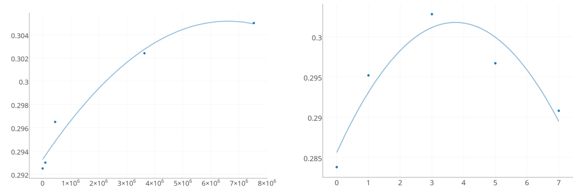


Fig. 2. The structure of EBDA and DSNN. (Parameters are explained in Section 4.)



(a) Size of unlabeled data in EBDA.

(b) Number of hidden layers in EBDA.

Fig. 3. Parameter analysis (in terms of F1-measure on average).

and $X_3' = \{0, X^t\}$. When training the autoencoder, we still expect it to recover all the three datasets into full features (i.e. $\hat{X} = \{X_1, X_1, X_1\}$). In this way, the EBDA learns the internal correlation between acoustic and textual features automatically.

Formally, supposing the EBDA has N_h layers, the recursion formula between two adjacent layers is:

$$h_i^{(l+1)} = \text{sigmoid}(W^{(l)} h_i^{(l)} + b^{(l)}) \tag{2}$$

where $h_i^{(l)}$ denotes the vector of lth hidden layers for v_i, $W^{(l)}$ and $b^{(l)}$ are the parameters between lth layer and $(l+1)$th layer and $sigmoid$ is the sigmoid function ($sigmoid(x) = \frac{1}{1+e^{-x}}$). Specially, $h_i^{(0)} = x_i$ and $\hat{x}_i = h_i^{(N_h+1)}$.

The cost function to evaluate the difference between x and $\hat{x}$ is defined as:

$$J(W, b) = \frac{\lambda_1}{2m} \sum_{i=1}^{m} ||x_i - \hat{x}_i||^2 + \frac{\lambda_2}{2} \sum_{l} (||W^{(l)}||_F^2 + ||b^{(l)}||_2^2) \tag{3}$$

where m is the number of samples, λ_1, λ_2 are hyperparameters and $|| \cdot ||_F$ denotes the Frobenius norm.

The first term in Equation 3 indicate average error of $\hat{x}$. The second term is a weight decay term for decreasing the values of the weights W and preventing overfitting [10]. The hyperparameters control the relative importance of the three terms. We define $\theta = (W, b)$ as our parameters to be determined. The training of EBDA is optimized to minimize the cost function:

$$\theta^* = \arg\min_{\theta} J(W, b) \tag{4}$$

The optimization method we adopt is Stochastic Gradient Descent Algorithm [11].

The complete algorithm for EBDA is summarized in Algorithm 1. After the training process, W is used as the initial parameters of Deep Sparse Neural Network.

B. Deep Sparse Neural Network

To infer emotion from the acoustic and textual features of utterances, we perform a supervised learning using Deep Sparse Neural Network to make the initial input extracted features $x_i = \{x_i^a, x_i^t\}$ classified to E_S. We first use parameters W learned from EBDA to initialize the lower layers of DSNN, then we finetune the network with back-propagation optimization with batch update. For the lower layers, the network is defined the same as Equation 2. For the highest layer, the hypothesis is defined as:

$$P_i = \text{softmax}(W^{(L)} h_i^{(L)}) \tag{5}$$

where $h_i^{(L)}$ and $W^{(L)}$ are the activation of the highest level feature neurons and parameter for v_i. P_i is the class probability. $softmax$ is the softmax function ($softmax(x) = \frac{e^x}{\sum_{k=1}^{S} e^{x_k}}$). The overall object of network is then given by

$$\min -\frac{1}{m} \sum_{i=1}^{m} \sum_{j=1}^{S} y_j^{(i)} \log P_j + \frac{\lambda}{2} \sum_{l} (||W^{(l)}||_F^2 + ||b^{(l)}||_2^2) + \beta \sum_{j=1} KL(\rho || \rho_j) \tag{6}$$

where m is the size of training utterances set, and $y_j^{(i)}$ is the ground truth indicating whether example(i) belongs to class j by zero for false and one for true. $W^{(l)}$ and $b^{(l)}$ are the parameters between lth layer and $(l+1)$th layer. 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 [12] given by

$$KL(\rho||\rho_j) = \rho \log \frac{\rho}{\rho_j} + (1 - \rho) \log \frac{1 - \rho}{1 - \rho_j} \tag{7}$$

We apply the Limited-memory Broyden Fletcher Goldfarb Shanno (L-BFGS) optimization algorithm [13] to train DSNN. By calculating the gradient all over the network with all the samples, we update the network in a batch with loop until it converges.

TABLE I
THE PRECISION, RECALL AND F1-MEASURE OF 5 METHODS FOR INFERRING EMOTIONS IN VDAS.

	Method	Happiness	Sadness	Anger	Disgust	Boredom	Neutral	Average
Precision	NB	0.2373	0.1182	0.1925	0.1487	0.1528	0.6983	0.2580
	RF	0.3231	0.2667	0.2917	0.1857	0.2821	0.5408	0.3150
	SVM	0.2958	0.2909	0.2900	0.2286	0.2025	0.5803	0.3147
	DSNN	0.2749	0.1748	0.2158	0.1736	0.2010	0.5856	0.2710
	DA+DSNN	0.2754	0.2714	0.2489	0.1708	0.2000	0.5899	0.2927
	EBDA+DSNN	0.3279	0.3118	0.3268	0.2194	0.2109	0.6011	**0.3330**
Recall	NB	0.2490	0.3015	0.4621	0.1574	0.2314	0.3184	0.2866
	RF	0.1728	0.0294	0.0966	0.0401	0.0431	0.9168	0.2165
	SVM	0.2160	0.1176	0.2000	0.0988	0.0627	0.8446	0.2566
	DSNN	0.2551	0.1324	0.1793	0.1543	0.1608	0.6678	0.2583
	DA+DSNN	0.2510	0.1397	0.1966	0.1265	0.1569	0.7153	0.2643
	EBDA+DSNN	0.2901	0.2132	0.2862	0.1327	0.1216	0.7545	**0.2997**
F1-Measure	NB	0.2430	0.1698	0.2718	0.1529	0.1841	0.4374	0.2432
	RF	0.2252	0.0530	0.1451	0.0660	0.0748	0.6803	0.2074
	SVM	0.2497	0.1675	0.2367	0.1379	0.0958	0.6880	0.2626
	DSNN	0.2647	0.1506	0.1959	0.1634	0.1786	0.6240	0.2629
	DA+DSNN	0.2626	0.1845	0.2197	0.1454	0.1758	0.6466	0.2724
	EBDA+DSNN	0.3079	0.2533	0.3051	0.1654	0.1542	0.6691	**0.3092**

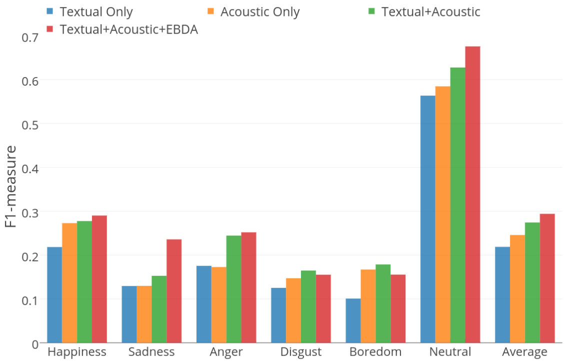


Fig. 4. Feature contribution analysis.

IV. EXPERIMENTS

A. Experimental setup

Dataset. We establish a corpus of voice data from Sogou Voice Assistant [1] (Chinese Siri) containing 7,534,064 Mandarin utterances recorded by 405,510 users in 2013. Every utterance is assigned with its corresponding speech-to-text information, query topic and user's location provided by Sogou Corporation.

We build a labeled dataset to do the emotion classification in DSNN. Due to the massive scale of our dataset, manually labeling the emotion for every utterance is not practical. Thus we randomly sample 3,000 utterances from the dataset and invite three people to annotate the emotion. The annotators are well trained and asked to label the emotion by listening to and reading the utterances simultaneously. When annotators have different opinions on the same utterance, they stop and discuss. If they cannot reach an agreement, the utterance is labeled *Unclear* and discarded. Finally, 2,942 utterances are labeled. The emotion distributions of these utterances are: *Neutral: 61.3%*, *Happiness: 13.2%*, *Disgust: 13.0%*, *Boredom: 4.8%*, *Anger: 3.9%* and *Sadness: 3.8%*. Besides, all the unlabeled data are employed to do feature learning in EBDA. Thus, the whole training process can be considered as a semi-supervised learning.

Comparison methods. To evaluate the effectiveness of our proposed method EBDA+DSNN, we compare the performance of emotion classification with some baseline methods, including Naive Bayes (NB) [14], Random Forest (RF) [15], Support Vector Machine (SVM) [16], and Deep Sparse Neural Network (DSNN). Employing the DSNN as classifier, we also compare the performance of different autoencoder settings for pretraining, including None (DSNN only), Deep Autoencoder (DA-DSNN) [17], Emotion-oriented Bimoal Deep Autoencoder (EBDA-DSNN).

Evaluation metrics. In all the experiments, we evaluate the performance in terms of F1-measure [18]. All the results reported are based on 5-fold cross validation.

B. Feature Extraction

To model the acoustic information of users' queries, we adopt the feature selection algorithm used in [5] to extract 113 acoustic features, including energy features (13), F0 features (13), MFCC features (26), LFPC features (24), spectral centroid features (13), spectral roll-off features (13), and syllable duration features (11).

To model the textual information of users' queries, Latent Dirichlet Allocation (LDA) [7] is widely used in the textual-based sentiment analysis and achieves good performance [19], [20]. We adopt the LDA method used in [20] to generate the textual features. Given utterance u's text t, it outputs a vector $g = \{g_1, g_2, ..., g_K\}$, where K is the length of the vector. K is an adjustable parameter, and in our work we set $K = 100$.

C. Experimental results

Performance of different classifiers and autoencoders. We make several comparisons among different classification models including NB, RF, SVM, and DSNN. Also, DA-DSNN is used as a baseline of EBDA. Table I shows the comparison results. In terms of F1-measure on average, the proposed EBDA-DSNN outperforms all the baseline methods: +5.9% compared with NB, +14.8% compared with RF, +4.4% compared with SVM, +2.8% compared with DSNN and +1.0% compared with DA-DSNN. From the comparison between DSNN and DA-/EBDA-DSNN, we can find that autoencoder pretraining strategy takes effect actually. Besides, the comparison between DA-DSNN and EBDA-DSNN indicates that EBDA's integration of acoustic and textual information does contribute to the results.

Comparing the prediction results of different categories, we find out the performance of boredom and disgust have lower performance than other categories. On one hand, these two categories have a small proportion in the labeled training data. Meanwhile, the same as [21] [6] report, utterances labeled boredom and disgust are often mixed together and difficult to distinguish. This phenomenon is obvious when compare to the results of anger which also only have a small proportion in the training data.

Feature contribution analysis. We discuss the contributions of acoustic and textual features. The F1-measure results of 4 methods (i.e. Textual Only, Acoustic Only, Textual + Acoustic, Textual + Acoustic + EBDA) for 6 emotion categories and their average are shown in Figure 4. We can see that the performance of "Textual Only" is far from satisfactory, while "Acoustic Only" performs better than "Textual Only". It indicates that when inferring emotion in the VDAs, acoustic features play more important role than textual features. Besides, "Textual + Acoustic" performs better than either "Textual Only" or "Acoustic Only" on average, proving the necessity of utilizing the two modalities simultaneously. Furthermore, we find that "Textual + Acoustic + EBDA" that combines acoustic and textual features by EBDA has the best performance, which proves the effectiveness of EBDA on modality fusion.

Parameter sensitivity analysis..We further test the parameter sensitivity about two key parameters in EBDA. 1) *Training data size*. From Figure 3(a), we can find that as the scale of unlabeled data increase, the performance gets better gradually. Thus utilizing large-scale unlabeled data does contribute to the results. 2) *Hidden layer number*. Theoretically, the description ability of EBDA can be improved by more layers. In Figure 3(b), the performance does increase with layer number less than 3, but gets worse when the number becomes larger due to overfitting. Therefore, we take 3 hidden layers in our experiments.

Error analysis. Finally, we analyze the possible sources of errors based on the emotion inferring results of the proposed EBDA-DSNN. 1) *Limited labeled data.* Due to the massive scale of our dataset, we are not able to label every utterance manually. Thus we only utilize 2,942 labeled utterances to train the DSNN, which may be not enough to get a well-trained model. 2) *Unbalanced data.* As 61.3% of the labeled utterances belong to Neural category, the data are extremely unbalanced, which has a negative impact on the results of classification definitely. 3) *Limited emotion categories.* Inferring emotion categories is a very difficult task, because emotion is highly subjective and complicated. At present, there is still no consensus on how to model emotion. Thus the 6 categories we adopt may not cover all the human feelings in the VDAs.

D. Case study

With the effective method we propose, we can apply several interesting case studies to discover some social phenomenons and mine the emotion pattern of public. We randomly sample 50,000 utterances and label their emotion categories using our model. Besides, the extra information of utterances including publishing time and talking topics are utilized to improve the analysis, provided by [1].

Time-emotion correlation. In Figure 5(a), the x-axis represents different time of a day, and the y-axis represents the proportion of the six kinds of emotion. From the figures, we summarize some interesting findings about time-emotion correlation as following.

- *Joy at night.* In Figure 5(b), the proportion of happiness from 17:00 to 20:00 is relatively high, indicating that people may feel more relaxed and comfortable when they finish the work during the day and start to enjoy the night.
- *Dull before dawn.* In Figure 5(c), the proportion of boredom is obviously higher and the proportion of anger is lower from 2:00 to 5:00, which is the regular time of sleeping. Such phenomenon can be explained by the people who suffer from insomnia. When people find it difficult to sleep on the early morning, they may feel bored and chat with VDA.

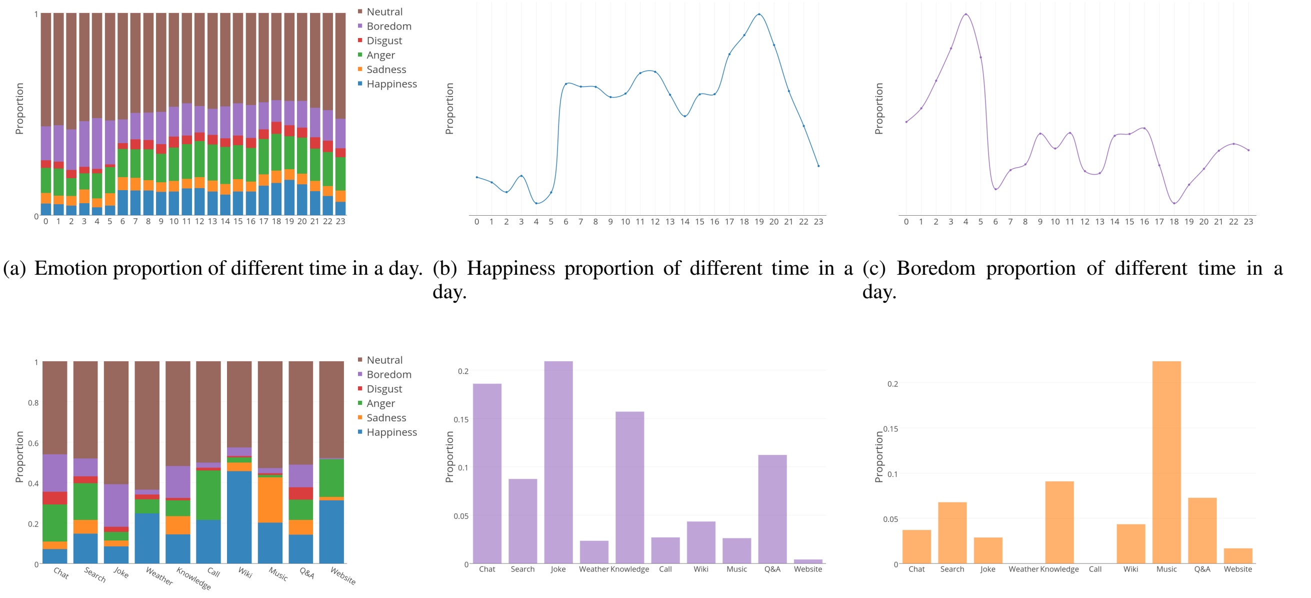
Topic-emotion correlation. In Figure 5(d), the x-axis represents ten different types of topics, and the y-axis represents the proportion of the six kinds of emotion. From the figures, we summarize some interesting findings about topic-emotion correlation as following.

- *Fun seeker.* In Figure 5(e), the proportion of boredom in topic "Chat" and "Joke" is relatively high, indicating that people may treat the VDA as a funny friend when they are bored.
- *Healing music.* In Figure 5(f), the proportion of sadness in topic "Music" is obviously higher than others, which indicates that music is a common way for people to comfort themselves when they are sad.

Employing the large-scale dataset that covers various kinds of data, our model is capable of inferring emotion for different utterances published by different users. It can support better analysis on people's emotion pattern and find more interesting emotion cases, which is useful for some social psychology studies.

V. Conclusion

In this paper, we construct a hybrid semi-supervised learning framework to do emotion inferring from large-scale Internet voice data. To integrate the acoustic and textual modalities of utterances, we propose an Emotion-oriented Bimodal Deep Autoencoder (EBDA), which also employs the large-scale unlabeled data for feature learning. Then we utilize a DSNN



(a) Emotion proportion of different time in a day.

(b) Happiness proportion of different time in a day.

(c) Boredom proportion of different time in a day.

(d) Emotion proportion of different topics.

(e) Boredom proportion of different topics.

(f) Sadness proportion of different topics.

Fig. 5. Findings of case studies.

initialized by EBDAs parameters to classify emotion, which employs the labeled dataset we build. The union structure of EBDA and DSNN is considered as semi-supervised learning. As shown in the experiment results and case studies, our framework turns out to be effective in speech emotion inferring. Furthermore, our work can be utilized in real-world applications. For instance, we can provide emotional response in the VDAs, which contributes to more humanized intelligent service.

Acknowledgment

This work is supported by the Innovation Method Fund of China (2016IM010200),and this work is supported by the National Natural, and Science Foundation of China (61602033).

References

[1] J. Deng, X. Xu, Z. Zhang, S. Frühholz, D. Grandjean, and B. Schuller, "Fisher kernels on phase-based features for speech emotion recognition," in *Dialogues with social robots*. Springer, 2017, pp. 195–203.

[2] P. Patel, A. Chaudhari, R. Kale, and M. Pund, "Emotion recognition from speech with gaussian mixture models & via boosted gmm," *International Journal of Research In Science & Engineering*, vol. 3, 2017.

[3] C. Busso, M. Bulut, C.-C. Lee, A. Kazemzadeh, E. Mower, S. Kim, J. N. Chang, S. Lee, and S. S. Narayanan, "Iemocap: Interactive emotional dyadic motion capture database," *Language resources and evaluation*, vol. 42, no. 4, p. 335, 2008.

[4] Z. Ren, J. Jia, Q. Guo, K. Zhang, and L. Cai, "Acoustics, content and geo-information based sentiment prediction from large-scale networked voice data," in *Multimedia and Expo (ICME), 2014 IEEE International Conference on*. IEEE, 2014, pp. 1–4.

[5] Z. Ren, J. Jia, L. Cai, K. Zhang, and J. Tang, "Learning to infer public emotions from large-scale networked voice data," in *International Conference on Multimedia Modeling*. Springer, 2014, pp. 327–339.

[6] B. Wu, J. Jia, T. He, J. Du, X. Yi, and Y. Ning, "Inferring users' emotions for human-mobile voice dialogue applications," in *Multimedia and Expo (ICME), 2016 IEEE International Conference on*. IEEE, 2016, pp. 1–6.

[7] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent dirichlet allocation," *JMLR*, vol. 3, pp. 993–1022, 2003.

[8] P. Ekman and W. V. Friesen, "The repertoire of nonverbal behavior: Categories, origins, usage, and coding," *semiotica*, vol. 1, no. 1, pp. 49–98, 1969.

[9] J. Ngiam, A. Khosla, M. Kim, J. Nam, H. Lee, and A. Y. Ng, "Multimodal deep learning," in *Proceedings of the 28th international conference on machine learning (ICML-11)*, 2011, pp. 689–696.

[10] A. Ng, "Sparse autoencoder," *CS294A Lecture notes*, vol. 72, pp. 1–19, 2011.

[11] L. Bottou, *Large-Scale Machine Learning with Stochastic Gradient Descent*. Physica-Verlag HD, 2010.

[12] J. C. Keegel, *Information Theory and Statistics*. WILEY, 1959.

[13] J. Nocedal, "Nocedal, j.: Updating quasi-newton matrices with limited storage. math. comp. 35, 773-782," *Mathematics of Computation*, vol. 35, pp. 773–782, 1980.

[14] F. Burkhardt, A. Paeschke, M. Rolfes, W. F. Sendlmeier, and B. Weiss, "A database of german emotional speech." in *Interspeech*, vol. 5, 2005, pp. 1517–1520.

[15] A. Liaw and M. Wiener, "Classification and regression with random forest," *R News*, vol. 23, no. 23, 2002.

[16] C.-C. Chang and C.-J. Lin, "Libsvm: a library for support vector machines," *ACM Transactions on Intelligent Systems and Technology (TIST)*, vol. 2, no. 3, p. 27, 2011.

[17] Q. Guo, J. Jia, G. Shen, L. Zhang, L. Cai, and Z. Yi, "Learning robust uniform features for cross-media social data by using cross autoencoders," *Knowledge-Based Systems*, vol. 102, no. C, pp. 64–75, 2016.

[18] D. M. Powers, "Evaluation: from precision, recall and f-measure to roc, informedness, markedness and correlation," 2011.

[19] C. Lin, Y. He, R. Everson, and S. Ruger, "Weakly supervised joint sentiment-topic detection from text," *IEEE TKDE*, vol. 24, pp. 1134–1145, 2012.

[20] Y. Yang, J. Jia, S. Zhang, B. Wu, Q. Chen, J. Li, C. Xing, and J. Tang, "How do your friends on social media disclose your emotions?" in *AAAI*, 2014, pp. 306–312.

[21] S. Ramakrishnan and I. M. El Emary, "Speech emotion recognition approaches in human computer interaction," *Telecommunication Systems*, pp. 1–12, 2013.