SocialNLP 2018 EmotionX Challenge Overview: Recognizing Emotions in Dialogues

Chao-Chun Hsu[1] **Lun-Wei Ku**[1,2]
[1] Academia Sinica, Taiwan.
{joe32140,lwku}@iis.sinica.edu.tw
[2] Joint Research Center for AI Technology and All Vista Healthcare, Taiwan.

Abstract

This paper describes an overview of the Dialogue Emotion Recognition Challenge, EmotionX, at the Sixth SocialNLP Workshop, which recognizes the emotion of each utterance in dialogues. This challenge offers the EmotionLines dataset as the experimental materials. The EmotionLines dataset contains conversations from Friends TV show transcripts (Friends) and real chatting logs (EmotionPush), where every dialogue utterance is labeled with emotions. Organizers provide baseline results. 18 teams registered in this challenge and 5 of them submitted their results successfully. The best team achieves the unweighted accuracy 62.48 and 62.5 on EmotionPush and Friends, respectively. In this paper we present the task definition, test collection, the evaluation results of the groups that participated in this challenge, and their approaches.

1 Introduction

Human emotion underlays in our daily interactions with other people, and study from Ekman(1987) shows that emotion is a universal phenomena across different cultures. An emotion detection system can improve mutual understanding between individuals by providing undetected emotion signal. For a common sense of human perception that emotion is inherently multi-modality including vision and speech, multi-modal emotion recognition plays an important role in emotion detection area(Sebe et al.; Kessous et al., 2010; Haq and Jackson, 2011). At the same time, studies in uni-modal emotion recognition also contribute in variety of modalities like vision(Ekman and Friesen, 2003), speech(Nwe et al., 2003) and text(Alm et al., 2005).

Chandler	Matthew Perry talking about signs in Las Vegas. (Neutral)
Chandler	I guess it must've been some movie I saw. (Neutral)
Chandler	What do you say? (Neutral)
Monica	***Okay! (Joy)***
Chandler	Okay! Come on! Let's go! All right! (Joy)
Rachel	Oh okay, I'll fix that to. What's her e-mail address? (Neutral)
Ross	Rachel! (Anger)
Rachel	All right, I promise. I'll fix this. I swear. I'll-I'll- I'll-I'll talk to her. (Non-neutral)
Ross	***Okay! (Anger)***
Rachel	Okay. (Neutral)

Table 1: "Okay!" of different emotions from Emotionlines dataset.

However, with the progress of social media and dialogue systems, especially the online customer services, textual emotion recognition has attracted more attention. In the social media, the hashtag and emoji are widely used and could provide substantial emotion clues(Qadir and Riloff, 2014; Kralj Novak et al., 2015). For the dialogue systems, instant emotion detection could help costumer service notice dissatisfaction of clients. Still, textual emotion recognition needs further exploration in dialogue systems for many reasons. For instance, a text segment can express various emotions given different context. Take the dialogue from Hsu et al.(2018) in Table 1 as an example, *Okay!* could be joy or anger in different scenarios. One more reason is that informal language and short sentence are everywhere in daily conversation. For instance, *lol* actually means *laugh out loud*. Therefore, emotion flow modeling and informal language understanding are essential for improving dialogue emotion recognition system.

For EmotionX shared task in SocialNLP 2018, we select an emotional dialogue dataset, Emo-

Proceedings of the Sixth International Workshop on Natural Language Processing for Social Media , pages 27–31,
Melbourne, Australia, July 20, 2018.

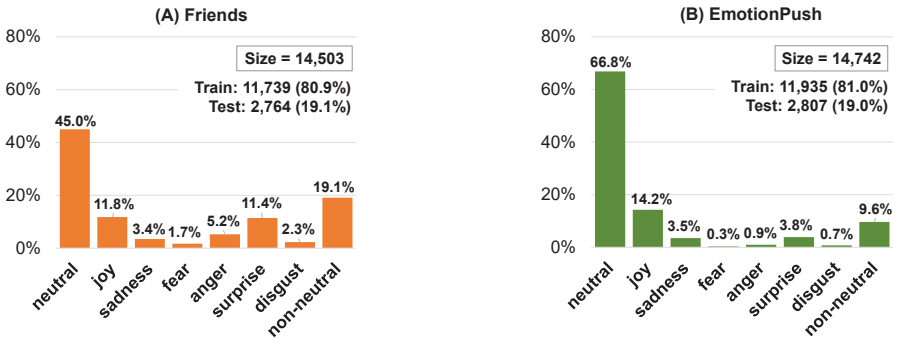

Figure 1: Emotion label distribution of Friends and EmotionPush datasets

tionlines, as the challenge dataset. A total of five teams presented their approaches, including feature-based and learning-based models, in this task. Neural models such as convolutional neural network(CNN) and recurrent neural network appear in all teams' work. The winning system achieves the unweighted accuracy 62.5% and 62.48% on Friends and EmotionPush dataset in the Emotionlines.

2 EmotionLines Dataset

EmotionLines is collected from two sources: Friends TV show transcripts (Friends) and Facebook messenger logs (EmotionPush). Dialogues are randomly selected from the raw data in four buckets of dialogue length [4-9], [10-14], [15-29], and [20-24], with 250 dialogues for each bucket. However, EmotionPush is a private chat log and releasing it may encounter privacy issues. To cope with this problem, Stanford Named Entity Recognizer (Manning et al., 2014) was adopted to replace the named entities in the corpus. In (Hsu et al., 2018), Amazon Mechanical Turk is utilized to label the emotion of every utterance. Following Ekman's(1987) six basic emotions and with neutral added, seven emotions are available for annotators in the labeling interface. To eliminate diverse emotion-labeled utterances, the utterance annotated with more than two emotions is considered as the non-neutral utterance. Finally, a total of eight emotion labels in both Friends and EmotionPush datasets are joy, anger, sadness, surprise, fear, disgust, neutral, and non-neutral. Figure 1 shows the emotion label distribution for these two datasets.

As we can see, more than 45% utterances are of neural emotion labels in both datasets, and the more severe emotion label imbalance in EmotionPush reflects the real situation that most of the utterances are neutral in daily conversations.

3 Challenge Setup

In shared task, each dataset is split into the training, the validation, and the testing set with 720, 80, 200 dialogues respectively. Due to the very few utterances of some emotions, we only evaluate the performance of recognizing four emotions: Joy, Anger, Sadness, Neutral, which was announced in the early announcement during the challenge. Generally speaking, recognizing strong emotions may provide more value than detecting the neutral emotion. To making a meaningful comparison in this challenge, we chose the unweighted accuracy(UWA) as our metric instead of the weighted accuracy(WA) as the latter is heavily compromised by the large proportion of the neutral emotion.

$$\text{WA} = \sum_{l \in C} s_l a_l \tag{1}$$

$$\text{UWA} = \frac{1}{|C|} \sum_{l \in C} a_l \tag{2}$$

where a_l denotes the accuracy of emotion class l and s_l denotes the percentage of utterances in emotion class l.

4 Submission

We receive 18 registrations and 5 teams submit their results successfully in the end. In the following, we summarize the approaches proposed by these 5 teams. More details could be found in their challenge papers.

Rank	Team	Model	Pre-trained Embedding	Other Resource	UWA (Friends)	UWA (EmotionPush)
1	**AR**	CNN	GloVe	Warriner′s, NRC, PERMA lexicons, formal list	62.5	62.48
2	**DLC**	LSTM+Attension	GloVe	-	59.65	55
2	**Area66**	Hierarchical LSTM +Attention+CRF	GloVe	-	55.38	56.73
4	**SmartDubai**	Logistic regression	fastText*	-	25.53	26.55
-	**JTML**	CNN+Attension	GloVe	-	33.35	46.75

Table 2: Overview of methods proposed by the participants and UWA of both datasets. JTML team is not in the ranking list because of late submission. * SmartDubai only used word and character TF-IDF as features for logistics regression. fastText is used by their other framework.

	Friends				EmotionPush			
	Neutral	Anger	Joy	Sadness	Neutral	Anger	Joy	Sadness
AR	68.3	**55.3**	**71.1**	**55.3**	76.3	**45.9**	**76**	51.7
DLC	90.1	49.1	68.8	30.6	94.2	24.3	70.5	31
Area66	73.5	39.8	57.6	50.6	88.2	21.6	63.1	**54**
SmartDubai	**99.5**	0	2.6	0	**99**	0	7.2	0
JTML	85.2	3.1	45.1	0	91.4	0	65.7	29.9

Table 3: Accuracy of four emotions on Friends and EmotionPush datasets.

DLC (Hang Seng Management College) A self-attentive BiLSTM network inspired by Transformer(Vaswani et al., 2017) is proposed. The self-attentive architecture on the top of BiLSTM could provide information between utterances and BiLSTM tries to model the word dependency in each utterance. Emoji symbols are converted to their meaning.

AR (Adobe Research) A CNN-DCNN autoencoder based emotion classifier is proposed. The latent feature of CNN-DCNN is augmented with linguistic features, such as lexical , syntactic, derived, and psycho-linguistic features as well as the formality list. The joint training of the classifer and the autoencoder improves generalizability, and linguistic features boost the performance on the minority class. AR is the only team that considers imbalance of emotions and also the only team that does not use the context information.

SmartDubai NLP (Smart Dubai Government Establishment) Multiple approaches are implemented by this team including logistic regression, Naive Bayes, CNN-LSTM, Xgboost, where they select TF-IDF, word vector, and some NLP features to train their models. In addition, the Internet slang is converted to its meaning e.g. *lol* is replaced by *lots of laughs*. Finally, logistic regression with TF-IDF of words and characters reached highest performance.

Area66 (TCS Research) A hierarchical attention network with a conditional random fields (CRF) layer on top of it is proposed. The word embeddings of the utterance are fed in to LSTM, then the attention mechanism captures the words with important emotion representations to form the sentence embedding. To model the context dependency, utterance embeddings of the dialogue are passed through another LSTM and CRF layer to predict emotion of utterances.

JTML (ESPOL University) A classifier using 1-dimensional CNN to extract utterance features with attention mechanism across utterances which obtains context information is provided. The proposed GRU-Attention model uses sequential GRU to learn relationship between previous utterances and current utterance. It achieves an improvement on UWA.

5 Evaluation Results

A brief summary of approaches proposed by teams participated in the EmotionX challenge and their corresponding final results are shown in Table 2. The performance varies across teams. Especially, in Table 3, we observed that SmartDubai and JTML obtained lower UWA scores because of the low accuracy on the minority emotion classes such as anger and sadness. In contrast, the winning team AR successfully reached a similar performance on four emotions on both datasets.

6 Discussion

6.1 Word Embedding

All teams used pre-trained word embedding: GloVe(Pennington et al., 2014) for four teams and fastText(Joulin et al., 2016) for one team. Area66 used GloVe-Tweet which is more related to informal language and the other teams did not mention the pre-trained data in their papers. Using pre-trained word embedding can reduce the unseen word issue in the testing phase especially for the relatively small dataset (Friends and EmotionPush only contain $\sim$ 14,000 utterances, which is small compared to the commonly used datasets for pre-training the embedding.)

6.2 Neural Network

Neural network architectures are adopted in all challenge papers. Acting as a universal feature extractor, neural network could minimize the feature engineering process. AR and JTML apply CNN to generate utterance embedding , and Area66 and DLC choose LSTM instead. By modeling context information in dialogue, DLC shows that self-attention improves UWA performance on both datasets. In addition, the AR team finds that adding a reconstruction loss of DCNN could improve generalizability.

6.3 Linguistic Features

Team AR combines latent feature of CNN-DCNN and linguistic features to prediction utterance emotion. Also, AR is the only team leveraging external resources, e.g. lexicons and the formal list. By adding linguistic features into neural model, the accuracy of anger is significantly boosted by 8.2% and 33.3% on Friends and EmotionPush, respectively. For the SmartDubai team, they use word and character TF-IDF independently with logistic regression. Results show it suppresses the Xgboost using TF-IDF and some linguistic features, e.g. sentence length and percentage of unique words, and outperforms CNN-BiLSTM using fastText word embedding, too.

6.4 Data Imbalance

Data imbalance directly harm the UWA performance. In Table 3, accuracy of minority emotions like anger and sadness are relatively low for SmartDubai and JTML, leading to low UWA performance. In contrast, AR is the only team considering data imbalance in the training process. They achieve balance accuracy on each emotion by applying weighed loss in the loss function , and ultimately obtain the best performance in the EmotionX challenge.

7 Conclusion

We have a succesfull dialogue emotion recognition challenge, EmotionX, in SocialNLP 2018. Many researchers have noticed this challenge and requested the datasets. Moreover, 5 teams successfully submitted their results this year. Various interesting approaches are proposed for this challenge, and the best performance achieves the unweighted accuracy 62.5% and 62.48% on Friends and EmotionPush dataset in the Emotionlines. We will continue organizing this challenge in SocialNLP 2019 and have planned to add the subtask of emotion dialogue generation, in the hope of encouraging and facilitating the research community to work on the emotion analysis on dialogues.

8 Acknowledgement

This research is partially supported by Ministry of Science and Technology, Taiwan, under Grant no. MOST 106-2218-E-002-043-, 107-2634-F-002-011- and 107-2634-F-001-004-. We especially thank Ting-Hao (Kenneth) Huang for deploying Amazon Mechanical Turk experiments and providing the figure of the emotion label distribution.

References

Cecilia Ovesdotter Alm, Dan Roth, and Richard Sproat. 2005. Emotions from text: machine learning for text-based emotion prediction. In *Proceedings of the conference on human language technology and empirical methods in natural language processing*, pages 579–586. Association for Computational Linguistics.

Paul Ekman and Wallace V Friesen. 2003. *Unmasking the face: A guide to recognizing emotions from facial clues*. Ishk.

Paul Ekman, Wallace V Friesen, Maureen O'sullivan, Anthony Chan, Irene Diacoyanni-Tarlatzis, Karl Heider, Rainer Krause, William Ayhan LeCompte, Tom Pitcairn, Pio E Ricci-Bitti, et al. 1987. Universals and cultural differences in the judgments of facial expressions of emotion. *Journal of personality and social psychology*, 53(4):712.

Sanaul Haq and Philip JB Jackson. 2011. Multimodal emotion recognition. In *Machine audition: principles, algorithms and systems*, pages 398–423. IGI Global.

Chao-Chun Hsu, Sheng-Yeh Chen, Chuan-Chun Kuo, Ting-Hao Huang, and Lun-Wei Ku. 2018. Emotionlines: An emotion corpus of multi-party conversations. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation, LREC 2018, Miyazaki, Japan, May 7-12, 2018.*

Armand Joulin, Edouard Grave, Piotr Bojanowski, Matthijs Douze, Hérve Jégou, and Tomas Mikolov. 2016. Fasttext.zip: Compressing text classification models. *arXiv preprint arXiv:1612.03651*.

Loic Kessous, Ginevra Castellano, and George Caridakis. 2010. Multimodal emotion recognition in speech-based interaction using facial expression, body gesture and acoustic analysis. *Journal on Multimodal User Interfaces*, 3(1-2):33–48.

Petra Kralj Novak, Jasmina Smailovi, Borut Sluban, and Igor Mozeti. 2015. Sentiment of emojis. *PLOS ONE*, 10(12):1–22.

Christopher D. Manning, Mihai Surdeanu, John Bauer, Jenny Finkel, Steven J. Bethard, and David McClosky. 2014. The Stanford CoreNLP natural language processing toolkit. In *Association for Computational Linguistics (ACL) System Demonstrations*, pages 55–60.

Tin Lay Nwe, Say Wei Foo, and Liyanage C De Silva. 2003. Speech emotion recognition using hidden markov models. *Speech communication*, 41(4):603–623.

Jeffrey Pennington, Richard Socher, and Christopher D. Manning. 2014. Glove: Global vectors for word representation. In *In EMNLP*.

Ashequl Qadir and Ellen Riloff. 2014. Learning emotion indicators from tweets: Hashtags, hashtag patterns, and phrases. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1203–1209.

Nicu Sebe, Ira Cohen, and Thomas S. Huang. *MULTIMODAL EMOTION RECOGNITION*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems 30*, pages 5998–6008. Curran Associates, Inc.

A Appendix

A.1 Registration Teams

EmotionX challenge obtained the attention of researchers including Amit Agarwal, Denis Lukovnikov, Egor Lakomkin, Fatiha Sadat, Gangeshwar Krishnamurthy, Gregory Grefenstette, Kushagra Singh, Pinelopi Papalampidi, Sashank Santhanam, Srishti Aggarwal, and Xiaolei Huang, who registered the challenge and obtained the dataset but failed to submit their results regretfully.