A Unified Framework for Emotion Identification and Generation in Dialogues

Avinash Madasu[*†] **Mauajama Firdaus**[*] **Asif Ekbal**
Indian Institute of Technology Patna, India
avinashmadasu17@gmail.com; {mauajama.pcs16,asif}@iitp.ac.in

Abstract

Social chatbots have gained immense popularity, and their appeal lies not just in their capacity to respond to the diverse requests from users, but also in the ability to develop an emotional connection with users. To further develop and promote social chatbots, we need to concentrate on increasing user interaction and take into account both the intellectual and emotional quotient in the conversational agents. In this paper, we propose a multi-task framework that jointly identifies the emotion of a given dialogue and generates response in accordance to the identified emotion. We employ a BERT based network for creating an empathetic system and use a mixed objective function that trains the end-to-end network with both the classification and generation loss. Experimental results show that our proposed framework outperforms current state-of-the-art models.

1 Introduction

One of the significant challenges of artificial intelligence (AI) is to endow the machine with the ability to interact in natural language with humans. The personal assistants in our mobile devices invariably assist in our day-to-day lives by answering a wide range of queries. Such assistants act as social agents that take care of the various activities of their users. Besides reacting passively to user requests, they also proactively anticipate user needs and provide in-time assistance, such as reminding of an upcoming event or suggesting a useful service without receiving explicit user requests (Sarikaya, 2017). The daunting task for these agents is that they have to work well in open domain scenarios as people learn to rely on them to effectively maintain their works and lives efficiently.

Building social chatbots to tackle the emotional needs is indeed of great benefit to our society. The

primary objective of these chatbots is not inherently to answer all the users' questions, but rather to be a virtual companion of the users. Therefore, it is essential to empower the conversational agents with the ability to perceive and express emotions to make them capable of interacting with the user at the human level. These agents help enhance user satisfaction (Prendinger et al., 2005), while reducing the breakdowns in conversations (Martinovski and Traum, 2003) and providing user retention. Hence, dialogue systems capable of generating replies while considering the user's emotional state is the most desirable advancement in Artificial Intelligence (AI). Previously, researchers have focused upon classifying user emotions (Poria et al., 2019; Chauhan et al., 2019) in conversations. For building an intelligent agent, sheer understanding of emotions is insufficient; hence several works (Song et al., 2019; Colombo et al., 2019) have concentrated in inducing emotions into the dialogue system. Most of these existing research have focused on generating emotionally aware (Rashkin et al., 2019; Lin et al., 2019) or emotionally controlled responses (Zhou et al., 2018; Huang et al., 2018).

In our current work, we focus on creating an end-to-end emotional response generation system that is capable of identifying the emotions and use the emotional information simultaneously for generating empathetic and emotionally coherent responses. The key contributions of our current work are threefold: (i) We propose a joint model that is able to simultaneously identify the emotions from the utterances and incorporate the emotional information for generation; (ii) We design a multi-task hierarchical framework with a mixed objective function that jointly optimize the classification and generation loss; (iii) Experimental analysis shows that the proposed multi-task framework is better in generating empathetic responses as opposed to single task emotion generation networks.

---

[*]Both authors contributed equally to this research.
[‡]Work done as an intern at IIT Patna.

73

2 Related Work

In (Welivita et al., 2020), the authors proposed a novel large-scale emotional dialogue dataset, consisting of 1M dialogues and annotated with 32 emotions and 9 empathetic response intents using a BERT-based fine-grained dialogue emotion classifier. Most of the early research on emotion classification was performed upon textual datasets mostly taken from twitter (Colneriĉ and Demsar, 2018; Ghosal et al., 2018). The authors in (Chen et al., 2018), propose a multi-party conversational dataset for emotions. Lately, emotional text generation has gained immense popularity (Huang et al., 2018; Li and Sun, 2018; Lin et al., 2019; Li et al., 2017; Ghosh et al., 2017; Rashkin et al., 2019). In (Zhou et al., 2018), an emotional chatting machine (ECM) was built based on seq2seq framework for generating emotional responses. Recently, a lexicon-based attention framework was employed to generate responses with a specific emotion (Song et al., 2019). Emotional embedding, affective sampling and regularizer were employed to generate the affect driven dialogues in (Colombo et al., 2019). The authors employed curriculum dual learning (Shen and Feng, 2020) for emotion controllable response generation. In (Asghar et al., 2018; Lubis et al., 2018; Zhong et al., 2019; Li et al., 2020), the authors proposed an end-to-end neural framework that captures the emotional state of the user for generating empathetic responses. Our present research differs from these as we propose and address a novel task of generating responses and simultaneously identifying the emotions in a multi-task framework.

3 Problem Definition

In this paper, we address the problem of identifying emotion from a user utterance and generate empathetic responses accordingly. In this setting, emotion remains the same throughout the conversation. Let $U_p = u_{p,1}, u_{p,2}, ..., u_{p,j}$ be the set of utterances in a conversation and E_p denotes the emotion for the conversation. We aim to identify E_p through our emotion classification sub-network and use this emotion information to generate responses to the user utterances U_p. Understanding emotion initially is very crucial to empathetic response generation and cannot be treated independently. Hence, we aim to propose an end-to-end model capable to identifying emotions and utilizes this emotion information for generation.
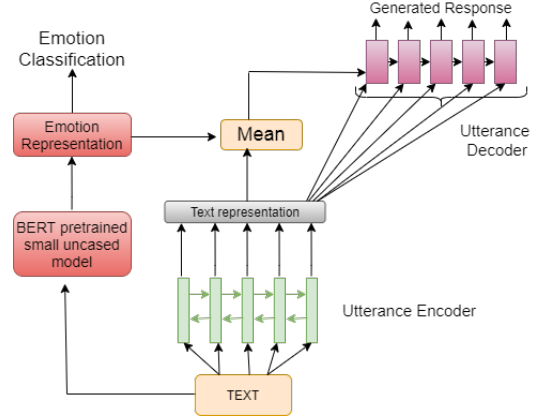

Figure 1: Architecture of the proposed model

4 Proposed Model

In this section, we present the proposed model. It has two sub-networks: Emotion classification sub-network and Generation sub-network.

4.1 Emotion classification sub-network

The input to this sub-network is a user utterance and the output is predicted emotion for this utterance. We employed a pre-trained BERT (Devlin et al., 2019) small uncased model. Instead of taking $[CLS]$ token from just the final layer, we learn a weighted representation on $[CLS]$ tokens from all the layers.

$$h_{cls}^l = \text{BERT}(U_p) l \in \{1, \ldots, 12\}$$
$$C_E = \sum_{i=1}^{l} \alpha_{t,i} h_{cls}^l \quad (1)$$
$$\alpha_{t,i} = softmax(h_{cls}^{l\ T} W_g h_{cls,t})$$
$$P(\tilde{y_{em}} \mid y_{em}) = softmax(W_E C_E),$$

y_{em} is the true emotion labels and $\tilde{y_{em}}$ is the predicted emotion label. The emotion classifier is trained by minimizing the negative log-likelihood

$$\mathcal{L}_{Emo} = - \sum_{em=1}^{N} y_{em} \log \tilde{y_{em}} \quad (2)$$

4.2 Generation sub-network

Utterance Encoder For a given utterance U_p, we employ a bidirectional Gated Recurrent Units (Bi-GRU) (Cho et al., 2014) to encode each word $u_{p,j}$, where $j \in (1, 2, 3, ...n)$ having d-dimensional embedding vectors into the hidden representation $h_{U_k,i}$. We concatenate the last hidden representation from both the unidirectional GRUs to form the

final hidden representation of a given utterance as follows:

$$h_{U_p,i}^{en} = [GRU(u_{p,j}, \overrightarrow{h_{U_p,i-1}}), GRU(u_{p,j}, \overleftarrow{h_{U_p,i-1}})] \quad (3)$$

Utterance Decoder For decoding the response to a given user utterance, we build a GRU which takes encoder last hidden state as the initial hidden state and words generated previously. To integrate emotion information, we compute the mean of emotion representation obtained using C_E and the encoder's last hidden state to initialize as decoder's initial state. We use the input feeding decoding along with the attention (Luong et al., 2015) mechanism for enhancing the performance of the model. Using the decoder state $h_{d,t}^{dec}$ as the query vector, we apply self-attention on the hidden representation of the utterance encoder. The decoder state and the encoder vector are concatenated and used to calculate a final distribution of the probability over the output tokens.

$$h_{d,t-1} = Mean(C_E, h_{c,t}^{en})$$
$$h_{d,t}^{dec} = GRU_d(y_{k,t-1}, h_{d,t-1})$$
$$c_t = \sum_{i=1}^{k} \alpha_{t,i} h_{c,p}^{ctx}, \quad (4)$$
$$\alpha_{t,i} = softmax(h_{c,p}^{en\ T} W_f h_{d,t})$$
$$\tilde{h}_t = tanh(W_{\tilde{h}}[h_{d,t}; c_t])$$
$$P(\tilde{y}_t \mid y_t^*) = softmax(W_V \tilde{h}_t)$$

where, W_f, W_V and $W_{\tilde{h}}$ are the trainable weight matrices. y_t^*, $\tilde{y}_t$ are the ground truth and generated words at each time-step respectively. We employ the teacher forcing (Williams and Zipser, 1989) algorithm at every decoding step to minimize the negative log-likelihood on the model distribution.

$$\mathcal{L}_{Gen} = -\sum_{t=1}^{m} \log p(y_t^* | y_1^*, \ldots, y_{t-1}^*) \quad (5)$$

4.3 Joint Training (JT)

To train an end-to-end model, we jointly optimize the emotion classification loss and the generation loss. The final loss of the entire model is:

$$\mathcal{L}_{JT} = \mathcal{L}_{Emo} + \mathcal{L}_{Gen} \quad (6)$$

5 Dataset and Experimentation

Dataset We conduct our experiment on the EmpatheticDialogues (Rashkin et al., 2019) dataset

which consist of 25k open-domain conversations grounded in emotional situations and provides 32 emotion classes. We used the train, test splits provided by the authors.

Implementation Details The hidden size of utterance encoder Bi-GRU is 768 and the hidden size of utterance decoder GRU is 768. A dropout of 0.3 is applied on both the utterance encoder and decoder layers. A dropout of 0.1 is applied on the weighted emotion representation obtained from 1 just before the final softmax layer. All the models are trained with a batch size of 32 for 10 epochs. AdamW is used as the optimizer with a learning rate of 0.0001. The maximum sentence length used is 80.

6 Baseline Methods and Metrics

We compare our model to the below SoTA models.
Pretrained:
It is a transformer model trained on 1.7 billion REDDIT conversations (Lin et al., 2019).
Fine-Tuned:
The above pre-trained model is fine-tuned on Emotion Dialogue Dataset (Lin et al., 2019).
MULTITASK:
It is an emotion classifier trained using a linear layer on the top of a transformer (Lin et al., 2019).
EmoPrepend-1:
A top-1 predicted emotion is appended to the beginning of input sentence (Lin et al., 2019).
ENSEM-DM:
The representations obtained from transformer encoder and pre-trained emotion classifier are concatenated and then fed to transformer decoder (Lin et al., 2019).
CAiRE:
It is the current State-of-the-art model. In this model, a large language model is jointly trained for multiple objectives response language modeling, response prediction, and dialogue emotion detection. It is then fine-tuned on EmphatheticDialogues dataset (Lin et al., 2019).
Bi-LSTM-Attn (JT):
To signify the importance of a BERT pre-trained model, we replace BERT with Bi-LSTM layer with attention applied on the top of it.
Metrics:
We use AVG BLEU, the average of BLEU-1, BLEU-2, BLEU-3, BLEU-4 (Papineni et al., 2002) and emotion F1 for comparison. As the dataset is unbalanced, we didn't use emotion accuracy for models' comparison. We used balanced emotion

labels during manual evaluation and hence we included emotion accuracy.

Table 1: Automatic evaluation results of emotion identification and generation. Here: JT-joint training

Models	AVG BLEU	EMO F1
Pretrained (Lin et al., 2019)	5.01	-
Fine-Tuned (Lin et al., 2019)	6.27	-
MULTITASK (Lin et al., 2019)	5.42	-
EmoPrepend-1 (Lin et al., 2019)	4.36	-
ENSEM-DM (Lin et al., 2019)	6.83	-
CAiRE (Lin et al., 2019)	7.03	-
Bi-LSTM (JT)	6.84	8.22
Bi-LSTM + Attn (JT)	6.13	19.89
BERT (JT)	**7.71**	**25.2**

7   Result and Analysis

In this section, we present the results of our proposed framework. We compute average BLEU scores for the model response, comparing against the gold label (the actual response). From the Table 1, it is evident that our proposed framework performs significantly better in comparison to the existing baselines[1]. As compared to CAiRE, we see an improvement of 0.68 in average BLEU score. Emotion identification is crucial for relevant empathetic responses and the generated responses should be in accordance to sentiment expressed by the user. Joint training helps in learning better emotion representations thereby generating contextually coherent, interactive, engaging and empathetic responses. The emotion classification results reported in Table 1, demonstrates effectiveness of joint training. From table 1, it is visible that the F1-score of BERT based classifier is better in comparison to the joint training Bi-LSTM and Bi-LSTM + Attn with a gain in performance of 17% and 6% respectively.

We recruit six annotators (in a similar manner as (Shang et al., 2015; Tian et al., 2019)) from a third party company, having high-level language skills. In Table 2 we present the results of human evaluation. We sampled 250 responses per model for evaluation with the utterance and the conversational history provided for generation. First, we evaluate the quality of the response on two conventional criteria: *Fluency* and *Consistency*. These are rated

---

[1]all the results are statistically significant. We perform statistical significance t-test (Welch, 1947), and it is conducted at 5% (0.05) significance level

Table 2: Human evaluation results of emotion identification and generation. Here: Flu-Fluency; Cons-Consistency; EmoAcc-Emotion Accuracy

Model	Flu	Cons	EmoAcc
Bi-LSTM (JT)	3.82	3.73	48%
Bi-LSTM + Attn (JT)	3.93	3.82	55%
BERT (JT)	**4.11**	**3.96**	**69%**

on a five-scale, where 1, 3, 5 indicate unacceptable, moderate, and excellent performance, respectively, while 2 and 4 are used for unsure. Secondly, we evaluate the emotion quotient of a response in terms of *Emotion Accuracy* metric that measures whether the emotion induced in the response is in accordance with the predicted emotion information and the dialogue history. Here, 0 indicates irrelevant or contradictory, and 1 indicates consistent with the predicted emotion and dialogue context. We compute Fleiss' kappa (Fleiss, 1971) to measure inter-rater consistency. The Fleiss' kappa for Fluency and Consistency is 0.53 and 0.49, indicating moderate agreement. For Emotion Accuracy, we get 0.65 as the kappa scores indicating substantial agreement. From the table, it is evident that our proposed framework performs better for all the human evaluation metrics. The responses generated are fluent, consistent to the dialogue history and the emotion quotient of the generated response is higher in comparison to the baselines. Also, the joint training mechanism proves to be efficient for simultaneously identifying the emotions and using the emotional information for generation.

8   Conclusion and Future Work

In our current work, we propose the task of jointly identifying the emotions in dialogues and use the emotional information for generating empathetic responses. For our proposed task, we design a BERT-based multi-task framework, that simultaneously identifies the emotion of the speaker and generate the response in accordance to the situation, conversational history and the predicted emotions. Experimental analysis on the EmpatheticDialogues dataset shows that the proposed framework achieved State-of-the-art results. In future, we look forward to designing sophisticated joint mechanisms to enhance empathetic response generation.

References

Nabiha Asghar, Pascal Poupart, Jesse Hoey, Xin Jiang, and Lili Mou. 2018. Affective neural response generation. In *Advances in Information Retrieval - 40th European Conference on IR Research, ECIR 2018, Grenoble, France, March 26-29, 2018, Proceedings*, pages 154–166. Springer.

Dushyant Singh Chauhan, Md Shad Akhtar, Asif Ekbal, and Pushpak Bhattacharyya. 2019. Context-aware interactive attention for multi-modal sentiment and emotion analysis. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 5651–5661.

Sheng-Yeh Chen, Chao-Chun Hsu, Chuan-Chun Kuo, Lun-Wei Ku, et al. 2018. Emotionlines: An emotion corpus of multi-party conversations. *arXiv preprint arXiv:1802.08379*.

Kyunghyun Cho, Bart Van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. 2014. Learning phrase representations using rnn encoder-decoder for statistical machine translation. *arXiv preprint arXiv:1406.1078*.

Niko Colnerič and Janez Demsar. 2018. Emotion recognition on twitter: Comparative study and training a unison model. *IEEE Transactions on Affective Computing*, 11(3):433–446.

Pierre Colombo, Wojciech Witon, Ashutosh Modi, James Kennedy, and Mubbasir Kapadia. 2019. Affect-driven dialog generation. *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, pages 3734–3743.

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 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Joseph L Fleiss. 1971. Measuring nominal scale agreement among many raters. *Psychological bulletin*, 76(5):378.

Deepanway Ghosal, Md Shad Akhtar, Dushyant Chauhan, Soujanya Poria, Asif Ekbal, and Pushpak Bhattacharyya. 2018. Contextual inter-modal attention for multi-modal sentiment analysis. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Brussels, Belgium, October 31 - November 4, 2018*, pages 3454–3466.

Sayan Ghosh, Mathieu Chollet, Eugene Laksana, Louis-Philippe Morency, and Stefan Scherer. 2017. Affect-lm: A neural language model for customizable affective text generation. *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, ACL 2017, Vancouver, Canada, July 30 - August 4, Volume 1: Long Papers*, pages 634–642.

Chenyang Huang, Osmar R Zaiane, 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, NAACL-HLT, New Orleans, Louisiana, USA, June 1-6, 2018, Volume 2 (Short Papers)*, pages 49–54.

Jingyuan Li and Xiao Sun. 2018. A syntactically constrained bidirectional-asynchronous approach for emotional conversation generation. *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Brussels, Belgium, October 31 - November 4, 2018*, pages 678–683.

Qintong Li, Hongshen Chen, Zhaochun Ren, Zhumin Chen, Zhaopeng Tu, and Jun Ma. 2020. Empgan: Multi-resolution interactive empathetic dialogue generation. *AAAI*.

Yanran Li, Hui Su, Xiaoyu Shen, Wenjie Li, Ziqiang Cao, and Shuzi Niu. 2017. Dailydialog: A manually labelled multi-turn dialogue dataset. *Proceedings of the Eighth International Joint Conference on Natural Language Processing, IJCNLP 2017, Taipei, Taiwan, November 27 - December 1, 2017 - Volume 1: Long Papers*, pages 986–995.

Zhaojiang Lin, Peng Xu, Genta Indra Winata, Zihan Liu, and Pascale Fung. 2019. Caire: An end-to-end empathetic chatbot. *arXiv preprint arXiv:1907.12108*.

Nurul Lubis, Sakriani Sakti, Koichiro Yoshino, and Satoshi Nakamura. 2018. Eliciting positive emotion through affect-sensitive dialogue response generation: A neural network approach. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, (AAAI-18), New Orleans, Louisiana, USA, February 2-7, 2018*, pages 5293–5300.

Minh-Thang Luong, Hieu Pham, and Christopher D Manning. 2015. Effective approaches to attention-based neural machine translation. *arXiv preprint arXiv:1508.04025*.

Bilyana Martinovski and David Traum. 2003. Breakdown in human-machine interaction: The error is the clue. In *Proceedings of the ISCA tutorial and research workshop on Error handling in dialogue systems*, pages 11–16.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics, July 6-12, 2002, Philadelphia, PA,*

USA, pages 311–318. Association for Computational Linguistics.

Soujanya Poria, Devamanyu Hazarika, Navonil Majumder, Gautam Naik, Erik Cambria, and Rada Mihalcea. 2019. MELD: A multimodal multi-party dataset for emotion recognition in conversations. *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*, pages 527–536.

Helmut Prendinger, Junichiro Mori, and Mitsuru Ishizuka. 2005. Using human physiology to evaluate subtle expressivity of a virtual quizmaster in a mathematical game. *International journal of human-computer studies*, 62(2):231–245.

Hannah Rashkin, Eric Michael Smith, Margaret Li, and Y-Lan Boureau. 2019. Towards empathetic open-domain conversation models: A new benchmark and dataset. In *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*, pages 5370–5381.

Ruhi Sarikaya. 2017. The technology behind personal digital assistants: An overview of the system architecture and key components. *IEEE Signal Processing Magazine*, 34(1):67–81.

Lifeng Shang, Zhengdong Lu, and Hang Li. 2015. Neural responding machine for short-text conversation. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, volume 1, pages 1577–1586.

Lei Shen and Yang Feng. 2020. Cdl: Curriculum dual learning for emotion-controllable response generation. *arXiv preprint arXiv:2005.00329*.

Zhenqiao Song, Xiaoqing Zheng, Lu Liu, Mu Xu, and Xuan-Jing Huang. 2019. Generating responses with a specific emotion in dialog. In *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*, pages 3685–3695.

Zhiliang Tian, Wei Bi, Xiaopeng Li, and Nevin L Zhang. 2019. Learning to abstract for memory-augmented conversational response generation. In *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*, pages 3816–3825.

Bernard L Welch. 1947. The generalization of student's' problem when several different population variances are involved. *Biometrika*, 34(1/2):28–35.

Anuradha Welivita, Yubo Xie, and Pearl Pu. 2020. Fine-grained emotion and intent learning in movie dialogues. *arXiv preprint arXiv:2012.13624*.

Ronald J Williams and David Zipser. 1989. A learning algorithm for continually running fully recurrent neural networks. *Neural computation*, 1(2):270–280.

Peixiang Zhong, Di Wang, and Chunyan Miao. 2019. An affect-rich neural conversational model with biased attention and weighted cross-entropy loss. In *The Thirty-Third AAAI Conference on Artificial Intelligence, AAAI 2019, The Thirty-First Innovative Applications of Artificial Intelligence Conference, Honolulu, Hawaii, USA, January 27 - February 1, 2019*, pages 7492–7500.

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, (AAAI-18), New Orleans, Louisiana, USA, February 2-7, 2018*, pages 730–739.

A Example Appendix

This is a section in the appendix.