2022 10th International Conference on Affective Computing and Intelligent Interaction Workshops and Demos (ACIIW) | 978-1-6654-5490-2/22/$31.00 ©2022 IEEE | DOI: 10.1109/ACIIW57231.2022.10086016

A chat with Dr. Jekyll and Mr. Hyde - Intent in chatbot communication

1st Jonas Poehler
University of Siegen
57076 Siegen, Germany
0000-0002-9942-8298

2nd Nadine Flegel
Trier University of Applied Sciences
D-54293 Trier, Germany

3rd Tilo Mentler
Trier University of Applied Sciences
D-54293 Trier, Germany
0000-0002-8138-6536

4th Kristof Van Laerhoven
University of Siegen
57076 Siegen, Germany
0000-0001-5296-5347

***Abstract*—What would happen if a chatbot tries to manipulate your emotions? Can a modern language model output stable sentiment oriented conversations which can manipulate the user? We propose a framework to explore chatbots which have hidden intentions in their interaction with the user.**

***Index Terms*—chatbot, sentiment, emotion elicitation**

I. Introduction

What if the computer you are communicating with has hidden intentions? Large scale language models like GPT-3 or the open source alternative GPT-Neo have sparked the development of chatbot applications that mimmick human communication. What is lacking in all these applications is the intent of the communication. In this paper, we describe the development and testing of an chatbot application that has hidden intentions when it is communicating with it's human counterpart. We propose that this chatbot application can be designed to elicite emotions using text communication.

II. Related Work

A. Large Scale Language Models

In recent years, a large number of different language models have been developed with the goal of learning natural language. For this purpose, these Deep Learning models were trained on a large number of different texts. Over the years, a transformer architecture has become established. Transformers are a further development of Recursive Neural Networks. They consist of encoder and decoder modules that take the input sequence in the form of embeddings and convert it into an output sequence using feed forward layers and attention functions. [1] The first model to apply this architecture to speech on a larger scale was BERT [2]. The innovations in the development of BERT included the use of unlabeled data as training for the model and the ability to use the trained model for a variety of different tasks in language processing. These models were further developed by OpenAI, among others, increasing the size of the model and the amount of training data, resulting in the GPT model. [3] The most recent iteration, GPT-3, has the ability to generate text that is difficult to distinguish from human-written text. Since the development of OpenAI is not public, EleutherAI is an open source movement that also develops large scale language models and makes them freely available. The most current model has 20 trillion parameters and is also capable of solving a variety of language-specific tasks. The performance of the model is comparable to that of the GPT-3 model [4].

B. Chatbots and Sentiment

Sentiment analysis - the determination of feelings and moods in text communications - has been used in many different ways in conjunction with chatbots. There are several different ways to perform sentiment analysis. On the one hand, there are rule-based frameworks like VADER where a score value is stored in a lexicon for different values and then the total score of a text can be determined. Furthermore, there are Deep Learning based frameworks like the DistilBERT from HuggingFace [5]. These can be used to determine whether the sentiment of a text is positive or negative. When used in conjunction with chatbots, sentiment analysis has so far been used primarily for evaluating the text the user enters. For example, An et. al. [6] use sentiment analysis to provide psychological tests and assistance to the user matching the sentiment of their responses in their chatbot which is intended to provide psychological support. In other domains, such as Nivethan et. al., sentiment analysis is used to determine if the chatbot's feedback is appropriate for the user [7]. But Sentiment Analysis can also be used for better fitting answers. For example, it is used by Lee et. al. to make the chatbot give more appropriate answers to the tone of the conversation [8]. Occasionally it is also applied to the answers of the bot. However, the only application so far seems to be to keep the answers of the bot positive as shown by Murali et. al. [9].

This project is funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) – 425868829 and is part of Priority Program SPP2199 Scalable Interaction Paradigms for Pervasive Computing Environments.

978-1-6654-5490-2/22/31.00©2022$IEEE$

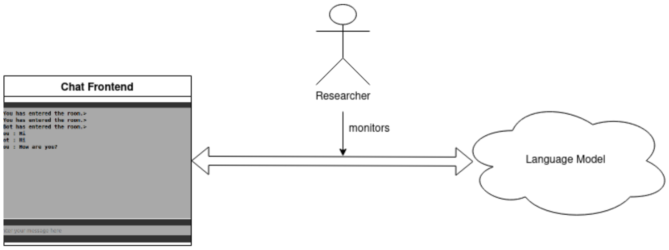

Fig. 1. Overview of the components of the chatbot application

III. Design

Several subcomponents are needed when interacting with the program. An overview of them is shown in figure 1. First, when the user visits the website, he is shown a chat where he can interact with the system. In the background, each user is randomly assigned to one of two instances of the system. One of them has benign intentions and the other has malicious intentions. The intentions of the system are mapped by the sentiment of its responses. The benign instance tries to give only answers with a positive sentiment while the malicious instance gives answers with a negative sentiment. The idea behind this is that the sentiment in the chatbot's answers is also reflected in the interaction with the user and that the user is also enticed to give answers with similar sentiment. To do this, the user's responses are now fed into the language model. To preserve the context of the conversation and to make the output more similar to a human conversation, not only the last answer is used as prompt for the model, but the complete interaction, i.e. also the last answers of the model of the last 20 messages are used. For performance reasons the GPT-Neo-1.3B from EleutherAI is used as the language model. It is comparable with the smaller instances of GTP-3 like Ada. Now 10 different possible answers of the model are sampled. For each possible answer, the sentiment of the text is calculated using the VADER model [10]. The response with the highest or lowest sentiment, depending on which instance is presented to the user, is then chosen as the response to the user. The algorithm is also shown as pseudocode in the algorithm 1.

The interaction with the chatbot is not unfiltered. All interactions are monitored by a researcher who has to approve the answers the bot is about to give to the user.

Algorithm 1 The chatbot algorithm

```
Capture user input
for i==0; i<10; i++ do
    Sample Output from Language Model
    Calculate sentiment using VADER
    Return answer with highest or lowest sentiment
end for=0
```

IV. Evaluation

For a preliminary evaluation, 60 chats were examined. This included 30 chats with the benign bot and 30 chats with the malicious bot. For each response, the sentiment was calculated using VADER. Figure 2 shows how the sentiment develops over time. It can be seen that there is a significant difference in the development of the sentiment between the two different bots. Both start in a normal to positive range but diverge significantly over time. The malicious bot usually shows a significant decrease in the sentiment of its responses, while the responses of the benign bot tend to get a more positive sentiment. It can also be observed that partly due to the user input there can be a significant increase or decrease in the sentiment when the bot reacts to the content of the user input.

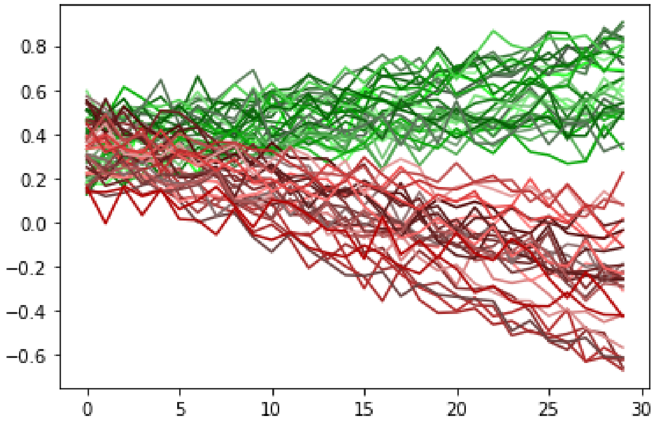

Fig. 2. Resulting sentiment of chatbot answers in 30 chat sessions. The y axis is depicting the answer sentiment on a 1 to -1 scale whereas the x axis shows the number of answers. A cutoff was introduced after 30 answers.

V. Limitations

The system is based on the assumption that the user will pick up and reflect the sentiment of the chatbot's responses in his answers. This requires a user who engages with this system and interacts with it in an open and curious manner. A user who tries to exploit and manipulate the system cannot be manipulated by the system with such a simple setup. Since the system takes the content of the user's messages and uses it for its own responses, it is also possible to induce the instance with benign intentions to make statements with negative sentiment. The next major limitation in the usability of this system is its performance. In the current configuration, it takes 10-15 seconds to compute a response. Thus, when used simultaneously by multiple users, the response time of the chatbot would increase a lot. This has both advantages and disadvantages. To some extent, a higher response time represents a more human communication than if the system's response appears in a fraction of a second, but if the response time increases to too high a level, the interaction becomes too cumbersome and the user will no longer want to interact with the system.

VI. Discussion

So far, the possibilities of the presented system are extremely limited, but this approach can still be extended in several directions. The most important extension that would have to be added next would be an evaluation of the success.

Accordingly, the user reacts as expected and the sentiment approaches the desired one. For this purpose, the sentiment of the user input could be analyzed as well and this could then be used to select the appropriate response of the chatbot. Likewise, the selection of the answers should be designed more intelligently, here a neural network is offered which is trained with the sentiment of the user input and the answer of the chatbot and thus should learn over the course of the conversation with which answers which change in the sentiment can be achieved with this user. Likewise, a larger text generation model would be an improvement as it can generate even more convincing texts and answers and thus better maintain the appearance of a real conversation. Furthermore, it would be a good idea to use a better way to determine sentiment that has a higher precision than the rule-based VADER model. However, both extensions are again at the expense of performance, so more computing power will be needed.

A. *Applications*

The most important question in the discussion of the presented system, however, is why such a chatbot offers added value. The history of social chatbots is longer and representatives like ELIZA have shown that they can fulfill the human need for communication. [11] However, in order for them to be more than just a mere occupation and pastime they must be able to evoke a change in the user. The presented project deals with the question if modern chatbots can cause such a change. Applications for a benign chatbot would be for example in possibilities of mental support to cope with stressful situations.

B. *Ethical Impact*

We are aware that a deliberate manipulation of the user brings many ethical problems with it. However, in the present configuration of the chatbot, where the interaction is monitored by a researcher and the answers must be Wizard of Oz-like unlocked, there is little danger. Also, due to the lack of a learning process, effects like those of the bot Tay cannot occur. [12] Based on this, we would classify the chatbot as borderline but ethical according to de Lima Salge et. al [13]. However, with any enhancements, particularly in the introduction of learning processes, this evaluation would need further consideration.

References

[1] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Kaiser, and I. Polosukhin, "Attention is All you Need," in *Advances in Neural Information Processing Systems*, vol. 30. Curran Associates, Inc., 2017.

[2] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," arXiv, Tech. Rep. arXiv:1810.04805, May 2019, arXiv:1810.04805 [cs] version: 2 type: article. [Online]. Available: http://arxiv.org/abs/1810.04805

[3] T. B. Brown, B. Mann, N. Ryder, M. Subbiah, J. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, S. Agarwal, A. Herbert-Voss, G. Krueger, T. Henighan, R. Child, A. Ramesh, D. M. Ziegler, J. Wu, C. Winter, C. Hesse, M. Chen, E. Sigler, M. Litwin, S. Gray, B. Chess, J. Clark, C. Berner, S. McCandlish, A. Radford, I. Sutskever, and D. Amodei, "Language Models are Few-Shot Learners," arXiv, Tech. Rep. arXiv:2005.14165, Jul. 2020, arXiv:2005.14165 [cs] type: article. [Online]. Available: http://arxiv.org/abs/2005.14165

[4] S. Black, S. Biderman, E. Hallahan, Q. Anthony, L. Gao, L. Golding, H. He, C. Leahy, K. McDonell, J. Phang, M. Pieler, U. S. Prashanth, S. Purohit, L. Reynolds, J. Tow, B. Wang, and S. Weinbach, "GPT-NeoX-20B: An Open-Source Autoregressive Language Model," arXiv, Tech. Rep. arXiv:2204.06745, Apr. 2022, arXiv:2204.06745 [cs] type: article. [Online]. Available: http://arxiv.org/abs/2204.06745

[5] V. Sanh, L. Debut, J. Chaumond, and T. Wolf, "DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter," arXiv, Tech. Rep. arXiv:1910.01108, Feb. 2020, arXiv:1910.01108 [cs] type: article. [Online]. Available: http://arxiv.org/abs/1910.01108

[6] S. H. An and O. R. Jeong, "A Study on the Psychological Counseling AI Chatbot System based on Sentiment Analysis," *Journal of Information Technology Services*, vol. 20, no. 3, pp. 75–86, Jun. 2021. [Online]. Available: https://doi.org/10.9716/KITS.2021.20.3.075

[7] Nivethan and S. Sankar, "Sentiment Analysis and Deep Learning Based Chatbot for User Feedback," in *Intelligent Communication Technologies and Virtual Mobile Networks*, ser. Lecture Notes on Data Engineering and Communications Technologies, S. Balaji, Rocha, and Y.-N. Chung, Eds. Cham: Springer International Publishing, 2020, pp. 231–237.

[8] C.-W. Lee, Y.-S. Wang, T.-Y. Hsu, K.-Y. Chen, H.-Y. Lee, and L.-S. Lee, "Scalable Sentiment for Sequence-to-Sequence Chatbot Response with Performance Analysis," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Apr. 2018, pp. 6164–6168, iSSN: 2379-190X.

[9] S. R. Murali, S. Rangreji, S. Vinay, and G. Srinivasa, "Automated NER, Sentiment Analysis and Toxic Comment Classification for a Goal-Oriented Chatbot," in *2020 Fourth International Conference On Intelligent Computing in Data Sciences (ICDS)*, Oct. 2020, pp. 1–7.

[10] C. Hutto and E. Gilbert, "VADER: A Parsimonious Rule-Based Model for Sentiment Analysis of Social Media Text," *Proceedings of the International AAAI Conference on Web and Social Media*, vol. 8, no. 1, pp. 216–225, May 2014. [Online]. Available: https://ojs.aaai.org/index.php/ICWSM/article/view/14550

[11] H.-y. Shum, X.-d. He, and D. Li, "From Eliza to XiaoIce: challenges and opportunities with social chatbots," *Frontiers of Information Technology & Electronic Engineering*, vol. 19, no. 1, pp. 10–26, Jan. 2018. [Online]. Available: https://doi.org/10.1631/FITEE.1700826

[12] G. Neff, "Talking to bots: Symbiotic agency and the case of tay," *International Journal of Communication*, 2016.

[13] C. A. de Lima Salge and N. Berente, "Is that social bot behaving unethically?" *Communications of the ACM*, vol. 60, no. 9, pp. 29–31, Aug. 2017. [Online]. Available: https://doi.org/10.1145/3126492