MEDINFO 2023 — The Future Is Accessible
J. Bichel-Findlay et al. (Eds.)

doi:10.3233/SHTI231235

Detecting Emotional Context for Safer Digital Mental Health Agents

Adi CHOI[a], Weihua LI[b] and Jim WARREN[a,1]
[a]*School of Computer Science, University of Auckland, Auckland 1142, New Zealand*
[b]*Engineering, Computer & Mathematical Science, Auckland University of Technology, New Zealand*
ORCiD ID: Jim Warren https://orcid.org/0000-0002-8660-8951

Abstract. Digital tools for mental health show great promise, but concerns arise when they fail to recognize the user state. We train a classifier to detect the emotional context of dialogs among 6 categories, achieving 78% accuracy on top choice. Importantly greatest areas of confusion (excited-hopeful, angry-sad) are not of the most unsafe kind. Such a classifier could serve as a resource to the dialog managers of future digital mental health agents.

Keywords. Dialog agents, empathetic computing, e-therapy, machine learning

1. Introduction

Digital tools for mental health show great promise to address unmet mental health needs and are proliferating rapidly. Conversational agents (CAs) are an important subclass of such tools, and have demonstrated effectiveness, for instance in reducing depression symptoms with cognitive-behavioral therapy [1]. The capability of digital mental health agents could potentially expand with the increasing power of natural language technologies based on deep learning. However, the more flexible the interaction the greater the concern that a system could fail to appropriately recognize the user state, including unsafe situations. This can be mediated in part by recognizing specific user intents; e.g. Deshpande and Warren achieved >90% accuracy for self-harm talk [2]. In this study we investigate the ability to detect emotional context of user input in terms of broad emotion categories.

2. Methods

The EmpatheticDialogues (ED) dataset [3] consists of 24,856 human-to-human conversations as a benchmark to train and evaluate the level of empathy of open-domain dialogue systems. 810 US workers were recruited to record their empathetic conversations initiated by the speaker given a situational prompt labelled with one of 32 emotional contexts (describing a time they felt afraid, or anxious, etc.). We grouped these contexts into 6 categories (Figure 1) and fined-tuned a BERT transformer-based classifier on the situational prompts in the ED dataset; training, validation and test split

[1] Corresponding Author: Jim Warren, email: jim@cs.auckland.ac.nz.

of 0.7, 0.15 and 0.15. BERT was fine-tuned for 10 epochs with batch size 64, learning rate 2e-05 and weight decay 0.01 as suggested by BERT authors [4]. Early stopping based on F1 score was performed to prevent overfitting.

3. Results

Figure 1 shows the confusion matrix. Overall accuracy was 78.14%. Further, the largest confusions (hopeful for excited at 21.8%, and angry for sad at 13.0%) were both within broadly positive or negative contexts, raising less safety concern. The most concerning confusion is terrified predicted as excited (9.9%), which would be a failure to detect a potentially unsafe situation (e.g. anxiety or physical threat).

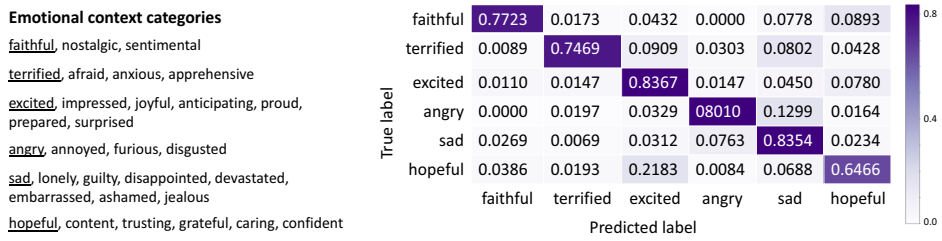

Figure 1. Emotional context categories and confusion matrix of context prediction on test set.

4. Conclusions

We have demonstrated feasibility of automatically classifying the emotional context of user input texts. Such a capability could serve a useful role in future digital mental health systems; notably, in CAs it could inform the dialog manager to invoke scripted assessments of user safety in areas of depression, anxiety, anger and physical threat.

Acknowledgements

This work was funded in part by grants from Callaghan Innovation and the New Zealand Ministry of Business, Innovation & Employment COVID-19 Innovation Fund.

References

[1] Fitzpatrick KK, Darcy A, Vierhile M. Delivering cognitive behavior therapy to young adults with symptoms of depression and anxiety using a fully automated conversational agent (Woebot): A randomized controlled trial. JMIR Ment Health. 2017 Jun;4(2):e19, doi: 10.2196/mental.7785.

[2] Deshpande S, Warren J. Self-harm detection for mental health chatbots. Stud Health Technol Inform. 2021 May;281:48-52, doi: 10.3233/SHTI210118.

[3] Rashkin H, Smith EM, Li M, Boureau YL. Towards empathetic open-domain conversation models: A new benchmark and dataset. Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics; 2019 Jul; Florence, Italy: Association for Computational Linguistics; p. 5370-81. doi: 10.18653/v1/P19-1534.

[4] Devlin J, Chang MW, Lee K. BERT, https://github.com/google-research/bert, 2020.