INFERRING EMOTIONS FROM HETEROGENEOUS SOCIAL MEDIA DATA: A CROSS-MEDIA AUTO-ENCODER SOLUTION

Shumei Zhang[1], *Jia Jia*[1,*], *Yishuang Ning*[1]

[1]Department of Computer Science and Technology, Tsinghua University, Beijing 100084, China
Tsinghua National Laboratory for Information Science and Technology (TNList),

vikeydr@126.com

jjia@mail.tsinghua.edu.cn, ningys13@mails.tsinghua.edu.cn

ABSTRACT

Social media is rocking the world in recent year, which makes modeling social media contents important. However, the heterogeneity of social media data is the main constraint. This paper focuses on inferring emotions from large-scale social media data. Tweets on social media platform, always containing heterogeneous information from different combinations of modalities, are utilized to construct a *cross-media* dataset. How to integrate cross-media information and solve the problem of modality deficiency are main challenges. To address those challenges, this paper proposes a Cross-media Auto-Encoder(CAE) to infer emotions on cross-media data, and CAE is designed to reconstruct missing modalities and integrate heterogeneous representations. In our experiments, We employ a dataset of 226,113 tweets to infer emotions of tweets, and our method outperforms several machine learning methods (+11.11% in terms of F1-measure). Feature contribution analysis also verifies the importance of adopting cross-media features.

Index Terms— Cross-media, Emotions prediction, Auto-Encoder

1. INTRODUCTION

With the rapid development of social media, people now are readily to share their opinions and moods on network such as Twitter, Facebook and Sina Weibo. Growing social media data which can indicate users' emotions laid the foundation of emotion prediction. Collecting and analyzing these data can benefit in many fields, such as understanding the underlying dynamics of public emotions[1], discovering the social contradictions[2, 3] and preventing social unrest[4, 5].

Recent years have witnessed increasing research efforts on emotion prediction utilizing different modalities of information. After [6, 7] verified the text shared in social networks does convey emotions of uploaders, [8] successfully used images and few social features in emotion prediction, and applied this method in understanding emotions of van Gogh's artworks; [9] showed a model for emotion prediction in viewers of images and [10] built a sentiment-analysis visual schema. However, these works mainly focused on investigating emotion-related features instead of utilizing multiple modalities.

Previous work also pay much attention to multi-modal machine learning. [11] proposed a unified probabilistic framework for emotion tagging on multimedia data. [12] fulfilled the cross-media sentiment classification task, by studying the impact different algorithms and data sources have on the accuracy. Auto-encoder was also used in sentiment cluster problem [13], which outperformed traditional methods a lot. However, although these works successfully integrate multi-modal and heterogeneous information, they seldom consider the common deficiency of multiple information which influences the accuracy of emotion prediction a lot.

In this paper, we focus on studying how to infer emotions on social media platform. Tweets on social media platform, always containing heterogeneous information from different combinations of modalities, are utilized to construct a *cross-media* dataset. However, not every tweet contains information of each modality. For example, on a random sample of 62,000 tweets from Dec.2011 to Feb.2012 in Sina Weibo, which is the largest tweet platform in China, we find that 53.5% contain images while only 27.8% contain social actions (attitudes, retweets or comments). For emotion prediction, the reconstruction of missing modalities and the integration of heterogeneous representation are two critical challenges.

To address the challenges above, we propose a Cross-media Auto-Encoders(CAEs) model to infer emotions on cross-media data. CAE is actually a variation of Auto-Encoder[14], which is broadly used in prediction and cluster problems. [15] firstly used CAE developed from deep neutral network for stress detection, and [16, 17] both proposed their deep models based on Auto-encoders to learn the high-level feature representation shared by multiple modalities for cross-media retrieval. Inspired by previous work, CAE in our method is specially designed to fit the problem which is able to reconstruct missing modalities and integrate heteroge-

neous representations. Thus, it builds a unified representation of every tweet to help optimize the emotion prediction results of classifiers.

With the proposed method, our experiments employ a dataset of 226,113 Sina Weibo tweets to infer emotions and our method can significantly improve the F1-measure over some machine learning methods (+11.11%).

2. FEATURE EXTRACTION

Social media data contains heterogeneous information from multiple modalities. The text and image features used in our work have been verified as having a strong correlation with emotions in [6, 9], and social features are first proposed in [15].

2.1. Text features

We define text information of a tweet as the text this user posted and text contained in retweet information. For the sake of accuracy, we removed two following parts: text between '#' which is defined as tag; other users' names probably contained in retweet information. With effective text, LTP-cloud is used for word segmentation.

Positive, neutral and negative emoticons(3 dimensions): For each tweet, we calculate the number of different kinds of emoticons.

Positive and negative emotion words of different degrees(14 dimensions): Based on Chinese emotion word and adverb dictionaries from Hownet, we calculate the number of different degrees of emotion words to distinguish users' emotion status.

2.2. Image features

Previous work has proved that image can directly influence publisher's and observer's emotions. Thus, we download static images in dataset and extract following features from them.

Five color theme(15 dimensions): a combination of five dominant in the RGB color space, representing the main color distribution of an image.

Saturation, Brightness, Warm or cool color, Clear or dull color(6 dimensions): These features are formulated in HSV color space. Generally, people with negative emotions prefer lower saturation and brightness with cool and dull color, while happy people will be more attracted by opposite ones.

2.3. Social features

When a user posts a tweet on social media platform, his/her friends will probably interact with him/her in three ways: comment, retweet or attitude. These social interactions also imply one's emotion status to some degree. 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 attitudes respectively. Then with the number of a tweet's comments, retweets and attitudes N_i, we define the variation characteristic of social features VC_i as:

$$VC_i = (N_i - M_i)^2 / V_i, 1 \leq i \leq 3 \quad (1)$$

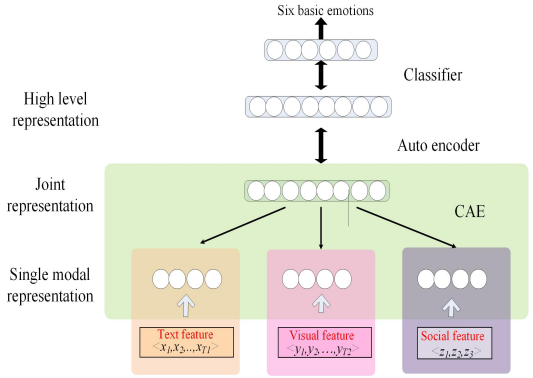

Fig. 1. The overall architecture of our method.

3. EMOTION PREDICTION MODELLING

Social media data, such as tweets, may not have complete feature modalities, including text, image, and social actions. For example, a user might not include an image in a tweet, or it may not elicit any retweets. Our dataset contains typical cross-media data. In order to integrate multiple modalities and reconstruct missing modalities of cross-media data , we propose a Cross-media Auto-Encoder (CAE) to generate a joint representation which is tolerant of missing modalities, which would optimize our prediction results a lot.

3.1. Overall Architecture

In the emotion prediction architecture, we first extract emotion-related features defined in Section 2. Then we feed these features to the input layer of CAE from three channels on a single modal layer. A hidden layer follows the single modal layer and takes the output normalized features from input layer, which both helps to reconstruct the missing modalities in the reconstruction layer and integrate heterogeneous representations to build a unified representation. The unified representation is benefit in jointly inferring emotions of cross-media data. At the top of the hierarchy, we employ a classifier to indicate one emotions category of a tweet based on the output unified representation. The overall architecture of our method is showed in Fig 1.

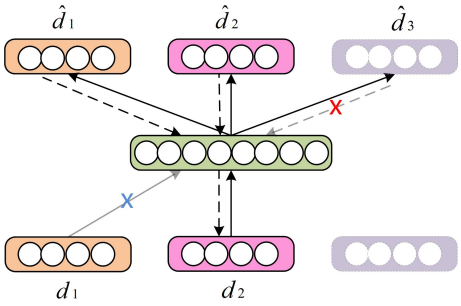

Fig. 2. Illustration of CAE for a training sample. The bottom blue cross means we disabled this modality in a sample from the input set, and the feedforward pass from the modality is blocked. The light color cylinders with dashed-line indicate the missing modality, while the upper red cross means that the backward pass from this modality is blocked.

3.2. CAE

A CAE for cross-media data is composed of three input channels, corresponding to the three modalities, a joint representation layer and three reconstruction channels.

The CAE can be formulated by the following equations:

$$\begin{cases} y_k = g(\sum_{m\in M}\sum_{j\in m} W_{kj}d_j + b), k = 1, 2, \ldots, K \\ \hat{d}_j = \hat{g}(\sum_k \hat{W}_{jk}y_k + \hat{b}), j = 1, 2, \ldots, J \end{cases} \tag{2}$$

where M is the set of modalities in the current sample, $d = \langle d_1, , d_J\rangle$ is the J-dimensional input, $y = \langle y_1, \ldots, y_K\rangle$ is the K-dimensional encoded representation, and $\hat{d} = \langle \hat{d}_1, \ldots, \hat{d}_J\rangle$ is the reconstruction, W and $\hat{W}$ are the weight matrix of encoder and decoder; b and $\hat{b}$ are corresponding biases; g and $\hat{g}$ are activation function respectively.

3.2.1. Training procedure

There are two principles for training CAE: 1) the representation should retain as much information as in the input data and 2) the representation should be equivalent for input of any combination of modalities. Based on these principles, we design the following training samples and functions.

Firstly, for the training samples, we build an augmented sample set which consists of the input set and the corresponding target set besides the original tweet features themselves. With training data containing all three modalities, we manually disable image features and/or social features when training the auto-encoder, while requiring it to reconstruct all three modalities, because text is the main modality that every tweet has. In this way, we follow the principles of training CAE and provide sufficient information for CAE.

Secondly, to maximally utilize data in training set, we calculate the reconstruction error only from available modalities for data with incomplete modalities. We formulate the target function as follows:

$$J = \frac{1}{2}\sum_{p\in P}\sum_{j\in p}(\hat{d}_j - d_j)^2 + \frac{\lambda}{2}\sum_{j,k}(W_{kj}^2 + \hat{W}_{kj}^2) \tag{3}$$

where the second term is the weight decay regularization with weight λ, P is the set of modalities available. In this way, we actually block the propagation of reconstruction error from the missing part since we do not have target for evaluation.

4. EXPERIMENTS

4.1. Dataset

For emotion prediction experiments, we choose Sina Weibo, the largest tweet platform in China, as the data source. Firstly, we download about 350,000,000 tweets related to several well-known social events posted from Oct.2009 to Oct.2012. For each tweet, we get its text, image and social action information– comments, retweets and attitudes. For a large-scale dataset, manually annotation is time-consuming and impractical. For this reason, we take the emotional words given by uploaders themselves in 'tags', which are contained in the text, as the standard of emotion classification. The emotion word lists are defined by Wordnet[18] based on synonyms. In this paper, we use the emotion space which is defined by [19]. This method was also used in [20]. Due to the high deficiency percent of 'tags' and emotional words in 'tags', we finally filter down 226,113 tweets, 11762 for anger, 14433 for fear, 37025 for sad, 79406 for happy, 33947 for disgust and 49540 for surprise.

4.2. Experimental Metrics

To evaluate the effectiveness of our proposed CAE , we use the Naive Bayes (NB), Support Vector Machine (SVM)[1], Random Forests (RF)[2] and Denoise Auto-Encoder (DAE)[3] as the comparison methods.

We compare the performance of the proposed method with baseline methods in terms of *F1-measure*. In our experiments, we use 5-fold cross validation.

4.3. Performance

The F1-measure results shown in Table 1 confirm the effectiveness of the proposed model. The following two abilities are the main advantages of our method:

1) Table 1 shows that the proposed method achieves a 5.78% improvement over DAE. Despite DAE is able to integrate heterogeneous features, it regards missing features as all zero vectors. When faced with cross-media data, CAE of our method is enforced to reconstruct the missing modalities

[1]LIBSVM: A library for support vector machines[21]
[2]NB and RF are provided by Matlab
[3]Compaired with CAE, it lacks in reconstruction ability.

Table 1. The F1-measure(%) of all methods

Method	Anger	Disgust	Fear	Happiness	Sadness	Surprise	Average
Naive Bayes	5.92	77.63	42.37	75.63	7.96	12.00	36.92
SVM	30.50	72.97	45.93	73.43	40.05	40.56	50.57
RF	44.12	82.57	36.40	78.45	45.26	38.59	54.23
DAE	45.28	67.47	59.57	58.30	60.19	35.21	54.34
CAE	41.40	81.73	57.92	77.92	53.21	48.56	**60.12**

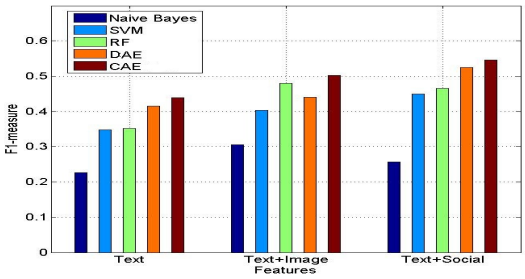

Fig. 3. Feature contribution in terms of F1-measure.

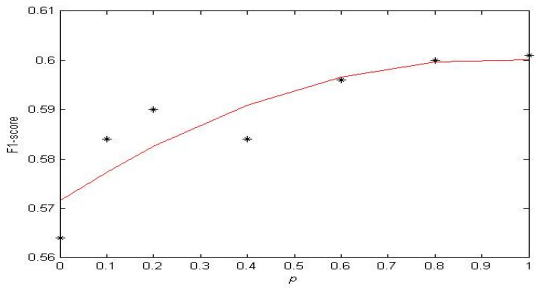

Fig. 4. Scalability of our method. ρ is the percentage of training data from whole dataset.

with any combination of input modalities, which makes sure that higher parts of the model can work with a uniform feature space. The ability to restore absent modalities is the main reason that our method outperforms DAE.

2) As we can see in Table 1, the proposed model outperforms Naive Bayes by 23.20% and SVM by 9.55% in terms of F1-measure. The main reason might be that the hidden layer of CAE is able to integrate heterogeneous representations and jointly infer emotions, while SVM and Naive Bayes simply mix all the features together and ignore the heterogeneity of modalities. Besides, although RF works well on balancing the data, it lacks in building a unified representation. And that's why our method achieves a 5.89% improvement over RF.

4.4. Feature contribution analysis

To explore how different features help in final results, we feed different combinations of features to our method and measure the contribution of each modality. The feature contribution here is shown in Fig 2, in terms of F1-measure.

The results are inspiring. On average, image and social features both make significant contributions to final performance. In term of F1-measure, image features contribute 7.04% and social features contribute 9.28%. The fact that different features achieve different contributions proves the necessity of capturing multiple modalities of information.

4.5. Scalable

To evaluate the scalability of our method, we conduct experiments on different scale of dataset and the result is showed in fig 4. At first, the performance increases with the scale of dataset. After that, the fitting curve indicates that the F1-measure of our method tends to be stable when ρ is bigger than 0.8, which verifies its scalability.

5. CONCLUSION

In this paper, we propose a novel solution, Cross-media Auto-Encoder, to deal with semantic analysis problem on heterogeneous social media data. The CAE here enables us to reconstruct deficient information and integrate heterogeneous modalities of features. To evaluate our proposed method, we take emotion inferring as an example and experimental results prove the effectiveness of our method.

Our method can benefit many fields, such as monitoring public opinions and providing sufficient advices for new products. In the future, we can adopt other modalities of information, especially user relationship. In addition, time sequence can also be considered in the method.

6. ACKNOWLEDGE

This work is supported by National Key Research and Development Plan (2016YFB1001200), the Innovation Method Fund of China (2016IM010200), the National Basic Research Program (973 Program) of China (2013CB329304), National Natural, and Science Foundation of China (61370023, 61602033).This work is partially supported by the major project of the National Social Science Foundation of China(13&ZD189).

7. REFERENCES

[1] Wendy M. Rahn, "A framework for the study of public mood," *Political Psychology*, vol. 17, no. 1, pp. 29–58, 1996.

[2] Xingjie Liu, Qi He, Yuanyuan Tian, Wang chien Lee, John Mcpherson, and Jiawei Han, "Event-based social networks: Linking the online and offline social worlds," *In Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 1032–1040, August 2012.

[3] Megan K. Torkildson, Kate Starbird, and Cecilia Aragon, *Analysis and Visualization of Sentiment and Emotion on Crisis Tweets.*, Cooperative Design, Visualization, and Engineering, 2014.

[4] Ryong Lee and Kazutoshi Sumiya, "Measuring geographical regularities of crowd behaviors for twitter-based geo-social event detection," *Proceedings of the 2nd ACM SIGSPATIAL International Workshop on Location Based Social Networks*, pp. 1–10, November 2010.

[5] Adam Bermingham and Alan F Smeaton, "On using twitter to monitor political sentiment and predict election results.," *In Proceedings of theWorkshop on Sentiment Analysis where AI meets Psychology*, pp. 2–10, November 2011.

[6] Mohamed Yassine and Hazem Hajj, "A framework for emotion mining from text in online social networks," *In Proceedings of the 2010 IEEE International Conference on Data Mining Workshops*, pp. 1136–1142, December 2010.

[7] Johan Bollen, Huina Mao, and Xiao-Jun Zeng, "Twitter mood predicts the stock market," *Journal of Computational Science*, vol. 2, no. 1, pp. 1–8, March 2011.

[8] Jia Jia, Sen Wu, Xiaohui Wang, Peiyun Hu, Lianhong Cai, and Jie Tang, "Can we understand van gogh's mood?: learning to infer affects from images in social networks," *In Proceedings of the 20th ACM international conference on Multimedia*, pp. 857–860, 2012.

[9] Yunhee Shin and Eun Yi Kim, "Affective prediction in photographic images using probabilistic affective model.," *In Proceedings of the ACM International Conference on Image and Video Retrieval*, pp. 390–397, July 2010.

[10] Jichang Zhao, Li Dong, Junjie Wu, and Ke Xu, "Moodlens: an emoticon-based sentiment analysis system for chinese tweets.," *In Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 1528–1531, August 2012.

[11] Shangfei Wang, Zhaoyu Wang, and Qiang Ji, "Multiple emotional tagging of multimedia data by exploiting dependencies among emotions," *Multimedia Tools Applications*, vol. 74, no. 6, pp. 1863–1883, 2015.

[12] rdia Sebaoun, Abdelhalim Rafrafi, Vincent Guigue, and Patrick Gallinari, "Cross-media sentiment classification and application to box-office forecasting," in *Conference on Open Research Areas in Information Retrieval*, 2013, pp. 201–208.

[13] Chunfeng Song, Yongzhen Huang, Feng Liu, Zhenyu Wang, and Liang Wang, "Deep auto-encoder based clustering," *Intelligent Data Analysis*, vol. 18, pp. S65–S76, 2014.

[14] Tara N. Sainath, Brian Kingsbury, and Bhuvana Ramabhadran, "Auto-encoder bottleneck features using deep belief networks," *Acoustics, Speech and Signal Processing (ICASSP), 2012 IEEE International Conference on*, pp. 4153 – 4156, March 2012.

[15] Huijie Lin, Jia Jia, Quan Guo, Yuanyuan Xue, Jie Huang, Lianhong Cai, and Ling Feng, "User-level psychological stress detection from social media using deep neural network," *IEEE International Conference on Multimedia & Expo 2014*, 2014.

[16] Xindi Shang, Hanwang Zhang, and Tat-Seng Chua, "Deep learning generic features for cross-media retrieval," in *International Conference on Multimedia Modeling*. Springer, 2016, pp. 264–275.

[17] X. Zhang, H. Zhang, Y. Zhang, Y. Yang, M. Wang, H. Luan, J. Li, and T. S. Chua, "Deep fusion of multiple semantic cues for complex event recognition.," *IEEE Transactions on Image Processing*, vol. 25, no. 3, pp. 1–1, 2015.

[18] George A. Miller, "Wordnet: A lexical database for english," *Communications of the ACM*, vol. 38, no. 11, pp. 39–41, 1995.

[19] Paul Ekman, "An argument for basic emotions," *Cognition & Emotion*, vol. 6, no. 3-4, pp. 169–200, 1992.

[20] Lexing Xie and Xuming He, "Picture tags and world knowledge: learning tag relations from visual semantic sources," *In Proceedings of the 21st ACM international conference on Multimedia*, pp. 967–976, October 2013.

[21] Chih-Jen Lin and Chih-Chung Chang, "Libsvm: a library for support vector machines," *ACM Transactions on intelligent systems and technology*, vol. 2, no. 3, pp. 389–396, 2007.