Sentiment-Based Neural Dialogue System for Social Media Assistance

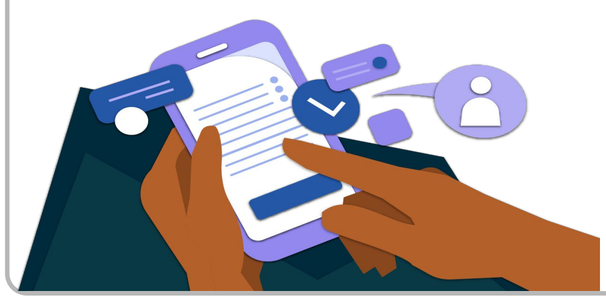
Group - 14
Jonnah Jagan Mohan, Shubham Jain, Chandini Velilani, Kavya Nayaka, Ajeya Hegde

USC Viterbi School of Engineering

Problem

Managing conversations in today's world is draining, and it can be hard to keep up with all the text messages and make sure replies are crisp and on time.

The proposed dialogue system acts as a proxy for the end-user responding to the incoming texts capturing the user's writing style.



Existing Systems

- Does not capture the writing style of the user.
- The sentiment of the incoming text is not considered.

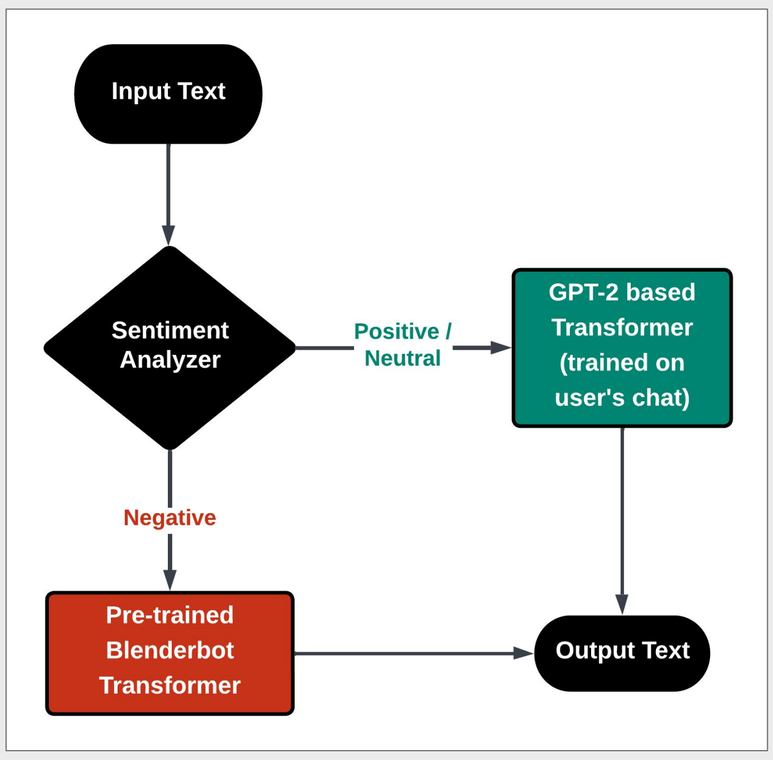
Corpus Processing

```
Human 1: Hi!
Human 2: How's it going?
Human 1: I'm so sleepy today!
Human 2: Not enough sleep last night?
Human 1: yeah was working all night on a homework
Human 2: Oh really? What class?
```

- The corpus is a human conversation dataset represented as a turn-based dialogue (obtained from Kaggle).
- Stop words were not removed.
- Emojis were represented by their corresponding textual meanings.

Solution

MODEL ARCHITECTURE



When the system receives a text, it is first parsed through a pre-trained sentiment analyzer model to examine the tone of the message.
Suppose the input text is friendly, or at least neutral; the fine-tuned transformer responds as the user would typically interact. Otherwise, we leverage the generic/discrete transformer.

How do we know if the model works?

We iterate through the dataset to obtain personality traits of the target human – their likes and dislikes such as them being adventurous, owning a pet cat, and use of informal abbreviations.

The response is evaluated against the above-mentioned traits to verify its coherence to the target human.

Not enough gold-references for automated evaluation metrics like BLEU, ROGUE to be useful.

```
Original line:  Human 1: same former owner?? 🙂
Transformed line:  Human 1: same former owner?? :slightly_smiling_face:
```

```
Original line:  Human 1: lol which cafe? let me guess. <REDACTED_TERM>?
Transformed line:  Human 1: lol which cafe? let me guess.
```

Results

We contrasted the base GPT-2-simple with our fine-tuned model.

Following are the results:

"Human 1" is our target…

Vanilla GPT-2-simple

```
print(row)

Question: Do you like roller coasters?
Response: Never ever!
```

```
print(row)

Question: Are you an extrovert?
Response: How did you come up with that whole thing?
```

Fine-tuned GPT-2-simple

```
print(res)

Sentiment Scores -
pos: 0.385
neu: 0.615
neg: 0.0

Calling Fine-Tuned Transformer for response -

Question: Do you like Roller Coasters?
Response: I think it's great! How could I not follow my dreams??
```

```
print(res)

Sentiment Scores -
pos: 0.0
neu: 1.0
neg: 0.0

Calling Fine-Tuned Transformer for response -

Question: Are you an extrovert?
Response: I am not! But I enjoy talking to people. What about you?
```

The discrete pre-trained transformer model reverts with a reasonable, neutral reply if the model receives a hateful prompt.

Test case of the system being self-censored.

```
print(res)

Sentiment Scores -
pos: 0.0
neu: 0.213
neg: 0.787

Calling Generic/Discrete Model for response -

Question: I hate you.
Response:  Why do you say that? I don't know what to say to make you feel better.
```

Conclusion

The model learns traits of 'Human 1' exhibited throughout the training set and the responses obtained from it are coherent with the personality of the target human. Whereas, the baseline is far more generic.

A concern beyond the score of this proposal is if there is a human commitment/decision to be made, the machine should not make a decision for the user.