Seq2Emo: A Sequence to Multi-Label Emotion Classification Model

Chenyang Huang, Amine Trabelsi, Xuebin Qin, Nawshad Farruque,
Lili Mou, Osmar Zaïane
Alberta Machine Intelligence Institute
Department of Computing Science, University of Alberta
{chenyangh,atrabels,xuebin,nawshad,zaiane}@ualberta.ca
doublepower.mou@gmail.com

Abstract

Multi-label emotion classification is an important task in NLP and is essential to many applications. In this work, we propose a sequence-to-emotion (Seq2Emo) approach, which implicitly models emotion correlations in a bi-directional decoder. Experiments on SemEval'18 and GoEmotions datasets show that our approach outperforms state-of-the-art methods (without using external data). In particular, Seq2Emo outperforms the binary relevance (BR) and classifier chain (CC) approaches in a fair setting.[1]

1 Introduction

Emotion classification from text (Yadollahi et al., 2017; Sailunaz et al., 2018) plays an important role in affective computing research, and is essential to human-like interactive systems, such as emotional chatbots (Asghar et al., 2018; Zhou et al., 2018; Huang et al., 2018; Ghosal et al., 2019).

Early work treats this task as *multi-class classification* (Scherer and Wallbott, 1994; Mohammad, 2012), where each data instance (e.g., a sentence) is assumed to be labeled with one and only one emotion. More recently, researchers relax such an assumption and treat emotion analysis as *multi-label classification* (MLC, Mohammad et al., 2018; Demszky et al., 2020). In this case, each data instance may have one or multiple emotion labels. This is a more appropriate setting for emotion analysis, because an utterance may exhibit multiple emotions (e.g., "angry" and "sad", "surprise" and "joy").

The *binary relevance* approach (BR, Godbole and Sarawagi, 2004) is widely applied to multi-label emotion classification. BR predicts a binary indicator for each emotion individually, assuming that the emotions are independent given the input sentence. However, evidence in psychotherapy suggests strong correlation among different emotions (Plutchik, 1980). For example, "hate" may co-occur more often with "disgust" than "joy."

An alternative approach to multi-label emotion classification is the classifier chain (CC, Read et al., 2009). CC predicts the label(s) of an input in an autoregressive manner, for example, by a sequence-to-sequence (Seq2Seq) model (Yang et al., 2018). However, Seq2Seq models are known to have the problem of *exposure bias* (Bengio et al., 2015), i.e., an error at early steps may affect future predictions.

In this work, we propose a sequence-to-emotion (Seq2Emo) approach, where we consider emotion correlations implicitly. Similar to CC, we also build a Seq2Seq-like model, but predict a binary indicator of an emotion at each decoding step of Seq2Seq. We do not feed predicted emotions back to the decoder; thus, our model does not suffer from the exposure bias problem. Compared with BR, our Seq2Emo model implicitly considers the correlation of emotions in the hidden states of the decoder, and with an attention mechanism, our Seq2Emo is able to focus on different words in the input sentence that are relevant to the current emotion.

We evaluate our model for multi-label emotion classification on SemEval'18 (Mohammad et al., 2018) and GoEmotions (Demszky et al., 2020) benchmark datasets. Experiments show that Seq2Emo achieves state-of-the-art results on both datasets (without using external data). In particular, Seq2Emo outperforms both BR and CC in a fair, controlled comparison.

2 Related work

Emotion classification is an activate research area in NLP. It classifies text instances into a set of emotion categories, e.g., angry, sad, happy, and surprise. Well-accepted emotion categorizations include the six basic emotions in Ekman (1984) and the eight primary emotions in Plutchik's wheel of emotions (1980).

[1]Our code is available at https://github.com/chenyangh/Seq2Emo

Early work uses manually constructed emotion lexicons for the emotion classification task (Tokuhisa et al., 2008; Wen and Wan, 2014; Shahraki and Zaiane, 2017). Such lexicon resources include WordNet-Affect (Strapparava and Valitutti, 2004), EmoSenticNet (Poria et al., 2014), and the NRC Emotion Intensity Lexicon (Mohammad, 2018).

Distant supervision (Mintz et al., 2009) has been applied to emotion classification, as researchers find existing labeled datasets are small for training an emotion classifier. For example, Mohammad (2012) finds that social media users often use hashtags to express emotions, and thus certain hashtags can be directly regarded as the noisy label of an utterance. Likewise, Felbo et al. (2017) use emojis as noisy labels for emotion classification. Such distant supervision can also be applied to pretrain emotion-specific embeddings and language models (Tang et al., 2014; Ghosh et al., 2017).

In addition, Yu et al. (2018) apply multi-task learning to combine polarity sentiment analysis and multi-label emotion classification with dual attention.

Different from the above studies that use extra emotional resources, our work focuses on modeling the correlations among emotions. This improves multi-label emotion classification without using additional data. A similar paper to ours is the Sequence Generation Model (SGM, Yang et al., 2018). SGM accomplishes multi-label classification by an autoregressive Seq2Seq model, and is an adaptation of classifier chains (Read et al., 2009) in the neural network regime. Our paper models emotion correlation implicitly by decoder hidden states and does not suffer from the drawbacks of autoregressive models.

3 Methodology

Consider a multi-label emotion classification problem. Suppose we have K predefined candidate emotions, and an utterance or a sentence $\mathbf{x}$ can be assigned with one or more emotions. We represent the target labels as $\boldsymbol{y} = (y_1, \cdots, y_K) \in \{0, 1\}^K$ with $y_i = 1$ representing that the ith emotion is on.

Our Seq2Emo is a Seq2Seq-like framework, shown as Figure 1. It encodes $\mathbf{x}$ with an LSTM, and iteratively performs binary classifications over y_i with another LSTM as the decoder.

Encoder. We use a two-layer bi-directional LSTM to encoder an utterance. Specifically, we use both token-level and contextual pretrained embeddings to represent a word in the sentence.

Formally, let a sentence be $\mathbf{x} = (\mathrm{x}_1, \cdots, \mathrm{x}_M)$. We first encode each word x_i with GloVe embeddings (Pennington et al., 2014), denoted by $\mathrm{GloVe}(\mathrm{x}_i)$. We further use the ELMo contextual embeddings (Peters et al., 2018), which processing the entire sentence $\mathbf{x}$ by a pretrained LSTM. The corresponding hidden state is used as the embedding representation of a word x_i in its context. This is denoted by $\mathrm{ELMo}(\mathbf{x})_i$.

We use a two-layer bi-directional LSTM on the above two embeddings. The forward LSTM, for example, has the form

$$\boldsymbol{h}_t^{\overrightarrow{E}} = \mathrm{LSTM}^{\overrightarrow{E}}([\mathrm{GloVe}(\mathrm{x}_t); \mathrm{ELMo}(\mathbf{x})_t], \boldsymbol{h}_{t-1}^{\overrightarrow{E}})$$

where the superscript E denotes the encoder. Likewise, the backward LSTM yields the representation $\boldsymbol{h}_t^{\overleftarrow{E}}$. They are concatenated as $\boldsymbol{h}_t^E = [\boldsymbol{h}^{\overrightarrow{E}}; \boldsymbol{h}^{\overleftarrow{E}}]$.

Here, we use BiLSTM for simplicity, following Sanh et al. (2019) and Huang et al. (2019). Other pretrained models, such as the Tranformer-based BERT (Devlin et al., 2019), may also be adopted. This, however, falls out of the scope of our paper, as we mainly focus on multi-label emotion classification. Empirical results on the GoEmotions dataset shows that, by properly addressing multi-label classification, our model outperforms a Transformer-based model (Table 2).

Decoder. In Seq2Emo, an LSTM-based decoder is used to make sequential predictions on every candidate emotion. Suppose a predefined order of emotions is given, e.g., "angry," "joy," and "sad." The decoder will perform a binary classification over these emotions in sequence. The order, in fact, does not affect our model much, as it is the same for all training samples and can be easily learned. In addition, we feed a learnable emotion embedding as input at each step of the decoder. This enhances the decoder by explicitly indicating which emotion is being predicted at a step.

Different from a traditional Seq2Seq decoder, we do not feed previous predictions back as input, so as to avoid exposure bias. This also allows Seq2Emo to use a bi-directional LSTM as the decoder, which implicitly model the correlation among different emotions.

Without loss of generality, we explain the forward direction of the decoder LSTM, denoted by $\mathrm{LSTM}^{\overrightarrow{D}}$. The hidden state at step j is given by

$$\boldsymbol{h}_j^{\overrightarrow{D}} = \mathrm{LSTM}^{\overrightarrow{D}}([\boldsymbol{e}_j; \tilde{\boldsymbol{h}}_{j-1}^{\overrightarrow{D}}], \boldsymbol{h}_{j-1}^{\overrightarrow{D}}) \quad (1)$$

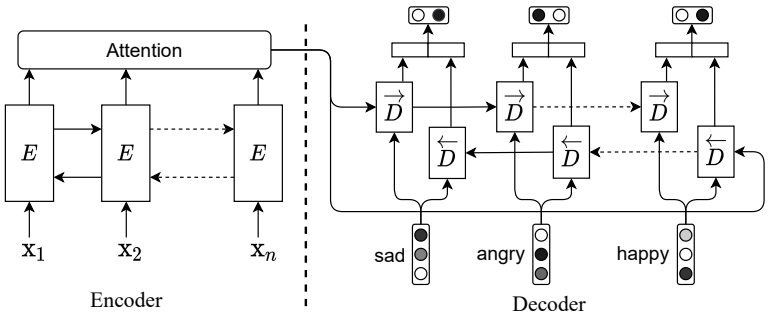


Figure 1: Overview of the Seq2Emo model.

where $\boldsymbol{e}_j$ is the embedding for the jth emotion, and $\tilde{\boldsymbol{h}}^{\overrightarrow{D}}_{j-1}$ is calculated by the attention mechanism in Luong et al. (2015).

Here, the attention mechanism dynamically aligns source words when predicting the specific target emotion at a decoding step. Let $\alpha^{\rightarrow}_{j,i}$ be the attention probability of the jth decoder step over the ith encoder step, computed by

$$s^{\rightarrow}_{j,i} = (\boldsymbol{h}^{\overrightarrow{D}}_j)^\top W^{\rightarrow}_a \boldsymbol{h}^E_i \quad (2)$$

$$\alpha^{\rightarrow}_{j,i} = \frac{\exp(s^{\rightarrow}_{j,i})}{\sum_{i=1}^{M} \exp(s^{\rightarrow}_{j,i})} \quad (3)$$

where M is the number of encoder steps, and $s^{\rightarrow}_{j,i}$ computes an unnormalized score for each pair of $\boldsymbol{h}^{\overrightarrow{D}}_j$ and $\boldsymbol{h}^E_i$ with a learnable parameter matrix $W^{\rightarrow}_a$. Then, we compute an attention-weighted sum of encoder hidden states as the context vector $\boldsymbol{c}^{\rightarrow}_j$:

$$\boldsymbol{c}^{\rightarrow}_j = \sum_{i=1}^{M} \alpha^{\rightarrow}_{j,i} \boldsymbol{h}^E_i \quad (4)$$

The context vector is concatenated with the LSTM hidden state as $\tilde{\boldsymbol{h}}^{\overrightarrow{D}}_j = [\boldsymbol{c}^{\rightarrow}_j; \boldsymbol{h}^{\overrightarrow{D}}_j]$. Likewise, we compute $\tilde{\boldsymbol{h}}^{\overleftarrow{D}}_j$ for the backward decoder LSTM. They are further concatenated for predicting the emotion in question:

$$P(y_j = 1|\mathbf{x}) = \sigma(\boldsymbol{w}_j^\top [\tilde{\boldsymbol{h}}^{\overrightarrow{D}}_j; \tilde{\boldsymbol{h}}^{\overleftarrow{D}}_j] + b_j) \quad (5)$$

where σ is a sigmoid function; $\boldsymbol{w}_j$ and b_j are the parameters for predicting the jth emotion. Notice that $\boldsymbol{w}_j$ and b_j are different at decoding different steps, because we are predicting different emotions. This treatment is similar to the binary relevance approach (BR, Godbole and Sarawagi, 2004).

Our Seq2Emo implicitly models the correlations among emotions through the decoder's bi-directional LSTM hidden states, which is more suited to multi-label classification than BR's individual predictions. Our Seq2Emo also differs from the classifier chain approach (CC, Read et al., 2009), which uses $\mathrm{softmax}$ to predict the next plausible emotion from all candidates. Thus, CC has to feed the previous predictions as input, and suffers from the exposure bias problem. By contrast, we predict the presence of all the emotions in sequence. Hence, feeding back previous predictions is not necessary, and this prevents the exposure bias. In this sense, our model combines the merits of both BR and CC.

4 Experimental Setup

Datasets. We conduct experiments on two multi-labeled emotion datasets: SemEval'18 (Affect in Tweets: Task E-c, Mohammad et al., 2018) and GoEmotions (Demszky et al., 2020). Compared with GoEmotions, SemEval'18 has fewer emotion categories, and is smaller in size. Both datasets come with standard train-dev-test splits. Appendix A shows the statistics of these datasets.

Metrics. Following Yang et al. (2018) and Mohammad et al. (2018), we use Jaccard Index (Rogers and Tanimoto, 1960), Hamming Loss (Schapire and Singer, 1999), Macro- and Micro-averaged F1 scores (Chinchor, 1992) as the evaluation metrics. Among them, Jaccard, Macro- and Micro-F1 are different ways of counting correctly predicted labels (the higher, the better); Hamming Loss (HL) counts the misclassifications (the lower, the better).

Baselines. On SemEval'18, we compare our system with the top submissions from the SemEval-2018 competition and recent development. NTUA-SLP (Baziotis et al., 2018) uses large amount of external emotion-related data to pretrain an LSTM-based model. TCS Research's system (Meisheri and Dey, 2018) uses the support vector ma-

chine with mannually engineered features: output from LSTM models, emotion lexicons (Mohammad and Kiritchenko, 2015), and SentiNeural (Radford et al., 2017). PlusEmo2Vec (Park et al., 2018) combines neural network models, which are pretrained by using emojis as labels (Felbo et al., 2017). Apart from the competition, Yu et al. (2018) propose DATN, which introduces sentiment information through dual-attention. These aforementioned systems are based on the BR approach. SGM (Yang et al., 2018), however, is a CC-based model for multi-label classification. We include it as a baseline by using its publicly released code.[2]

Since GoEmotions dataset is fairly recent, we only include the results originally reported by Demszky et al. (2020).

Settings. For the encoder, we set the two-layer bi-directional LSTM's dimension to 1200. Given the small number of emotions to embed, we set the dimension of decoder LSTM to 400. The GloVe embedding is 300 dimensional, and the ELMo embedding is 1024 dimensional. We use the Adam optimizer (Kingma and Ba, 2015), where the learning rate is set to 5e-4 initially and decayed with cosine annealing. The batch size is set to 16 for SemEval'18, and set to 32 for GoEmotions for efficiency concerns.

We perform 5-fold cross-validation on the combined train-dev split for each experiment. Within each fold, we apply early stopping to prevent overfitting and return the best model based on Jaccard accuracy for testing. We then merge the predicted results over the test set by majority voting. Additionally, we repeat each 5-fold experiment 5 times to further improve reduce noise.

5 Results

Overall performance. Table 1 presents the results on the SemEval'18 dataset. The proposed Seq2Emo outperforms the top submissions of the SemEval-2018 shared task in general. Compared with the median submission, Seq2Emo outperforms over 10% in the Jaccard accuracy. Admittedly, Seq2Emo performs slightly lower (but comparably) with NTUA-SLP and DATN, both introducing extra emotion/sentiment information through transfer learning. Our work, however, focuses on modeling the multi-label classification problem for emotion analysis and achieves high performance.

[2] https://github.com/lancopku/SGM

	Jaccard ↑	Micro F. ↑	Macro F. ↑	HL ↓
Random	18.50	30.70	28.50	–
SVM-Unigrams	44.20	57.00	44.30	–
SGM	45.14	55.11	–	0.1668
*Median**	47.10	59.90	46.40	–
[+] PlusEmo2Vec	57.60	69.20	49.70	–
[+] TCS Research	58.20	69.30	53.00	–
[+] NTUA-SLP	**58.80**	**70.10**	52.80	–
[+] DATN	58.30	–	**54.40**	–
BR†	57.64	68.89	50.32	0.1262
BR-att†	58.13	69.49	51.60	0.1237
CC†	58.16	69.19	51.07	0.1381
Seq2Emo (uni)†	58.22	69.60	50.98	0.1229
Seq2Emo†	**58.67**	**70.02**	**51.92**	**0.1214**
t-test	$p < 0.1$	$p < 0.01$	$p < 0.1$	$p < 0.01$

Table 1: Results on the SemEval'18 dataset. *Median refers to the median score reported among the submissions. [+] denotes additional emotion/sentiment information is used. † denotes the results obtained by our implementations.

While both NTUA-SLP and DATN are based on the BR approach, we implement additional baselines for fair comparison. In particular, we implement BR and BR-att variants, where the latter uses an attention mechanism when predicting the emotions, similar to our Seq2Emo. In the same spirit, we also implement a CC-based baseline, which is a Seq2Seq model predicting the next emotion among all candidates. For fair comparison, all of the BR, BR-att, and CC variants are trained with the same setting as our Seq2Emo. In this controlled setting, we observe that the proposed Seq2Emo consistently outperform BR, BR-att, and CC on the SemEval'18 dataset in all metrics.

For the GoEmotions dataset, we show the results in Table 2. Since it is a very new dataset, we can only find previous reported results from Demszky et al. (2020). In addition, we include BR, BR-att, and CC for fair comparison. Results show that Seq2Emo outperforms other models on most of the metrics, except that Seq2Emo is worse than CC on Jaccard accuracy. This is understandable, as we have quite a few metrics with different datasets.

It is worth noting that the model of Demszky et al. (2020) is based on BERT (Devlin et al., 2019). We replicate their approach to obtain all the evaluation metrics. We observe that our replication achieves a similar Macro-F1 to Demszky et al. (2020), and thus our replication is fair. The results show that our Seq2Emo achieves comparable or higher performance than the BERT-based model.

We run one-sided t-tests to compare Seq2Emo with the best competing model that does not use additional data, shown in Tables 1 and 2. Results ver-

#	Model	Jaccard ↑	Micro F. ↑	Macro F. ↑	HL ↓
1	BERT (Demszky et al., 2020)	–	–	46.00	–
2	BERT (our implementation)†	53.06	58.49	46.23	0.0312
3	BR†	52.76	58.21	45.38	0.0312
4	BR-att†	53.35	58.53	45.11	0.0310
5	CC†	**55.61**	58.38	43.92	0.0352
6	Seq2Emo (uni)†	53.07	58.76	45.30	0.0306
7	Seq2Emo†	53.79	**59.57**	**47.28**	**0.0302**
	t-test	$p < 0.05$	$p < 0.05$	$p < 0.01$	$p < 0.01$

Table 2: Results on the GoEmotions dataset. † denotes the results obtained by our implementations. t-test compares Row 7 with the best model in Rows 3–6 in each metric.

ify that most of the comparisons are statistically significant (although some are more significant than others). The two experiments provide consistent evidence on the effectiveness of our Seq2Emo.

Seq2Emo with an uni-directional decoder. One of the virtues of Seq2Emo is that it can use a bi-directional LSTM decoder. To show its effectiveness, we perform experiments on Seq2Emo with an uni-directional decoder, denoted as "Seq2Emo (uni)." We show the results in Tables 1 and 2 for SemEval'18 and GoEmotions datasets, respectively. We first observe that Seq2Emo performs better than Seq2Emo (uni), which in turn is better than BR-att that predicts emotions individually. This confirms that our Seq2Emo is able to implicitly model the correlation of different emotions, and that a bi-directional decoder is better than a uni-directional one.

Order of emotions. Both Seq2Emo and the classifier chain (CC) predict emotions sequentially. The difference is that our Seq2Emo predicts the presence (or not) of an emotion in a predefined order. CC predicts the next salient emotion autoregressively, it learns the emotion order from the training data. We try different orders, including the original order in the dataset and the ascending/descending order based on emotion frequency. We also try an order where the emotion frequency first increases and then decreases (concave-down), and vice versa (concave-up). We perform experiments on SemEval'18 and report the Jaccard accuracy and the standard deviations in Table 3.

The results show that Seq2Emo is the least affected by the order of the emotions, whereas the performance of CC varies largely. This verifies that the emotion order does not affect Seq2Emo much as it can be easily learned. CC is more sensitive to emotion order and has a larger variance, as it suffers from the exposure bias problem.

Case study. We conduct case studies in Appendix B. Results show that our Seq2Emo can attend to relevant words when predicting the emotion of interest.

	Seq2Emo	Seq2Emo (uni)	CC
Dataset order	58.67	58.22	58.16
Desending	58.42	58.23	57.86
Ascending	58.54	58.14	58.11
Concave-up	58.48	58.12	57.58
Concave-down	58.40	57.93	58.49
STD	0.110	0.120	0.341

Table 3: Analysis on the order of emotions. The results are the Jaccard accuracy on SemEval'18.

6 Conclusion

In this work, we propose Seq2Emo for multi-label emotion classification. Our approach implicitly models the relationship of different emotions in its bi-directional decoder, and is shown to be better than an individual binary relevance (BR) classifier. Our model does not suffer from the exposure bias problem and also outperforms the classifier chain (CC). In general, we achieve state-of-the-art performance for multi-emotion classification on the SemEval'18 and GoEmotions datasets (without using additional emotion labels).

Acknowledgments

We acknowledge the support of the Natural Sciences and Engineering Research Council of Canada (NSERC) under grant Nos. RGPIN-2020-04465 and RGPIN-2020-04440. Chenyang Huang is supported by the Borealis AI Graduate Fellowship Program. Lili Mou and Osmar Zaïane are supported by the Amii Fellow Program and the Canada CIFAR AI Chair Program. This research is also supported in part by Compute Canada (www.computecanada.ca).

References

Nabiha Asghar, Pascal Poupart, Jesse Hoey, Xin Jiang, and Lili Mou. 2018. Affective neural response generation. In *European Conference on Information Retrieval*, pages 154–166.

Christos Baziotis, Athanasiou Nikolaos, Alexandra Chronopoulou, Athanasia Kolovou, Georgios Paraskevopoulos, Nikolaos Ellinas, Shrikanth Narayanan, and Alexandros Potamianos. 2018. NTUA-SLP at SemEval-2018 Task 1: Predicting affective content in Tweets with deep attentive RNNs and transfer learning. In *Proceedings of the 12th International Workshop on Semantic Evaluation*, pages 245–255.

Samy Bengio, Oriol Vinyals, Navdeep Jaitly, and Noam Shazeer. 2015. Scheduled sampling for sequence prediction with recurrent neural networks. In *Advances in Neural Information Processing Systems*, pages 1171–1179.

Nancy Chinchor. 1992. MUC-4 evaluation metrics. In *Fourth Message Uunderstanding Conference*, pages 22–29.

Dorottya Demszky, Dana Movshovitz-Attias, Jeongwoo Ko, Alan Cowen, Gaurav Nemade, and Sujith Ravi. 2020. GoEmotions: A dataset of fine-grained emotions. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4040–4054.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4171–4186.

Paul Ekman. 1984. Expression and the nature of emotion. *Approaches to Emotion*, 3:19–344.

Bjarke Felbo, Alan Mislove, Anders Søgaard, Iyad Rahwan, and Sune Lehmann. 2017. Using millions of emoji occurrences to learn any-domain representations for detecting sentiment, emotion and sarcasm. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 1615–1625.

Deepanway Ghosal, Navonil Majumder, Soujanya Poria, Niyati Chhaya, and Alexander Gelbukh. 2019. DialogueGCN: A graph convolutional neural network for emotion recognition in conversation. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing*, pages 154–164.

Sayan Ghosh, Mathieu Chollet, Eugene Laksana, Louis-Philippe Morency, and Stefan Scherer. 2017. Affect-LM: A neural language model for customizable affective text generation. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 634–642.

Shantanu Godbole and Sunita Sarawagi. 2004. Discriminative methods for multi-labeled classification. In *Pacific-Asia Conference on Knowledge Discovery and Data Mining*, pages 22–30.

Chenyang Huang, Amine Trabelsi, and Osmar Zaïane. 2019. ANA at SemEval-2019 Task 3: Contextual emotion detection in conversations through hierarchical LSTMs and BERT. In *Proceedings of the 13th International Workshop on Semantic Evaluation*, pages 49–53.

Chenyang Huang, Osmar Zaïane, Amine Trabelsi, and Nouha Dziri. 2018. Automatic dialogue generation with expressed emotions. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 49–54.

Diederik P. Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *Proceedings of the International Conference on Learning Representations*.

Minh-Thang Luong, Hieu Pham, and Christopher D Manning. 2015. Effective approaches to attention-based neural machine translation. In *Proceedings of Conference on Empirical Methods in Natural Language Processing*, pages 1412–1421.

Hardik Meisheri and Lipika Dey. 2018. TCS research at SemEval-2018 Task 1: Learning robust representations using multi-attention architecture. In *Proceedings of the 12th International Workshop on Semantic Evaluation*, pages 291–299.

Mike Mintz, Steven Bills, Rion Snow, and Daniel Jurafsky. 2009. Distant supervision for relation extraction without labeled data. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing*, pages 1003–1011.

Saif Mohammad. 2012. #Emotional Tweets. In **SEM 2012: The First Joint Conference on Lexical and Computational Semantics*, pages 246–255.

Saif Mohammad. 2018. Word affect intensities. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation*, pages 174–181.

Saif Mohammad, Felipe Bravo-Marquez, Mohammad Salameh, and Svetlana Kiritchenko. 2018. SemEval-2018 Task 1: Affect in Tweets. In *Proceedings of the 12th International Workshop on Semantic Evaluation*, pages 1–17.

Saif M Mohammad and Svetlana Kiritchenko. 2015. Using hashtags to capture fine emotion categories from Tweets. *Computational Intelligence*, 31(2):301–326.

Ji Ho Park, Peng Xu, and Pascale Fung. 2018. PlusEmo2Vec at SemEval-2018 Task 1: Exploiting emotion knowledge from emoji and #hashtags. In *Proceedings of the 12th International Workshop on Semantic Evaluation*, pages 264–272.

Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. GloVe: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing*, pages 1532–1543.

Matthew Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2227–2237.

Robert Plutchik. 1980. A general psychoevolutionary theory of emotion. *Theories of Emotion*, pages 3–33.

Soujanya Poria, Alexander F. Gelbukh, Erik Cambria, Amir Hussain, and Guang-Bin Huang. 2014. EmoSenticSpace: A novel framework for affective common-sense reasoning. *Knowledge-Based Systems*, 69:108–123.

Alec Radford, Rafal Jozefowicz, and Ilya Sutskever. 2017. Learning to generate reviews and discovering sentiment. *arXiv preprint arXiv:1704.01444*.

Jesse Read, Bernhard Pfahringer, Geoff Holmes, and Eibe Frank. 2009. Classifier chains for multi-label classification. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 254–269.

David J Rogers and Taffee T Tanimoto. 1960. A computer program for classifying plants. *Science*, 132(3434):1115–1118.

Kashfia Sailunaz, Manmeet Dhaliwal, Jon G. Rokne, and Reda Alhajj. 2018. Emotion detection from text and speech: A survey. *Social Network Analysis Mining*, 8(1):28:1–28:26.

Victor Sanh, Thomas Wolf, and Sebastian Ruder. 2019. A hierarchical multi-task approach for learning embeddings from semantic tasks. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 6949–6956.

Robert E. Schapire and Yoram Singer. 1999. Improved boosting algorithms using confidence-rated predictions. *Machine Learning*, 37(3):297–336.

Klaus R Scherer and Harald G Wallbott. 1994. Evidence for universality and cultural variation of differential emotion response patterning. *Journal of Personality and Social Psychology*, 66(2):310.

Ameneh Gholipour Shahraki and Osmar R. Zaiane. 2017. Lexical and learning-based emotion mining from text.

Carlo Strapparava and Alessandro Valitutti. 2004. WordNet Affect: an affective extension of WordNet. In *Proceedings of the Fourth International Conference on Language Resources and Evaluation*, pages 1083–1086.

Duyu Tang, Furu Wei, Nan Yang, Ming Zhou, Ting Liu, and Bing Qin. 2014. Learning sentiment-specific word embedding for Twitter sentiment classification. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1555–1565.

Ryoko Tokuhisa, Kentaro Inui, and Yuji Matsumoto. 2008. Emotion classification using massive examples extracted from the Web. In *Proceedings of the 22nd International Conference on Computational Linguistics*, pages 881–888.

Shiyang Wen and Xiaojun Wan. 2014. Emotion classification in microblog texts using class sequential rules. In *Proceedings of the Twenty-Eighth AAAI Conference on Artificial Intelligence*, pages 187–193.

Ali Yadollahi, Ameneh Gholipour Shahraki, and Osmar R Zaiane. 2017. Current state of text sentiment analysis from opinion to emotion mining. *ACM Computing Surveys*, 50(2):1–33.

Pengcheng Yang, Xu Sun, Wei Li, Shuming Ma, Wei Wu, and Houfeng Wang. 2018. SGM: Sequence generation model for multi-label classification. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 3915–3926.

Jianfei Yu, Luís Marujo, Jing Jiang, Pradeep Karuturi, and William Brendel. 2018. Improving multi-label emotion classification via sentiment classification with dual attention transfer network. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1097–1102.

Hao Zhou, Minlie Huang, Tianyang Zhang, Xiaoyan Zhu, and Bing Liu. 2018. Emotional chatting machine: Emotional conversation generation with internal and external memory. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence*, pages 730–739.

A Dataset Statistics

Table 4 shows the statistics of both SemEval'18 and GoEmotions datasets. Noticeably, the majority of the data samples in SemEval'18 are labeled with at least two emotions. The GoEmotions dataset is mostly annotated with one label for an utterance, although multiple emotions do exist. This suggests that SemEval'18 may contain more correlated emotions on average.

Dataset	# emo.	# sample	% multi-emo.	# avg. emo.
SemEval'18	11	10690	86.1	2.37
GoEmotions	24	54263	16.2	1.17

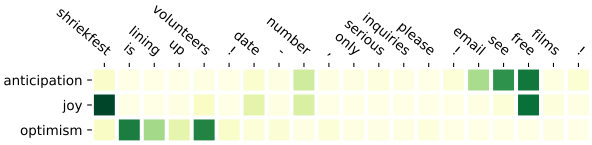
Table 4: Data statistics: the number of the emotion categories, the number of data samples, the percentage of multi-labeled samples, and the average number of emotions per utterance.

B Case Study

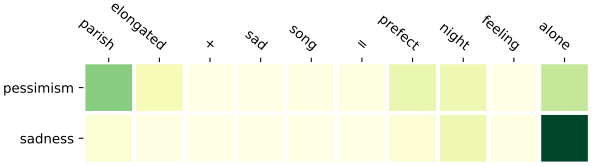
In Figure 2, we visualize the attention layer of Seq2Emo by plotting the heat map over the attention scores. The emotions shown in each example are the groundtruth labels of the corresponding utterance.

We observe that Seq2Emo is able to focus on relevant words when predicting the emotion of interest. In Case 3, for example, the emotions *joy* and *love* highly resemble each other, both focusing on the word "laughter." On the other hand, the decoder of Seq2Emo can focus on entirely different words if the emotions are different. In Case 1, we see the emotion *anticipation* mainly focuses on "see free," whereas the emotion *optimism* mainly focuses on "is lining up volunteers."

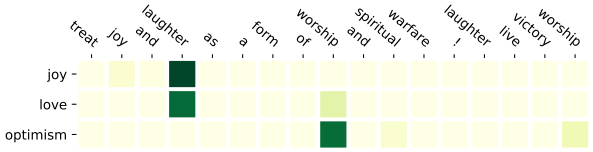
Case 1: shriekfest is lining up volunteers ! date - number , only serious inquiries please ! email see free films !



Case 2: parish elongated + sad song = prefect night feeling alone



Case 3: treat joy and laughter as a form of worship and spiritual warfare ! laughter live victory worship



Case 4: user ' operation echoes ' is gathering momentum . . . tense feel sick excited

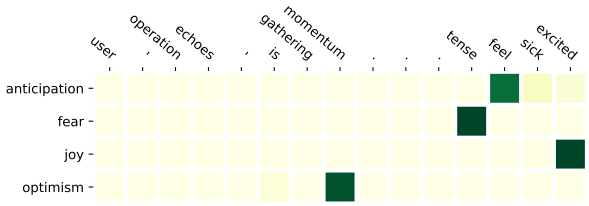


Figure 2: Case study.