assignment1

September 2, 2024

1 Getting started with Generative-AI

Submitted By: Jenish Twayana

Submitted Date: 2nd September, 2024

1.1 Exercise 1: Sentiment Analysis and Moderation

Objective: Analyze the sentiment of given texts and identify any content that may require moderation based on predefined guidelines. Display the output in JSON format.

1.1.1 Set up the client and function

[ ]:
```
#install openai package
%pip install openai
```

[42]:
```
#importing necessary packages
from openai import OpenAI
import json
import os
```

[43]:
```
#initialize openai client
openai_api_key = os.getenv("OPENAI_API_KEY")
client = OpenAI(api_key=openai_api_key)
model = "gpt-3.5-turbo"
```

[44]:
```
#function for calling api
def get_completion(system_prompt, user_prompt, model="gpt-3.5-turbo"):
    messages = [
    {"role": "system", "content": system_prompt},
    {"role": "user", "content": user_prompt}
    ]
    response = client.chat.completions.create(
    model=model,
    messages=messages,
    temperature=0,
    )
    return response.choices[0].message.content
```

1.1.2 Step 1: Sentiment Analysis

Task: Given a list of text snippets, analyze the sentiment of each snippet (positive, negative, or neutral).

Texts:

1. "I had an amazing experience at the new restaurant downtown!"
2. "The service was terrible, and I will never go back."
3. "It was an okay movie, nothing special."
4. "The product quality is outstanding, I highly recommend it."
5. "I'm extremely disappointed with my purchase."
6. "The weather is nice today."

Instructions for Sentiment Analysis:

For each text in Step 1, determine whether the sentiment is positive, negative, or neutral. Explain the reasoning behind the classification.

[45]:
```
#list of sentiment analysis messages
sentiment_analysis_messages = [
"I had an amazing experience at the new restaurant downtown!",
"The service was terrible, and I will never go back.",
"It was an okay movie, nothing special.",
"The product quality is outstanding, I highly recommend it.",
"I'm extremely disappointed with my purchase.",
"The weather is nice today."
]
```

[46]:
```
#desired response in valid JSON format
sentiment_analysis_response_example = '''
{
"Text": "I had an amazing experience at the new restaurant downtown!",
"Sentiment": "Positive",
"Reasoning": "The use of words like "amazing" and "experience" in a
positive context indicates a positive sentiment."
}
'''
```

[47]:
```
#system prompt for sentiment analysis
sentiment_analysis_task_system_prompt = f'''
Classify the sentiment of the following text into one of the following:␣
 ↪positive, negative or neutral.
Provide the reasoning for the classification.
Here is an example:
```

```
Prompt: "I had an amazing experience at the new restaurant downtown!"
Desired Response in valid JSON format: {sentiment_analysis_response_example}
'''
```

[48]:
```
#generate responses and store in list
sentiment_analysis_responses_json = []
for message in sentiment_analysis_messages:
    response = get_completion(sentiment_analysis_task_system_prompt, message)
    print(response)
    sentiment_analysis_responses_json.append(json.loads(response))
```

```
{
"Text": "I had an amazing experience at the new restaurant downtown!",
"Sentiment": "Positive",
"Reasoning": "The use of words like 'amazing' and 'experience' in a positive
context indicates a positive sentiment."
}
{
"Text": "The service was terrible, and I will never go back.",
"Sentiment": "Negative",
"Reasoning": "The use of words like 'terrible' and 'never go back' express a
negative sentiment towards the service received."
}
{
"Text": "It was an okay movie, nothing special.",
"Sentiment": "Neutral",
"Reasoning": "The use of words like 'okay' and 'nothing special' suggests a
neutral sentiment, indicating a lack of strong positive or negative feelings
towards the movie."
}
{
"Text": "The product quality is outstanding, I highly recommend it.",
"Sentiment": "Positive",
"Reasoning": "The use of words like 'outstanding' and 'highly recommend' convey
a positive sentiment towards the product quality, indicating a positive
sentiment overall."
}
{
"Text": "I'm extremely disappointed with my purchase.",
"Sentiment": "Negative",
"Reasoning": "The use of words like 'disappointed' and 'extremely' conveys a
negative sentiment towards the purchase."
}
{
"Text": "The weather is nice today.",
"Sentiment": "Positive",
"Reasoning": "The word 'nice' indicates a positive sentiment towards the
```

```
weather."
}
```

1.1.3 Step 2: Content Moderation

Task: Identify any text snippets that may contain harmful, offensive, or inappropriate content and suggest moderation actions.

Texts:

1. "I can't believe they hired such an incompetent person!"
2. "This is a wonderful community, and I love being part of it."
3. "The event was a disaster; the organizers did a horrible job."
4. "You are such an idiot for thinking that!"
5. "I support everyone who works hard to achieve their dreams."
6. "Get lost, no one wants you here!"

Instructions for Content Moderation:

> For each text in Step 2, identify if the content is harmful, offensive, or inappropriate. Suggest a moderation action (e.g., warning, content removal, etc.) and explain why.

[49]:
```
#list of content_moderation messages
content_moderation_messages = [
"I can't believe they hired such an incompetent person!",
"This is a wonderful community, and I love being part of it.",
"The event was a disaster; the organizers did a horrible job.",
"You are such an idiot for thinking that!",
"I support everyone who works hard to achieve their dreams.",
"Get lost, no one wants you here!"
]
```

[50]:
```
#desired response in valid JSON format
content_moderation_response_example = '''
{
"Text": "You are such an idiot for thinking that!",
"Identified Issue": "Offensive content",
"Moderation Action": "Warning or content removal"
"Reasoning": "The text includes name-calling and derogatory language,
which is inappropriate and harmful."
}
'''
```

[56]:
```
#system prompt for content moderation
content_moderation_task_system_prompt = f'''
```

```
Identify if the following text is offensive, harmful or inappropriate.
Provide the reasoning as well as the appropriate moderation action. The␣
 ↪moderation action can be either a warning or content removal.
Here is an example:
Prompt: "You are such an idiot for thinking that!"
Desired Response in valid JSON format: {content_moderation_response_example}
The "Identified Issue" and "Moderation Action" fields should be "N/A" if the␣
 ↪text is not offensive, harmful or inappropriate.
'''
```

[57]:
```
#generate responses and store in list
content_moderation_responses_json = []
for message in content_moderation_messages:
    response = get_completion(content_moderation_task_system_prompt, message)
    print(response)
    content_moderation_responses_json.append(json.loads(response))
```

```
{
"Text": "I can't believe they hired such an incompetent person!",
"Identified Issue": "Offensive content",
"Moderation Action": "Warning",
"Reasoning": "The text contains derogatory language ('incompetent person') which
can be harmful and offensive."
}
{
"Text": "This is a wonderful community, and I love being part of it.",
"Identified Issue": "N/A",
"Moderation Action": "N/A",
"Reasoning": "The text is positive and does not contain any offensive, harmful,
or inappropriate content."
}
{
"Text": "The event was a disaster; the organizers did a horrible job.",
"Identified Issue": "N/A",
"Moderation Action": "N/A",
"Reasoning": "The text expresses a negative opinion about the event and the
organizers, but it does not contain any offensive, harmful, or inappropriate
language."
}
{
"Text": "You are such an idiot for thinking that!",
"Identified Issue": "Offensive content",
"Moderation Action": "Warning or content removal",
"Reasoning": "The text includes name-calling and derogatory language, which is
inappropriate and harmful."
}
```

```
{
"Text": "I support everyone who works hard to achieve their dreams.",
"Identified Issue": "N/A",
"Moderation Action": "N/A",
"Reasoning": "The text is positive and supportive, promoting hard work and
achievement."
}
{
"Text": "Get lost, no one wants you here!",
"Identified Issue": "Offensive content",
"Moderation Action": "Warning",
"Reasoning": "The text is hostile and dismissive, creating a negative and
unwelcoming environment. It can be hurtful and harmful to the recipient."
}
```