Emotionally Aligned Responses through Translation

Anonymous ACL submission

Abstract

Emotional response generation is an area of particular interest within conversational AI. However, many approaches lack control over the response. Potentially, due in part to the widely adopted approach of reflecting the users emotion in the response. As such this paper proposes an independent, but adaptable, emotion for the conversational agent that is separate from the user's, using Valence, Arousal, and Dominance scores which are updated based on user input. Additionally, by treating the alignment of the response as a matter of translation, a set of fine tuned sequence to sequence models are used to translate an initially generated response into one aligned with the agent emotion. This work provides a unique perspective on the topic of emotional response generation and showcases that potential means for improved consistency and controlability may yet be discovered beyond traditional methods.

1 Introduction

Producing dialogue possessing emotion has been a topic of focus within generative artificial intelligence in the last few years. This prevalence is due in large part to the fact that emotion and emotional understanding are crucial in human conversation, and subsequently for the success of a conversational agent(Rashkin et al., 2019; Zhou et al., 2018; Liu et al., 2022). Moreover, such emotional capability can increase the perception of friendliness and intelligence of an agent(Wang and Wan, 2018) and is often expected by the user(Ensi Chen and Wang, 2022; Fung et al., 2016).

Emotional responses, sometimes referred to as empathetic responses, usually incorporate emotional information from the users utterances in order to create an aligned response. Many of these approaches, rely singularly on the emotion conveyed in the user input, following the empathy strategy that the input and response should have consistent emotions(Qian et al., 2023).

Additionally, with conversational generation, it can be difficult to maintain control over the emotion of the generated response (Zhou and Wang, 2018) with emotional consistency also being noted as necessary for creating an appropriate response (Zhou et al., 2018). Sensitivity to input(Hamad et al., 2024) and a lack of control over responses (Li et al., 2024) have additionally been encountered in conversational generation and subsequently can be anticipated in emotional generation.

The contribution of this paper is twofold. Firstly, an investigation into the use of Valence, Arousal, and Dominance scores to provide a consistent but adaptable emotion for the agent, and secondly, presenting emotional alignment as a matter of translation between an initial utterance and one in the targeted emotion. The approach, dubbed ETC (Emotional Translation for Consistency), is outlined in the following sections, including the implementation which is made avaliable on GitHub [1] and experiments conducted, as well as a discussion of the outcomes.

2 Similar Works

Many works in the area of emotional generated dialogue responses follow the principle that the produced response should reflect or be similar to the emotion of the user (Ensi Chen and Wang, 2022) or uses the users emotion in order to determine the emotion of the agents response (Liu et al., 2022; Pang et al., 2024). A notable example of which is (Majumder et al., 2020) MIME, which argued that the degree to which the response should mimic the users, emotion depended on the positivity or negativity of the utterance. Gao et al. (2023) conversely, employed dual latent variables to capture the emotions of both interlocutors, which then influences the generated response.

[1] https://github.com/codesubmissionanon112/ACLAnonSubmission

Yang et al. (2024) meanwhile, used smaller models to help large language models improve emotional understanding, using an emotion prediction strategy. Shin et al. (2021) Not only investigates creating emotional responses, but also uses reinforcement learning to reward the model if the users emotion improves. Shin et al. (2021) is not the only work to extend beyond a one-to-one emotional response generation. It is also becoming increasingly common to encounter approaches that extend beyond emotional integration often with common sense information (Ensi Chen and Wang, 2022; Sabour et al., 2022) or dialogue and emotional history being leveraged to generator response (Cai et al., 2024).

MoEL (Lin et al., 2019) introduced a framework where an emotion distribution was created based on the user emotion, which different listeners optimized to certain emotions used to create output states. These output states were then combined in order to create an empathetic response. While ETC bears some similarities to MoEL (Lin et al., 2019), as both employ emotion specific modifiers to the generation, ETC differs in that the emotion of the agent is tracked through numeric scores. Additionally, fine-tuned models were used for ETC instead of independent decoders for each emotion. The notion of modifying initially generated response to better align with a target emotion has also been investigated in Qian et al. (2023). Despite these similarities, ETC presents a unique approach with the potential of improving consistency and controllability, by maintaining a separate agent emotion.

3 Methods

At runtime the ETC framework begins by handing the user input to both the Multioutput Regression Model (MRM), which supplies the input's VAD value, and the initial response generator. The input's VAD score is averaged with that of the agent to provide the agents updated VAD score. Nearest neighbor comparison (Buitinck et al., 2013) is then used to determine which of the five emotions was most similar to the agent score. The selected emotion then indicates which of the five fine-tuned models to use. The initially generated response is then handed to the selected translator to produce the translated response. Figure 1 provides a visual overview of the framework.

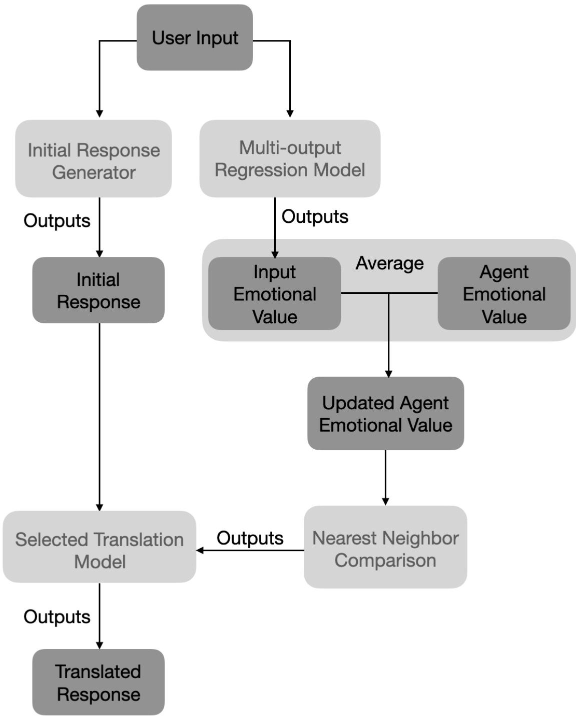

Figure 1: The user input's VAD score is averaged with the agents to provide an updated agent score for that turn of the conversation, on the start of a new conversation, the agent's score is reset to calm.

3.1 Managing The Emotion Values

Valence, Arousal, and Dominance scores(Russell and Mehrabian, 1977) provide means for representing emotion in a three-dimensional space. These values were selected over labels in order to record the agents emotional value with more precision than that of a label. Additionally, by using the scores it is possible to average them with the score for the input, allowing for incremental modification that is reactionary without completely discarding the agents emotion from the previous turn.

The VAD scores generated at runtime were produced by the MRM, which accepted text and returned a VAD score, the idea for which was inspired by Park et al. (2021). The model was created by using a MultiOutput Regressor wrapper on a Gradient Boosting Regressor(Pedregosa et al., 2011; Buitinck et al., 2013), which applies a regressor per target which, in this case corresponded to V, A, and D.

The data sets used for training the MRM were EmoBank(Buechel and Hahn, 2017a,b) and NRC VAD Lexicon(Mohammad, 2018). EmoBank provides utterances and associated VAD scores, while the NRC VAD Lexicon provides individual words and their associated scores. The combination of the two data sets somewhat reduced the likelihood that the vectoriser used with the model would encounter completely unseen words. Additionally, while it is not uncommon to encounter data sets that supply text and an associated emotion like Gupta et al. (2018) and Rashkin et al. (2019) the selected data sets are some of the few that use VAD values to indicate the emotion.

In order to handle the agents VAD value, at the beginning of each conversation, it would be reset to calm, the value for which is [.875, .1, .282] in the NRC VAD Lexicon data set (Mohammad, 2018) . Following an input, the agents value was then averaged together with the value assigned to the user input by the MRM. This allows for the agents emotion to adapt to the conversation, while maintaining consistency and somewhat reducing sensitivity to input.

3.2 Emotionally Aligned Response Generation

In the 1970s Ekman and Friesen (1971) proposed that there were six main labels chosen to explain emotional facial expression. These emotions are anger, disgusted, fear, happiness, sadness, and surprise, now often referred to as Ekman's universal basic emotions Wortman (2024). These emotions were selected for the translators to reduce complexity.

Facebook's BlenderBot (Roller et al., 2021) 400 million parameter distill model [2] provided the initial generated responses. BlenderBot was selected as it is capable of producing human like responses, and contains no explicit emotional handling beyond fine-tuning on the EmpatheticDialogues dataset(Rashkin et al., 2019). This provides the opportunity to modify those responses to instead align with the emotion of the agent without attempting to override the efforts of an emotion module.

The initial response was then supplied to the selected translator model based on which of the five emotions was most similar to that of the agent emotion after adjustment based on the input. Each of the translation models was a fine tuned GODEL(Peng et al., 2022) seq2seq model[3]. This model was selected as sequence to sequence models are widely used in machine translation.

[2] https://huggingface.co/facebook/blenderbot-400M-distill

[3] https://huggingface.co/microsoft/GODEL-v1_1-base-seq2seq

The data used to find tune these models was a combination of the EmpatheticDialogues(Rashkin et al., 2019) and Emotion DatasetSaravia et al. (2018). As the EmpatheticDialogues data contained disgust but the Emotion Dataset did not, it was ultimately omitted in order to reduce an imbalance in the amount of data for each emotion. The EmpatheticDialogues data set was selected as it is widely used both for training and benchmarking (Lin et al., 2019; Sabour et al., 2022; Cai et al., 2024; Yang et al., 2024).

As no data set existed which contained an utterance with the same meaning in two different emotions, for each of the five selected emotions, a subset of the data set was created containing only the target emotion. Each utterance in the subset was then compared against the rest of the data to find the most semantically similar utterance in a different emotion using Semantic Textual Similarity from Reimers and Gurevych (2019). The utterance in a different emotion was then used as the initial value and the target emotion utterance as the target for the fine-tuning.

Though in other endeavors manipulating the data like this would generally be somewhat inadvisable, for the purpose of this project, the intention was to use the fine-tuned models to ensure that the emotional response contains words more frequently seen with the target emotion. Therefore, even if the utterances are not terribly similar in regards to their content, the models should still learn to use the emotionally coded words for the target emotion more frequently.

4 Experiments and Results

The validation subset of the dialogsum dataset (Chen et al., 2021) was used, with conversations containing more than two interactants or an unbalanced number of utterances being omitted, the responses were also cleaned of padding tokens prior to evaluation.

Approach	BLEU	METEOR	ROUGE L
ETC	0.007236	0.11122	0.11140
Untuned	0.00477	0.10664	0.10216
BlenderBot	0.01580	0.17393	0.12986

Table 1: The untuned model is a GODEL seq2seq model, which is the same type as the emotional translators.

It has been established that automated transla-

Approach	Utterance 1	Utterance 2
Initial utterance	I really want to take a nap. I feel very sleepy today.	I fell asleep very late. It was almost two o'clock in the morning when I finally fell asleep.
Data set response	What's the matter? Didn't you get enough sleep last night?	Are you worried about something? Why couldn't you sleep?
ETC	i feel like i am a little tired from work	I was so tired I had to go to sleep.
Untuned	I like to nap at night. I'm usually up at 5 or 6 in the morning	I did, but I was so tired I couldn't sleep.
BlenderBot	I love a good nap, especially after a long day of work. What time do you want to nap?	Oh no! I hate when that happens. Did you get up and go back to sleep?

Table 2: The initial utterance and data set response are from the dialogsum train subset.

Approach	Utterance 1 VAD	Utterance 2 VAD
Agent Emotion	[0.71126, 0.33504, 0.42927]	[0.64523, 0.43851, 0.51128]
ETC	[0.51598, 0.51605, 0.51834]	[0.55561, 0.57950, 0.57430]
Untuned	[0.59914, 0.62133, 0.62106]	[0.59674, 0.61308, 0.63007]
BlenderBot	[0.59914, 0.62133, 0.62106]	[0.59674, 0.61308, 0.63007]

Table 3: While the proposed approach is unique, both the Untuned and BlenderBot VAD scores are identical.

tion metrics such as do not align well with human judgment in regards to evaluating machine generated text, and that such means should only be used alongside human evaluation (Liu et al., 2016), however, this is not always an available option.

The ETC was compared with BlenderBot and BlenderBot responses sent through an untuned sequence to sequence model (noted as untuned in the tables). BlenderBot was used to provide insight on how the translation modified an initial response, and the responses modified by an untuned seq2seq model were to determine whether the translational fine-tuning had an impact on the sort of words used in the response. The selected metrics for comparison were BLEU(Papineni et al., 2002), ROUGE(Lin, 2004), and METEOR(Banerjee and Lavie, 2005) the outcome from these investigations is shown in Table 1.

A comparison of the responses and subsequent VAD scores produced by each of the aforementioned approaches to a set utterance, and how they differed from the agents target emotion score was also conducted. Table 2 shows the different approaches responses to selected conversation, and Table 3 indicates the associated VAD value for each of the responses as produced by the MRM.

5 Discussion

As shown by the evaluation, ETC did not perform as well as the initial response generator, though it did improve over that of the untuned seq2seq approach. Unfortunately, without human evaluation, it is difficult to determine whether the translated responses would be preferred by a human user, though from the output in Table 2 the responses that have been translated appear to be less conversational than those of the initial generator.

In the example provided in table 2 the ETC responses differ on average from the agent's target emotion by .1265 while the BlenderBot and untuned seq2seq model differed by .1553. Therefore, is possible that this approach would lead to improve the alignment with an created agent emotion therefore, providing a sort of control over the emotion of the response. However, it is difficult to determine whether the ETC responses would be preferred over that of the initial response generator.

Despite the somewhat poor results of automated metrics, ETC provides a potential means for creating and maintaining an independent but influenceable emotion for conversational AI. With further work the incorporation and use of VAD scores for agent emotion could provide a valuable way for enhancing consistency and control over the emotion of a generated response.

6 Limitations

As this work provides an initial investigation into the use of an independently maintained VAD score for the agent, emotion and translation for emotional alignment, there are a number of limitations.

The primary limitation is present in the quality of the models used in order to create the framework. As the focus was on investigating the theory as a whole, less time was devoted to ensuring the quality of the independent models. For instance, though different Regression Models such as a Linear Regressor and Decision Tree Regressor were investigated for the Multioutput Regression Model, the first construction of the seq2seq translation models were used despite potential ability to be improved. Moreover, the evaluation was conducted against a limited data set that may not truly reflect the framework's capability or lack there of.

Secondly, this work does not include human evaluation, which as noted above, is crucial for determining the actual quality of generated text. Additionally, as the focus is on handling emotions for conversation, the true quality of this approach cannot be determined by automated metrics.

Despite these limitations, this paper still provides a valuable theoretical contribution to emotional alignment in generated text by exploring a novel means for controlled and consistent emotional text generation.

6.1 Ethical Considerations

When working with emotions and text generation designed to appear human, it is important to be mindful of the ethical ramifications.

Some (Ghotbi, 2023) argue that present AI cannot accurately assess emotional data due to the complexity and nuances in its expression, and separately could reinforce stereotyping. Additionally, Stark and Hoey (2021) notes the importance of models of emotion and data used in regards to ethical appropriateness an emotional AI. As this work incorporates emotion recognition on the user input, it is crucial to acknowledge the risk of bias, sensitivity of the data, and a risk of harm that might arise from the technologies (Katirai, 2023).

Generative AI presents its own set of ethical challenges. An agent presenting a emotion could potentially be used to influence human decisions, and as well as concerns over manipulation (Klenk, 2024). Oniani et al. (2023) proposes a set of ethical principles, though intended directly for healthcare, could be used to guide ethical generative AI across fields.

When used ethically, these technologies have potential to support their users. This could be in the form of the ability to emotionally support users or detect if a user is at risk of harming themselves. Despite the aforementioned ethical concerns in this portion of the field, it is the hope of the authors that the ETC framework can inform future research, which will use such technology in an ethical manner to support its users.

References

Satanjeev Banerjee and Alon Lavie. 2005. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, pages 65–72, Ann Arbor, Michigan. Association for Computational Linguistics.

Sven Buechel and Udo Hahn. 2017a. EmoBank: Studying the impact of annotation perspective and representation format on dimensional emotion analysis. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 578–585, Valencia, Spain. Association for Computational Linguistics.

Sven Buechel and Udo Hahn. 2017b. Readers vs. writers vs. texts: Coping with different perspectives of text understanding in emotion annotation. In *Proceedings of the 11th Linguistic Annotation Workshop*, pages 1–12, Valencia, Spain. Association for Computational Linguistics.

Lars Buitinck, Gilles Louppe, Mathieu Blondel, Fabian Pedregosa, Andreas Mueller, Olivier Grisel, Vlad Niculae, Peter Prettenhofer, Alexandre Gramfort, Jaques Grobler, Robert Layton, Jake VanderPlas, Arnaud Joly, Brian Holt, and Gaël Varoquaux. 2013. API design for machine learning software: experiences from the scikit-learn project. In *ECML PKDD Workshop: Languages for Data Mining and Machine Learning*, pages 108–122.

Mingxiu Cai, Daling Wang, Shi Feng, and Yifei Zhang. 2024. EmpCRL: Controllable empathetic response generation via in-context commonsense reasoning and reinforcement learning. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 5734–5746, Torino, Italia. ELRA and ICCL.

Yulong Chen, Yang Liu, Liang Chen, and Yue Zhang. 2021. DialogSum: A real-life scenario dialogue summarization dataset. In *Findings of the Association* *for Computational Linguistics: ACL-IJCNLP 2021*, pages 5062–5074, Online. Association for Computational Linguistics.

P Ekman and W V Friesen. 1971. Constants across cultures in the face and emotion. *Journal of personality and social psychology*, 17(2):124–129.

Bo Li Xupeng Zha Haoqian Wang Ensi Chen, Huan Zhao and Song Wang. 2022. Affective feature knowledge interaction for empathetic conversation generation. *Connection Science*, 34(1):2559–2576.

Pascale Fung, Anik Dey, Farhad Bin Siddique, Ruixi Lin, Yang Yang, Yan Wan, and Ho Yin Ricky Chan. 2016. Zara the Supergirl: An empathetic personality recognition system. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Demonstrations*, pages 87–91, San Diego, California. Association for Computational Linguistics.

Pan Gao, Donghong Han, Rui Zhou, Xuejiao Zhang, and Zikun Wang. 2023. Cab: Empathetic dialogue generation with cognition, affection and behavior. In *Database Systems for Advanced Applications*, pages 597–606, Cham. Springer Nature Switzerland.

Nader Ghotbi. 2023. The ethics of emotional artificial intelligence: A mixed method analysis. *Asian Bioethics Review*, 15(4):417–430.

Umang Gupta, Ankush Chatterjee, Radhakrishnan Srikanth, and Puneet Agrawal. 2018. A sentiment-and-semantics-based approach for emotion detection in textual conversations. *Preprint*, arXiv:1707.06996.

Omama Hamad, Khaled Shaban, and Ali Hamdi. 2024. ASEM: Enhancing empathy in chatbot through attention-based sentiment and emotion modeling. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 1588–1601, Torino, Italia. ELRA and ICCL.

Amelia Katirai. 2023. Ethical considerations in emotion recognition technologies: a review of the literature. *AI and Ethics*.

Michael Klenk. 2024. Ethics of generative AI and manipulation: a design-oriented research agenda. *Ethics and Information Technology*, 26(1):9.

Bobo Li, Hao Fei, Fangfang Su, Fei Li, and Donghong Ji. 2024. Integrating discourse features and response assessment for advancing empathetic dialogue. *Information Processing Management*, 61(5):103803.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Zhaojiang Lin, Andrea Madotto, Jamin Shin, Peng Xu, and Pascale Fung. 2019. MoEL: Mixture of empathetic listeners. 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 121–132, Hong Kong, China. Association for Computational Linguistics.

Chia-Wei Liu, Ryan Lowe, Iulian Serban, Mike Noseworthy, Laurent Charlin, and Joelle Pineau. 2016. How NOT to evaluate your dialogue system: An empirical study of unsupervised evaluation metrics for dialogue response generation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2122–2132, Austin, Texas. Association for Computational Linguistics.

Yuhan Liu, Jun Gao, Jiachen Du, Lanjun Zhou, and Ruifeng Xu. 2022. Empathetic response generation with state management. *Preprint*, arXiv:2205.03676.

Navonil Majumder, Pengfei Hong, Shanshan Peng, Jiankun Lu, Deepanway Ghosal, Alexander Gelbukh, Rada Mihalcea, and Soujanya Poria. 2020. MIME: MIMicking emotions for empathetic response generation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8968–8979, Online. Association for Computational Linguistics.

Saif Mohammad. 2018. Obtaining reliable human ratings of valence, arousal, and dominance for 20,000 English words. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 174–184, Melbourne, Australia. Association for Computational Linguistics.

David Oniani, Jordan Hilsman, Yifan Peng, Ronald K Poropatich, Jeremy C Pamplin, Gary L Legault, and Yanshan Wang. 2023. Adopting and expanding ethical principles for generative artificial intelligence from military to healthcare. *npj Digital Medicine*, 6(1):225.

Zi Haur Pang, Yahui Fu, Divesh Lala, Keiko Ochi, Koji Inoue, and Tatsuya Kawahara. 2024. Acknowledgment of emotional states: Generating validating responses for empathetic dialogue. *Preprint*, arXiv:2402.12770.

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*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Sungjoon Park, Jiseon Kim, Seonghyeon Ye, Jaeyeol Jeon, Hee Young Park, and Alice Oh. 2021. Dimensional emotion detection from categorical emotion. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages

4367–4380, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Baolin Peng, Michel Galley, Pengcheng He, Chris Brockett, Lars Liden, Elnaz Nouri, Zhou Yu, Bill Dolan, and Jianfeng Gao. 2022. Godel: Large-scale pre-training for goal-directed dialog. arXiv.

Yushan Qian, Bo Wang, Shangzhao Ma, Wu Bin, Shuo Zhang, Dongming Zhao, Kun Huang, and Yuexian Hou. 2023. Think twice: A human-like two-stage conversational agent for emotional response generation. In *Proceedings of the 2023 International Conference on Autonomous Agents and Multiagent Systems*, AAMAS '23, page 727–736, Richland, SC. International Foundation for Autonomous Agents and Multiagent Systems.

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 Annual Meeting of the Association for Computational Linguistics*, pages 5370–5381, Florence, Italy. Association for Computational Linguistics.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Stephen Roller, Emily Dinan, Naman Goyal, Da Ju, Mary Williamson, Yinhan Liu, Jing Xu, Myle Ott, Eric Michael Smith, Y-Lan Boureau, and Jason Weston. 2021. Recipes for building an open-domain chatbot. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 300–325, Online. Association for Computational Linguistics.

James A Russell and Albert Mehrabian. 1977. Evidence for a three-factor theory of emotions. *Journal of Research in Personality*, 11(3):273–294.

Sahand Sabour, Chujie Zheng, and Minlie Huang. 2022. Cem: Commonsense-aware empathetic response generation. *Proceedings of the AAAI Conference on Artificial Intelligence*, 36(10):11229–11237.

Elvis Saravia, Hsien-Chi Toby Liu, Yen-Hao Huang, Junlin Wu, and Yi-Shin Chen. 2018. CARER: Contextualized affect representations for emotion recognition. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3687–3697, Brussels, Belgium. Association for Computational Linguistics.

Jamin Shin, Peng Xu, Andrea Madotto, and Pascale Fung. 2021. Generating empathetic responses by looking ahead the user's sentiment. *Preprint*, arXiv:1906.08487.

Luke Stark and Jesse Hoey. 2021. The ethics of emotion in artificial intelligence systems. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '21, page 782–793, New York, NY, USA. Association for Computing Machinery.

Ke Wang and Xiaojun Wan. 2018. Sentigan: Generating sentimental texts via mixture adversarial networks. In *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence, IJCAI-18*, pages 4446–4452. International Joint Conferences on Artificial Intelligence Organization.

Benjamin Wortman. 2024. *Models of Human Emotion and Artificial Emotional Intelligence*, pages 3–21. Springer International Publishing, Cham.

Zhou Yang, Zhaochun Ren, Wang Yufeng, Shizhong Peng, Haizhou Sun, Xiaofei Zhu, and Xiangwen Liao. 2024. Enhancing empathetic response generation by augmenting llms with small-scale empathetic models. *Preprint*, arXiv:2402.11801.

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 and Thirtieth Innovative Applications of Artificial Intelligence Conference and Eighth AAAI Symposium on Educational Advances in Artificial Intelligence*, AAAI'18/IAAI'18/EAAI'18. AAAI Press.

Xianda Zhou and William Yang Wang. 2018. MojiTalk: Generating emotional responses at scale. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1128–1137, Melbourne, Australia. Association for Computational Linguistics.

7 Appendix

7.1 Dataset Details

All the data sets used are of the English language, and this work falls within the bounds of their licenses. Links, licenses, and URLs are provided in Table 4.

7.2 Experimental Details

The artifacts from scikit learn fall under the BSD License (new BSD), the HuggingFace libraries use the Apache Software License (Apache 2.0 License), and SentenceTransformers also uses the Apache Software License (Apache License 2.0), and the particular model used was the "all-MiniLM-L6-v2". Where there is a paper relevant to the model it is

Data set	Reference	License	URL
EmoBank	Buechel and Hahn (2017a,b)	CC-BY-SA 4.0	https://github.com/JULIELab/EmoBank/tree/master/corpus
NRC VAD Lexicon	Mohammad (2018)	non-commercial research and educational purposes	https://saifmohammad.com/WebPages/nrc-vad.html
EmpatheticDialogues	(Rashkin et al., 2019)	Attribution-NonCommercial 4.0 International	https://huggingface.co/datasets/facebook/ empathetic_dialogues
Emotion Dataset	Saravia et al. (2018)	educational and research purposes	https://huggingface.co/datasets/dair-ai/emotion
dialogsum	(Chen et al., 2021)	CC BY-NC-SA 4.0	https://huggingface.co/datasets/knkarthick/ dialogsum

Table 4: Dataset details.

Component	URL	API
MultiOutput Regressor wrapper	https://scikit-learn.org/stable/modules/generated/sklearn.multioutput.MultiOutputRegressor.html	scikit-learn
Gradient Boosting Regressor	https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.GradientBoostingRegressor.html	scikit-learn
vectoriser	https://scikit-learn.org/stable/modules/generated/sklearn.feature_extraction.text.TfidfVectorizer.html	scikit-learn
Nearest Neighbors	https://scikit-learn.org/stable/modules/neighbors.html	scikit-learn
BlenderBot	https://huggingface.co/facebook/blenderbot-400M-distill	HuggingFace transformers
GODEL	https://huggingface.co/microsoft/GODEL-v1_1-base-seq2seq	HuggingFace transformers
Semantic Textual Similarity	https://www.sbert.net/docs/sentence_transformer/usage/semantic_textual_similarity.html and https://huggingface.co/sentence-transformers/all-MiniLM-L6-v2	SentenceTransformers
BLEU	https://huggingface.co/spaces/evaluate-metric/bleu	HuggingFace evaluate
ROUGE	https://huggingface.co/spaces/evaluate-metric/rouge	HuggingFace evaluate
METEOR	https://huggingface.co/spaces/evaluate-metric/meteor	HuggingFace evaluate

Table 5: Where components are reused from other sources both the API and URL are listed above.

cited in the main text of the paper. The models used their default hyperparameters for the purpose of this project, the BlenderBot model has 400 million parameters, all-MiniLM-L6-v2 has 22.7 million parameters, and the GODEL model has 117 million parameters. Further details are provided in Table 5.

Any other libraries used, as well as their versions, are outlined in the code repository[4] .

It is estimated that in total the project required around 36 hours for computation, including fine-tuning, data processing, and evaluation. the computation was run on an Apple M2 Max computer using CPU.

As the proposed translation method is not a model, but rather a framework, the evaluation was conducted by first collecting predicted responses to the data set from each model, these responses were run through the various evaluation metrics.

[4]https://github.com/codesubmissionanon112/ACLAnonSubmission