Leia: An Emotional Support Chatbot System

Ling Gan, Lingzi Liu, Jiehui Luo

ABSTRACT

Have you ever stared at your contact list during sleepless nights and found no one available to talk to? Thanks to technology, we can talk to virtual machines who are available 24/7. In this paper, we introduce an emotional support chatbot system, Leia, aiming to fill this blank of conversational chatbot systems. By applying the bag-of-words model with the TF-IDF algorithm and sentiment analysis on the DailyDialog corpus, we train Leia to understand the content of the user inputs and generate positive responses that ease the stress of users. In addition, we describe our exploration process of the seq2seq model with LSTM architecture in terms of chatbot text generation. At last, we present a human evaluation experiment to measure Leia's communication ability From the observed users' feedback, Leia delivers a performance that followed our initial motivations.

INTRODUCTION

Chatbot, designed to parallel humans in message exchanges, has been around for decades. The early chatbots include Eliza (Weizenbaum, 1966), which uses pattern matching algorithms to generate responses and thus ensures fluency of conversation. Entering the twenty-first century, research on dialogue systems has embraced a remarkable growth that products like Apple Siri, Amazon Alexa, and Microsoft Xiaoice have been parts of people's daily lives. However, these dialogue systems mostly focus on information retrieval and aim for finding the most useful information to the user, where the importance of users' emotion feedback may be put into a secondary position.

To address this problem, we create a new chatbot system and referred to it as Leia. The chatbot is designed to be text-based to avoid possible embarrassment talking to a chatbot in public. The purpose of our chatbot system is to provide encouragement and comfort by replying with sentences in positive tones rather than providing knowledge or information to specific questions.

In this paper, we first describe the two datasets we used and demonstrated the approaches we tried. Then we evaluate our models by conducting user studies with quantitative and qualitative questions. As a result, we observed that users are more willing to have a longer conversation with Leia, and they generally experienced an increase in mood after having such a conversation.

RELATED WORK

Early work on dialogue systems like Eliza was based mainly on states and rules hand-crafted by human experts, while modern dialogue systems typically follow a hybrid architecture, combining hand-crafted states and rules with statistical machine learning algorithms (Serban et al., 2017). For modern dialogue systems, in particular, we examined the recent work on chatbot systems and found out they could be roughly divided into two categories.

The first genre of chatbots is designed to provide better support in terms of customer services, like AliMe Chat proposed by Alibaba Group (Qiu et al., 2017) and a new conversational system created by IBM Research (Xu et al., 2017). Alime Chat is an open-domain chatbot engine that answers millions of customer questions in the E-commerce industry per day, and it uses a hybrid approach that integrates both informational retrieval and generation models. The IBM conversational system is designed to generate responses for users' requests on social media automatically. It is integrated with state-of-the-art deep learning techniques and is trained by nearly 1M Twitter conversations between users and agents from over 60 brands. In our starting stage, these high-level research offered us a general overview of the popular techniques used to design a chatbot system. However, as their nature of being supporting customer services, their work is not precisely what we are looking for, which focuses on human emotions instead.

The second genre of chatbots is designed for counseling services, like the conversational agent developed by Danielle Elmasri and Anthony Maeder (2016), and the chat assistant presented by Dongkeon Lee, Kyo-Joong Oh, and Ho-Jin Choi (2017). The conversational agent investigated the suitability of chatbots for mental health intervention, specifically alcohol drinking habits assessment, while the chat assistant understands the content of conversation based on recent natural language processing methods with emotion recognition, and senses emotional flow through the continuous observation of conversation. Although both pieces of research aligned with our purpose of making the chatbot emotional, their applications are too specifically defined within the context of identifying emotional hardships on the level of medical crises, which disaccord with our goal of providing support to general users. Thus, we still need to conduct our own research.

For evaluation, we first looked into the research work previously done by Chia-Wei Liu, Ryan Lowe, Iulian V. Serban, Michael Noseworthy, Laurent Charlin, Joelle Pineau (2016). According to Liu's group, automatic evaluation metrics such as BLEU and METEOR correlates very weakly with human judgments in the non-technical Twitter domain, and not at all in the technical Ubuntu domain. Thus, evaluation metrics are not optimal for the evaluation of the unsupervised dialogue response generation model. We then investigated ChatEval, a tool for chatbot evaluation created by the NLP group of the University of Pennsylvania (2019). It is a unified framework for the human evaluation of chatbots that augments existing tools and provides a web-based hub for researchers to share and compare their dialog systems. Although this evaluation method seems solid and scientific, it asks a specific format of the model as an input. Thus, for the initial evaluation stage, we decided to keep using human judgments in the form of user studies for analyzing and comparing.

METHODOLOGY

I. Dataset

The first dataset we used is Cornell Movie-Dialogs Corpus collected by Cristian Danescu-Niculescu-Mizil and Lillian Lee (2011). The corpus contains a metadata-rich collection of fictional conversations extracted from raw movie scripts, which involves 9,035 characters from 617 movies and includes 220,579 conversational exchanges between 10,292 pairs of movie characters. For each conversation in the dataset, we extracted every sentence except the last as questions, and every sentence except the starting one as answers. Within the 304,713 utterances in total, we obtained 221,616 pairs of questions and answers.

The second dataset we used is DailyDialog Corpus (Li et al., 2017), which is a human-written and less noisy dataset. The dialogues in the dataset include our daily communications and cover various topics such as school life, relationships, work, and health. It was also manually labeled the developed dataset with communication intention and emotion information. The DailyDialog dataset contains 11318 transcribed dialogues. The dataset is manually labeled with different emotion numbers: 0 represents no emotion, 1 presents anger, 2 represents disgust, 3 represents fear, 4 represents happiness, 5 represents sadness, and 6 represents surprise.

II. Template-Based Model (Eliza)

We chose to implement a template-based chatbot system for our baseline model. We explored one of the earliest natural language processing computer programs called ELIZA, which is a chatbot system operating on pattern matching. The algorithm consists of a list of keywords and their associative transformation rules called SCRIPT. The original ELIZA takes the SCRIPT as data, which means that ELIZA is not restricted to any specific patterns. However, for our study, we implemented a specific set of patterns in our model and conducted matching with regular expressions. Our version of ELIZA could lead conversations by asking questions, in this case, specifically by asking questions about users' family to prompt further discussions.

III. TF-IDF Model

For our second approach, we implemented the term frequency-inverse document frequency model (TF-IDF). The model first accounts for the term frequency, which represents the count of the number of times a word appears in a document, the essential ideal of the bag-of-words algorithm. Then the model added inverse document frequency, which is a measure of how common a word is across documents in a given corpus. We first vectorized each sentence in our datasets and the user input. We then calculated the cosine similarity score between our inputs and the sentences. Our algorithm would go through all the scores and output the sentence with the highest similarity.

Initially, we trained our model with Cornell Movie Corpus Datasets. However, due to the theatricality and specificness of this dataset, the responses generated by our chatbot system contained many movie-specific terms like character names. They reflected a tone that was more dramatic and not necessarily appropriate for daily conversations. Therefore, we decided to adopt a new dataset, DailyDialog Corpus, which would be a better reflection of everyday conversations.

By using DailyDialog datasets, the model could successfully provide us with feedback whenever we typed in a sentence, but the outputs were generally not optimal, nor were they conversational. To improve our TF-IDF model, we first conducted a lemmatization of our data. Through lemmatizing the words, the inflected terms could be grouped and analyzed as a single item. Such lemmatization benefited our calculation of similarity scores in the context of two sentences with different forms of the same word.

Moreover, to achieve a sense of conversation, we split our data into questions and answers. Questions and answers lists were constructed by separating the dataset based on conversational exchanges: Each sentence was considered a *question* and its response, which was the sentence next inline, was considered an *answer*. After the model found the most similar question to the user's input, we outputted the corresponding answer in the corpus to the question. In this way, the output became more context-related and conversational.

Then we conducted a sentiment analysis using the annotation from DailyDialog datasets. We filtered out the sentences annotated with *happiness*, which was a score of 4, and its next sentences to construct our lists of questions and answers. By doing so, our dataset was reduced to 11829 lines of sentences per list. Then we conducted a round of word-level analysis using the AFINN word list (Rowe et al., 2011). AFINN word list contains 2477 words and phrases, where each word has a score ranging from negative five to positive five. Given the wordlist, we calculated a score for each sentence and only kept the sentence with positive overall scores. This step further reduced our dataset into 11315 sentences per list.

IV. LSTM Model

We have also explored the sequence to sequence and the Long Short-Term Memory model (Hochreiter and Schmidhuber, 1997) in this project. So far, for the TF-IDF approach, the responses we generated were all derived from the dataset. Although we did some cleaning and processing, we were still building on the existing sentence from the database.

However, for the LSTM approach, we tried to have the machine generates response word by word from scratch. The Seq2Seq model works as every sentence we speak has a sequential structure, and thus we have to understand each word based on our understanding of previous words. Such a model takes every word embedding as an input, encodes it with an LSTM cell, extracts features to feed the next cell, and thus generates a response using a greedy beam search procedure (Graves, 2013).

For implementation, we utilized the Seq2Seq encoder-decoder framework in TensorFlow and data from DailyDialog datasets. For outputs, our model was trained to predict the next word in the response for a given length. We did get a cluster of word predictions as a result; unfortunately, the words were incoherent and hardly maintained the structure of a sentence. Given the time limit, we decided to keep using the improved TF-IDF model for the evaluation.

EVALUATION

We conducted a user study to evaluate and compare the performance of our chatbot systems. Participants were first introduced to our project given the motivation and some basic instructions to use the chatbot systems. Then the evaluation was divided into three sessions. In the first session, we asked the participants to chat with Eliza (See Appendix I for demo screenshot) then complete the survey we prepared. In the second session, we asked the participants to chat with the initial TF-IDF model named Robo(See Appendix II for demo screenshot) , and then fill out the second part of the survey. In the third session, we asked the participants to chat with Leia(See Appendix III for demo screenshot), an improved version of our TF-IDF model, and then finish the last session of our survey.

In each session, we have prepared 7 questions as follows:

1. How many rounds have you chatted with the chatbot? (we count 1 input from the user and 1 output from chatbot as a single round)
2. (a) What was your mood before you started conversing with the chatbot? (Rate from 1-5, 1 represents negative/sad and 5 represents the positive/happy)
   (b) What was your mood after you finished conversing with the chatbot? (Rate from 1-5, 1 represents negative/sad and 5 represents the positive/happy)
3. What score will you give to the replying speed of the chatbot? (Rate from 1-5, 1 represents the slowest and 5 represents the fastest)
4. Do you think this chatbot gives a reasonable response? (Rate from 1-5, 1 represents disagree and 5 represents agree)
5. How do you like this chatbot overall? (Rate from 1-5, 1 represents just so-so and 5 represents fantastic)
6. What is the most impressive thing you find about this chatbot? (Answer in short answer)
7. How can we improve on this chatbot? (Answer in short answer)

We had 15 college students range in age from 20 to 25 participate in the user study. Each study lasted for 30 to 45 minutes. From the results we gathered for the first question, on average, our users chatted around 9 rounds with Eliza, 7 rounds with Robo, and 14 rounds with Leia. The results of this question imply that our users are more willing to chat with Leia in general because Leia has the maximum rounds on average. Robo gets the minimum rounds on average based on users' responses.

Then we did an analysis on question 2(a) and question 2(b). We used the mood score after chatting with chatbot minus the mood score before chatting with a chatbot to get the difference in mood changing. In this case, the positive score implies that the user is feeling better after the conversation, and the negative score implies that the user is feeling worse. The score is ranging from negative four to positive four. From the chart below, we can see that 4 out of 15 participants felt more positive after talking to Eliza, while 5 participants felt worse. For Robo, we had 5 participants who got a better mood, and 6 got worse. For Leia, we did not have any negative scores. All 15 participants felt happier after chatting (See Appendix IV for more details of the graphs).

$

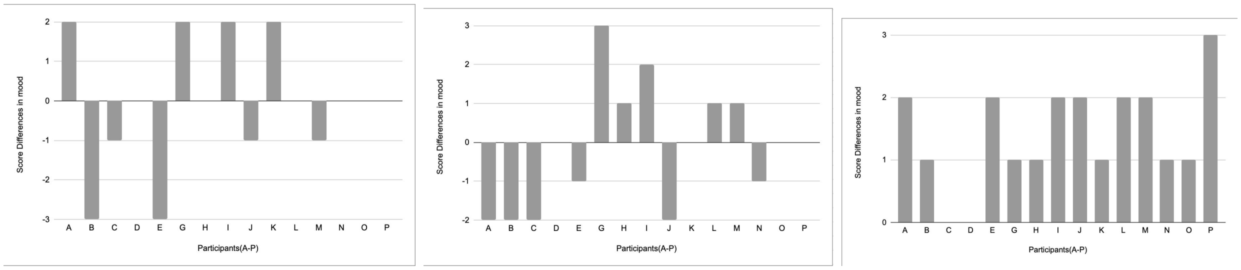


Figure 1: Results of the users' mood difference before/after chatting with the chatbot

For the replying speed, we can see that Eliza had the fastest replying speed and Robo performed badly on this part. Among the 15 participants, 13 users gave a score lower than 3 on the replying speed of Robo, while Leia got 10 rates which were higher than 3 on replying speed (See Appendix V). Robo has the slowest due to the nature of the TF-IDF algorithm, where cosine-similarity scores of every document have to be calculated individually with each input. When the dataset is no-filtered, the size of the dataset is enormous and thus causes a delay in the generation of responses.

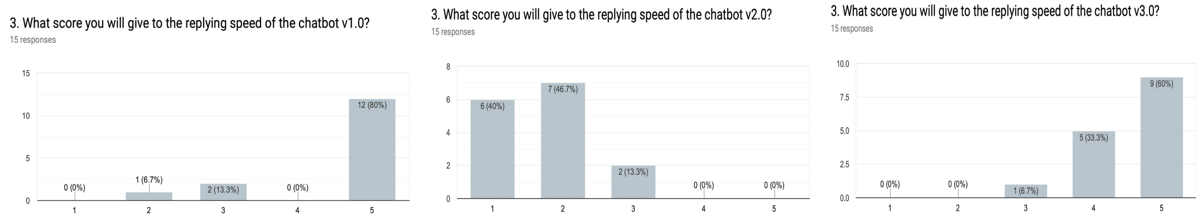


Figure 2: Ratings of the replying speed of the three chatbot models

On the evaluation of the response rationality, the results showed that Robo and Leia gave more reasonable responses than Eliza (See Appendix VI).

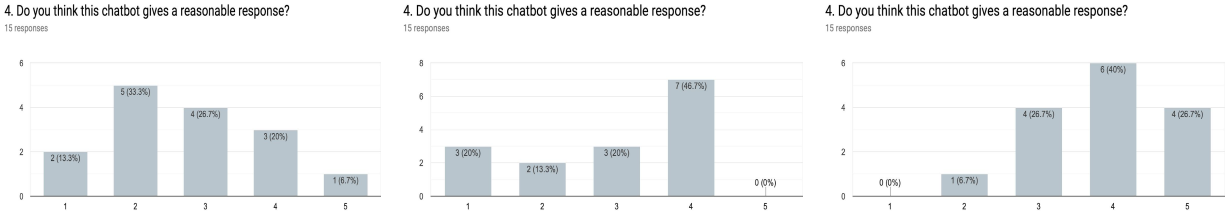


Figure 3: Ratings of the response rationality of the three chatbot models

Rating our three chatbot systems overall, Leia got the highest score, while Eliza got the second-highest, and Robo got the last (See Appendix VII).

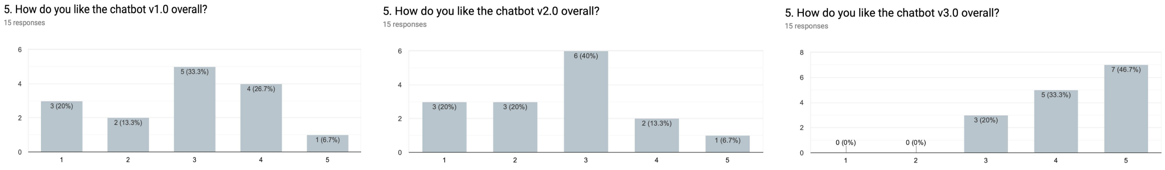


Figure 4: Overall ratings of the three chatbot models

According to the responses we got on the last two qualitative questions, all the three chatbot systems had their own highlights and shadows. For example, users pointed out Eliza always asked for more details in users' inputs but did not offer solutions or more feedback on the dialogue. Some participants pointed out the major problem of Robo is the replying speed and thus emphasized the importance of replying speed on the chatbot system. Overall, the improved version chatbot Leia overperformed the other two models by giving relatively reasonable responses in a relatively fast replying pace.

DISCUSSION

For this project, we implemented three models: Template-based, TF-IDF, and LSTM. For LSTM, our model did not do well in producing reasonable responses. As our model was constructing sentences on a word level, the words generated by each node could not form a consistent sentence. We inspected the model with care but yet had not been able to identify the failure. Therefore, we left it out for user studies.

In the course of development, we faced many challenges. First, due to our unique goal of providing emotional companionship, we are not able to find existing datasets of daily conversation that is built for this purpose. The available options are either just general everyday conversation, most of which are not annotated based on emotions, or topic-specific corpus built for analyzing specific scenarios.

Therefore, we had to do preliminary data cleaning using sentiment analysis to construct our own version of the dataset that reflected only positive emotions. This unique take on the dataset set our project different from others. However, the sentiment analysis we conducted has space for future improvement.

One limitation of our project is the assumption we made about the transitivity of positive emotions. We have taken for granted that receiving positive responses with necessarily increase the other party's mood. Luckily, our current user studies' results are in line with such assumptions. However, the sample size of our user study is still relatively small. Thus, our results are not necessarily sufficient to make any assertive inference in this case.

Another limitation regarding our sentiment analysis is the way we utilized the AFINN word list. We classified the sentences to be positive or negative based on the emotion score of each word. Such word-level generalization ignored the cases that negative-scored sentences could be conveying a sense of empathy to the users, which is another effective way of providing emotional support. For example, based on the AFINN word list, the sentence "I am so sorry to hear that" is negative-scored but it is conveying a supportive intention. Our approach filtered such sentences. Thus the response we generated did not contain all the possible emotional supportive answers.

Based on the feedback we gathered and the limitation we saw, there is still a big margin for improvements. For future steps, we plan 1) Optimize the selection of the response from datasets. Currently, we have tried to output the N best answers and found out that among approximately ten answers, there must exist an answer that is perfectly humanoid. However, we are still working on creating criteria that could extract the best answer out without human judgments so that the perfect answers could be outputted automatically for each input. 2) Apply more advanced generation models like RNN to have the chatbot generate its own sentence, rather than picking sentences as a response from the existing data. 3) Develop a better measurement of "an emotional supportive response." 4) Design a front-end interface to intrigue the users, and thus transform the conversation process more similar to a real conversation between friends. 5) Personalize the chatbot. In addition to naming the chatbot, we could also assign age, character, background history, and stories to the chatbot to make it more human-like. Moreover, since the users' background vary in age, gender, and personality, it would be optimal if the chatbot could switch tones and provide personalized responses when facing different users. 6) Collect feedback from more users to have a more representative picture of the general users.

Through developing a chatbot system with a goal of providing emotional support in mind, we explored the current landscape of this field and encountered many innovative approaches and humanistic motivations. We understood more profoundly that the domain of natural language processing extends beyond the scope of our classroom and even the boundaries of computer science as a discipline. In our process of development, our choices of datasets played one of the most crucial roles. It influenced our outcome in an even more profound way than the algorithms we chose to implement for our models. We were also challenged by the scale and inclusiveness of our datasets: Instead of taking the existing datasets for granted, we filtered out much of the irrelevant data to make our dataset more precise to serve our motivation. Most importantly, from this project, we are able to apply the knowledge we learned from class to solve some real-life problems and had chances to explore many more advanced state-of-the-art models alone the way.

REFERENCES

Danescu-Niculescu-Mizil, C., Lee, L. (2011). Chameleons in imagined conversations: A new approach to understanding the coordination of linguistic style in dialogs. *Workshop on Cognitive Modeling and Computational Linguistics, ACL 2011*.

Elmasri, D., Maeder, A. (2016). A Conversational Agent for an Online Mental Health Intervention. *Proceedings of International Conference on Brain and Health Informatics*, pp. 243-251.

Graves. A. (2013). Generating sequences with recurrent neural networks. *arXiv preprint arXiv:1308.0850.*

Hochreiter, S., Schmidhuber, J. (1997). Long short-term memory. *Neural Computation*, 9(8):1735–1780.

Lee D., Oh K., Choi H. (2017). The chatbot feels you - a counseling service using emotional response generation. *2017 IEEE International Conference on Big Data and Smart Computing*. DOI: 10.1109/BIGCOMP.2017.7881752.

Li, Y., Su, H., Shen, X., Li, W., Cao, Z., Niu, S. (2017). DailyDialog: A Manually Labelled Multi-turn Dialogue Dataset. *Proceedings of the Eighth International Joint Conference on Natural Language Processing*, Volume 1, 986-995.

Liu, C. W., Lowe, R., Serban V.I., Noseworthy, M., Charlin, L., Pineau, J. (2016).How not to evaluate your dialogue system: An empirical study of unsupervised evaluation metrics for dialogue response generation. *EMNLP*.

Qiu, M., Li, F., Wang, S., Gao, X., Chen, Y., Zhao, W., Chen, H., Huang, J., Chu, W. (2017). AliMe Chat: A Sequence to Sequence and Rerank based Chatbot Engine. *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics*, Volume 2, 498-503. 10.18653/v1/P17-2079.

Rowe, M., Stankovic, M., Dadzie, A., Hardey, M. (2011). A new evaluation of a word list for sentiment analysis in microblogs. *Proceedings of the ESWC2011 Workshop on 'Making Sense of Microposts': Big things come in small packages*, Volume 718, 93-98, http://ceur-ws.org/Vol-718/paper_16.pdf

Sedoc, J., Ippolito, D., Kirubarajan, A., Thirani, J., Ungar, L., Callison-Burch, C. (2019). ChatEval: A Tool for Chatbot Evaluation. *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics.*

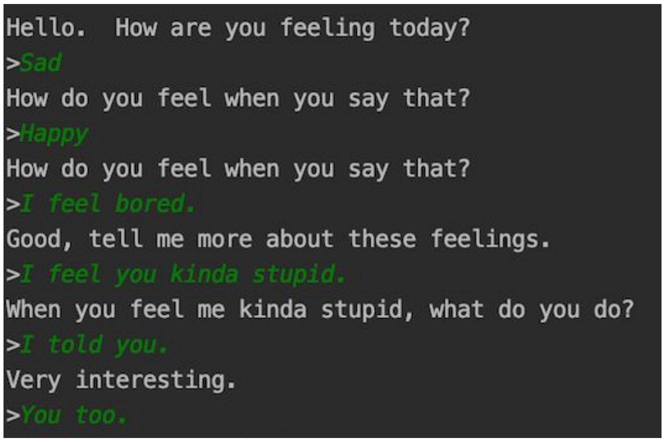
Serban, V. I., Sankar, C., Gemain, M., Zhang, S., et al. (2017). A Deep Reinforcement Learning

Chatbot. arXiv preprint. arXiv:1709.02349.

Weizenbaum, J. (1966). Eliza - A computer program for the study of Natural Language Communication between man and machine, *ACM* 9(1).

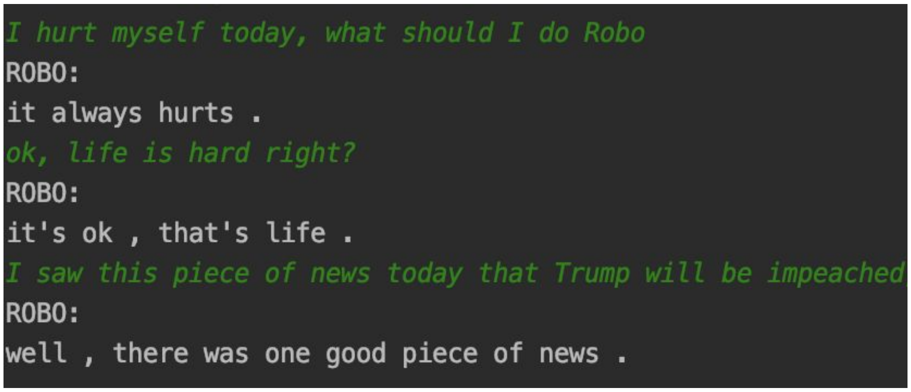
Xu, A., Liu, Z., Guo, Y., Sinha, V., Akkiraju, R. (2017). A new chatbot for customer service on social media. *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*, 3506-3510.

Appendix I: Demo Screenshot for Template-Based Model (Eliza)



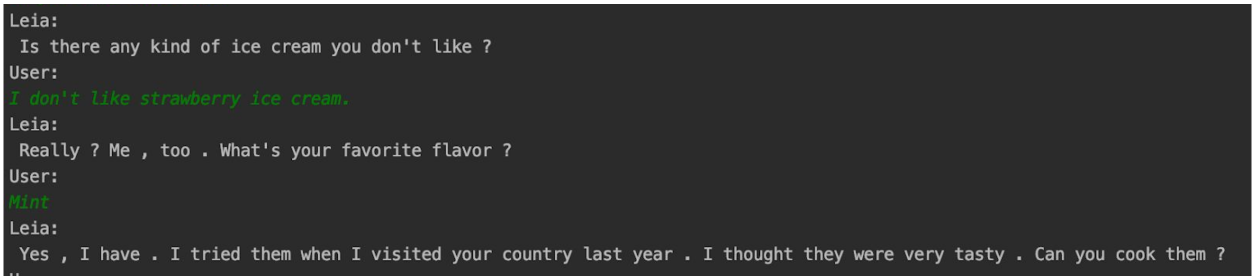
```
Hello.  How are you feeling today?
>Sad
How do you feel when you say that?
>Happy
How do you feel when you say that?
>I feel bored.
Good, tell me more about these feelings.
>I feel you kinda stupid.
When you feel me kinda stupid, what do you do?
>I told you.
Very interesting.
>You too.
```

Appendix II: Demo Screenshot for Initial TF-IDF Model (Robo)



```
I hurt myself today, what should I do Robo
ROBO:
it always hurts .
ok, life is hard right?
ROBO:
it's ok , that's life .
I saw this piece of news today that Trump will be impeached
ROBO:
well , there was one good piece of news .
```

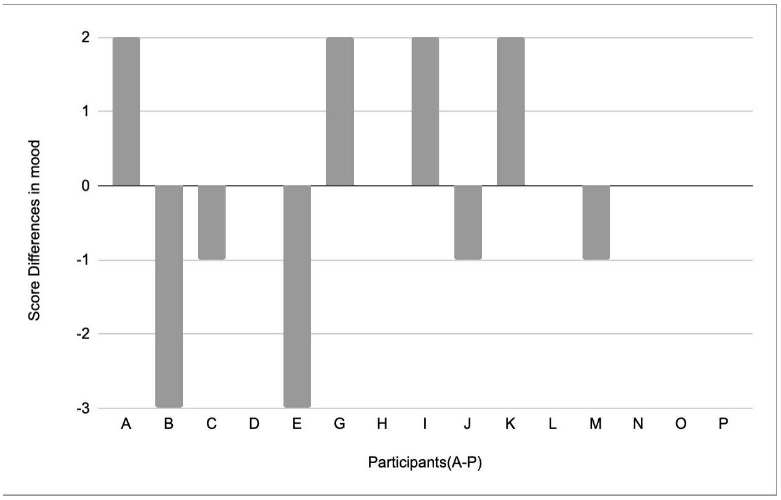
Appendix III: Demo Screenshot for Improved TF-IDF Model (Leia)



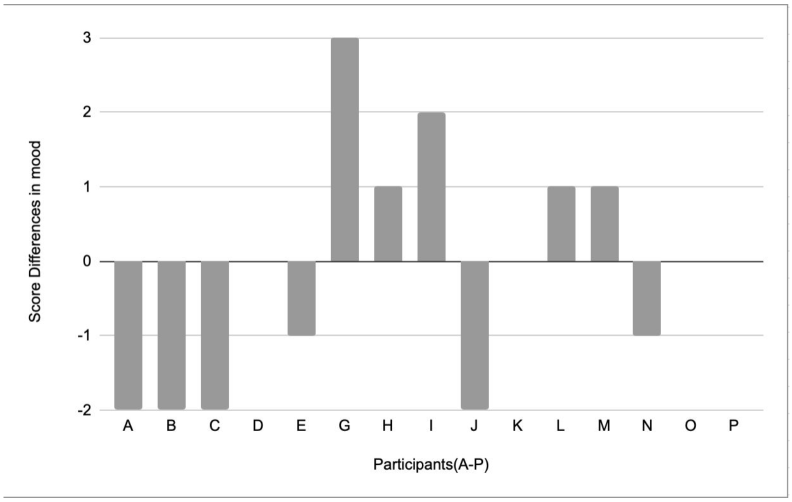
```
Leia:
 Is there any kind of ice cream you don't like ?
User:
I don't like strawberry ice cream.
Leia:
 Really ? Me , too . What's your favorite flavor ?
User:
Mint
Leia:
 Yes , I have . I tried them when I visited your country last year . I thought they were very tasty . Can you cook them ?
```

Appendix IV: Results of the users' mood difference before/after chatting with the chatbot

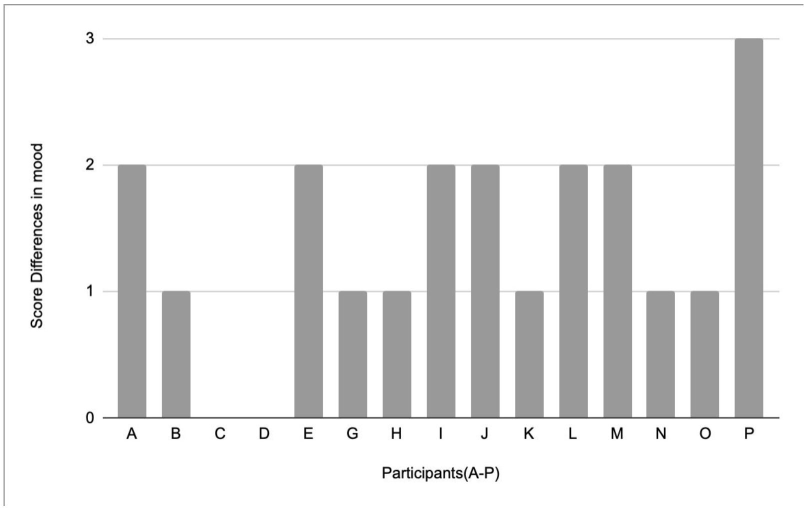
Eliza



Robo

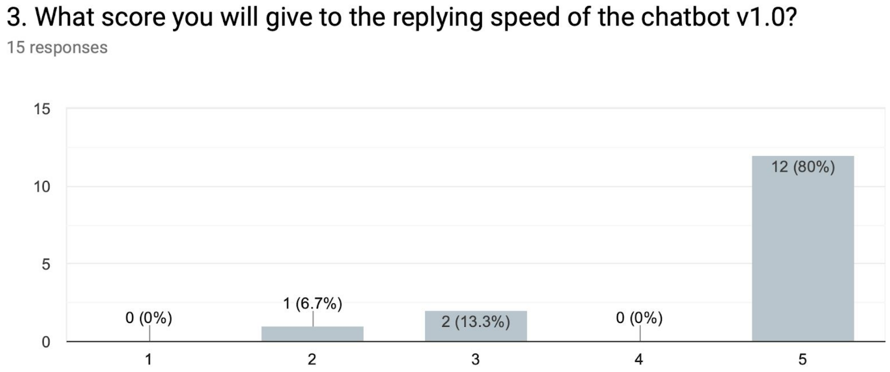


Leia



Appendix V: Ratings of the replying speed of the three chatbot models

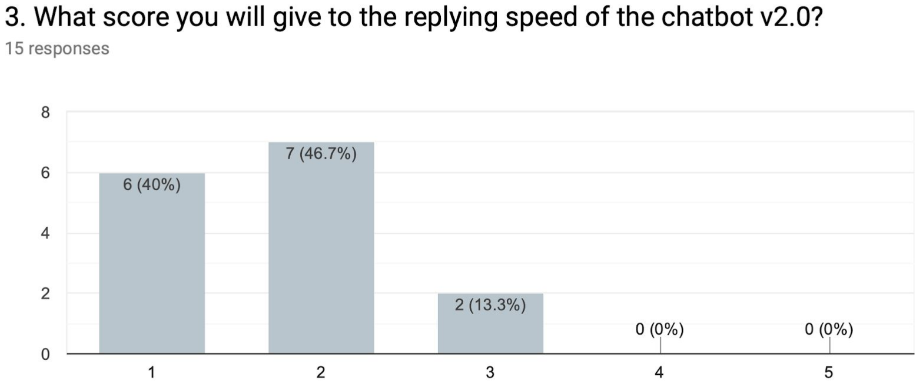
Eliza



3. What score you will give to the replying speed of the chatbot v1.0?

15 responses

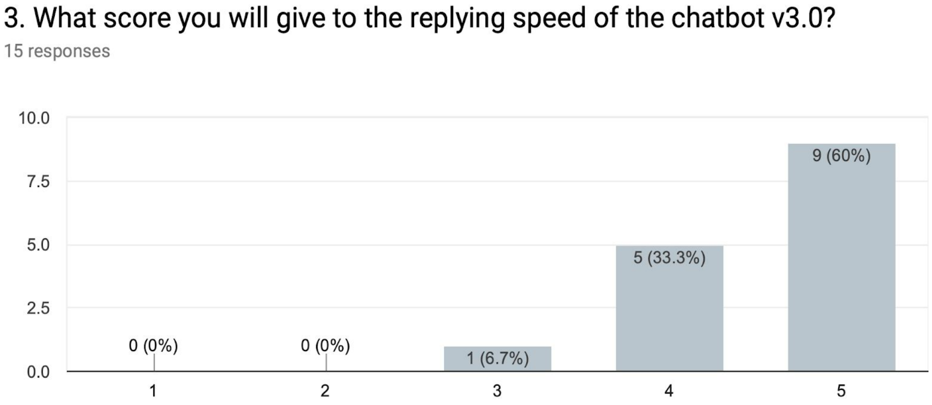
Robo



3. What score you will give to the replying speed of the chatbot v2.0?

15 responses

Leia

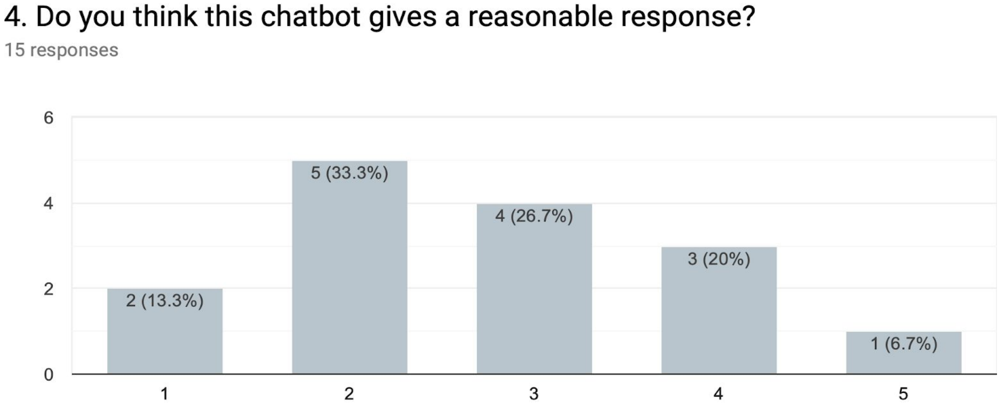


3. What score you will give to the replying speed of the chatbot v3.0?

15 responses

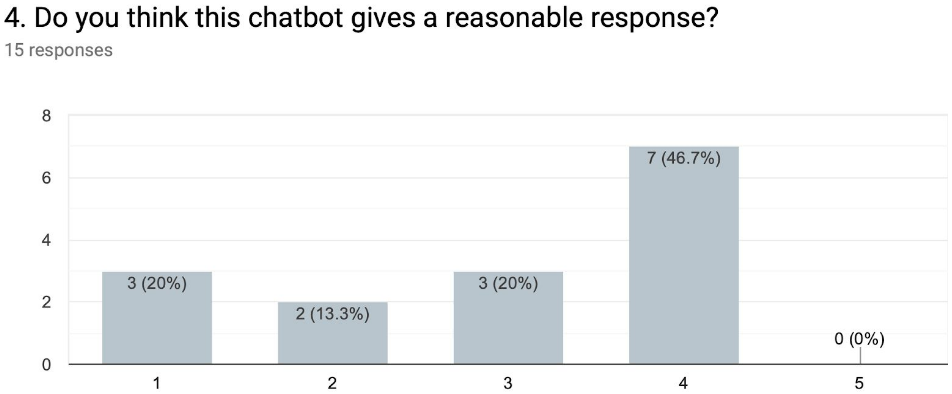
Appendix VI: Ratings of the response rationality of the three chatbot models

Eliza


4. Do you think this chatbot gives a reasonable response?

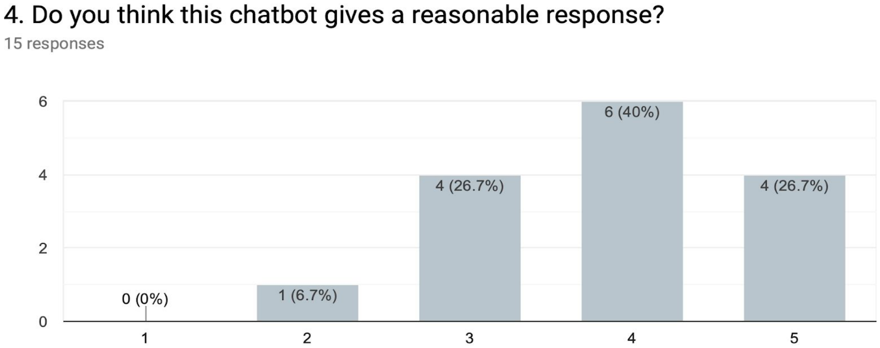
15 responses

Robo


4. Do you think this chatbot gives a reasonable response?

15 responses

Leia


4. Do you think this chatbot gives a reasonable response?

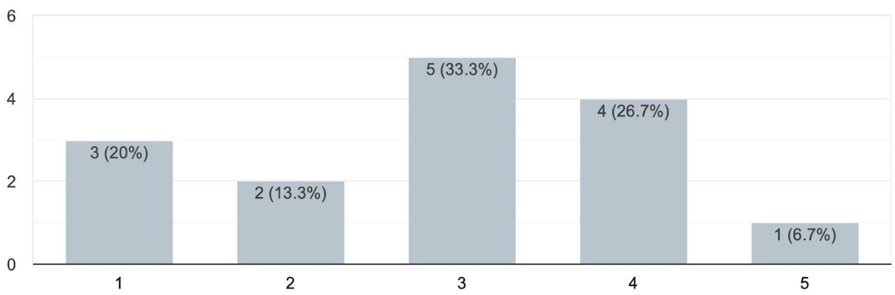
15 responses

Appendix VII: Overall Ratings of the three chatbot models

Eliza

5. How do you like the chatbot v1.0 overall?

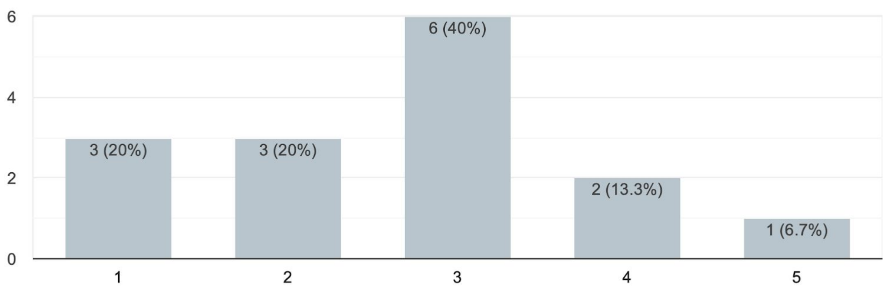
15 responses



Robo

5. How do you like the chatbot v2.0 overall?

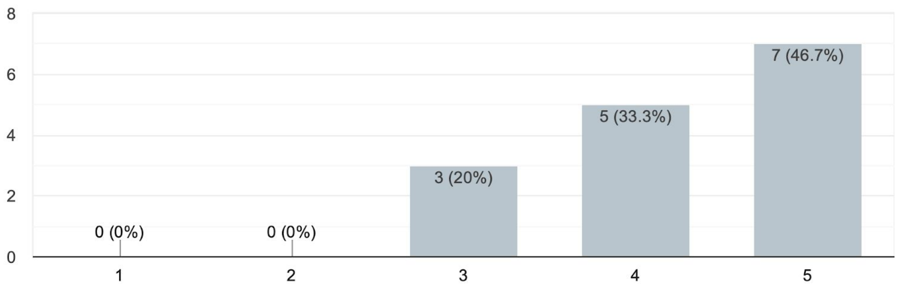
15 responses



Leia

5. How do you like the chatbot v3.0 overall?

15 responses


8
6
4
2
0
0 (0%)
0 (0%)
3 (20%)
5 (33.3%)
7 (46.7%)
1
2
3
4
5