CMSC 491/691: Interactive Fiction and Text Generation Fall 2024

Emotion-Consistent AI Characters in IF

Sukhbir Singh Sardar- ssardar1@umbc.edu | Maitri Mistry- mmistry3@umbc.edu | Mukesh Kumar Vidam- mukeshv1@umbc.edu

UMBC

Fine-tuning GPT-3.5 Turbo with DailyDialog dataset demonstrated higher emotional accuracy

Introduction

Interactive fiction relies on emotionally consistent characters. Our work aims to correct the frequent emotional misalignments in AI-generated narratives, enhancing both immersion and coherence.

Methods

The **DailyDialog** dataset, containing 13,118 dialogues annotated with:

- emotional states [no emotion (0), anger (1), disgust (2), fear (3), happiness (4), sadness (5) and surprise (6)] and
- communicative intents [__dummy__ (0), inform (1), question (2), directive (3) and commissive (4)].

A sample dialogue of dataset:

dialog: [’I was scared stiff of giving my first performance.’ ’Were you ? your performance was excellent.’ ’Thank you for your kindly words.’]

act: [1 1 1]

emotion: [3 4 4]

Addressed **data imbalance** by filtering out excessive 'no emotion' labels, resulting in a training set of 1,331 examples and a testing set of 1,000 examples.

Structured the dataset in a chat format to align with the input requirements of GPT-3.5 Turbo as shown in Figure 1 and Figure 2.

Fine-tuned GPT-3.5 Turbo to generate responses, and performance was assessed using VADER Sentimental Analysis, Cosine Similarity and Human Feedback

Results

Metric	GPT-3.5 Turbo scores
mean	0.1524
median	0.0824
std	0.2214

Table 1: Cosine Similarity Statistics

Cosine similarity scores revealed a low mean and median values, indicating lower similarity between generated and actual dialogues. Notably, this method measures word overlap without considering context, potentially underestimating true semantic similarities as it overlooks nuanced meanings behind word combinations.

Metric	GPT-3.5 Turbo scores
mean difference	0.2955
median difference	0.2394
std difference	0.2906

Table 2: VADER Analysis Statistics

Our analysis shows a mean difference of 0.2955, reflecting good sentiment similarity between generated and actual responses. With a median difference of 0.2394, most response pairs show minor discrepancies. While the system effectively captures overall sentiment, there is still notable potential for improvement.

Actual Emotion	Cosine Similarity ↑	VADER Difference ↓
Anger	0.1554	0.4418
Disgust	0.0389	0.2849
Happiness	0.1857	0.2571
No Emotion	0.1405	0.3052
Sadness	0.0979	0.4424
Surprise	0.1273	0.2641

Table 3: Analysis of emotion-wise performance comparing Cosine Similarity and VADER Difference metrics

The analysis reveals that the system performs best with happiness, both textually and emotionally, but struggles with negative emotions such as anger and sadness. Disgust shows somewhat better emotional handling but has the lowest textual similarity. These findings indicate a need for targeted improvements in managing negative and complex emotions to enhance response accuracy.

Emotion Tested	Average Naturalness	Average Coherence
Fear	4.2	4.0
Happiness	4.5	4.6
Sadness	3.8	3.9
Surprise	4.0	4.1
Anger	3.5	3.7
No Emotion	4.3	4.4

Table 4: Human Evaluation Results

The table presents human evaluation results of AI-generated dialogues, rating their naturalness and coherence across various emotions. Responses to happiness scenarios scored highest, indicating strong alignment with expected emotions. Conversely, dialogues involving anger and sadness received lower scores, highlighting areas where the AI struggles to maintain emotional accuracy and coherence, suggesting a need for further refinement in these emotional contexts.

```
{
  "messages": [
    {
      "role": "system",
      "content": "You are an AI that maintains
          emotional consistency in dialogue.
          Generate responses that match the
          expected emotion."
    },
    {
      "role": "user",
      "content": "Character A: I hear Mary and
          Sally often go there to play pingpong.
          Perhaps we can make a foursome with
          them . [Emotion: happiness]\nCharacter
           B: Sounds great to me ! If they are
          willing , we could ask them to go
          dancing with us.That is excellent
          exercise and fun , too . [Emotion:
          happiness]\nCharacter A: Good. Let's
          go now . [Emotion: happiness]"
    },
    {
      "role": "assistant",
      "content": "Character B: \"All right ..[
          Expected Emotion: happiness]\""
    }
  ]
}
```

Figure 1: Sample fine-tuning prompt in chat format

```
{
  "messages": [
    {
      "role": "system",
      "content": "You are an AI that maintains
          emotional consistency in dialogue.
          Generate responses that match the
          expected emotion."
    },
    {
      "role": "user",
      "content": "Character A: I think I heard
          someone outside the window. [Emotion:
          fear]\nCharacter B: Are you sure? Did
          you see anything? [Emotion: no emotion
          ]\nCharacter A: No, but it sounded
          like footsteps. [Emotion: fear]\
          nCharacter B: That's terrifying!
          Should we call someone? [Emotion: fear
          ]"
    }
  ]
}
```

Figure 2: Sample testing prompt in chat format