Understand Travel Mode Changes through LLMs and In-context Learning Final Report

Xinyue Zhang
Master of Science in Operations Research, Columbia University
xz3250@columbia.edu
May 12, 2024

Abstract

This project explores the application of Large Language Models (LLMs) to extract insights on travel modes, sentiments, and reasons from unstructured tweet data. Employing advanced techniques such as in-context learning, chain-of-thought reasoning, and demonstrations, we enhanced the precision and relevance of our outputs. Challenges like managing structured response formats and addressing overfitting were tackled through continuous prompt refinement and the integration of human verification, which proved crucial for accuracy. Our findings indicate improved satisfaction rates due to more effective sentiment categorization, although extraction rates decreased. Future work will focus on employing diverse LLMs for cross-verification to mitigate bias and improve reliability.

1. Introduction

This project explores the potential of LLMs, specifically GPT-3.5-turbo, to analyze unlabelled travel patterns through social media (Twitter) data. By utilizing advanced techniques such as in-context learning and chain-of-thought reasoning, this study aims to extract the travel mode from tweets, assess sentiments towards these travel modes, and uncover underlying reasons for these sentiments. Additionally, the project seeks to understand changes in travel mode preferences over time.

The methodology integrates several approaches: in-context learning for initial data extraction, self-verification using GPT-3.5-turbo, and cross-verification with GPT-4-turbo and LLAMA 2 to ensure accuracy. Human verification is also employed to identify any discrepancies in the LLM outputs. Through iterative refinement of prompts and continuous model tuning, the project endeavors to enhance the accuracy of the extracted data.

The report is structured as follows: Section 2 delves into a literature review, discussing methodologies such as sentiment analysis, Chain-of-Thought, and In-Context Learning. Section 3 provides a summary of the coding sources and the progress made. Section 4 outlines the methodologies employed, detailing the use of LLMs. Section 5 presents experimental results and discusses the iterative process of prompt refinement. Section 6 compares the results across different experiments, focusing on travel mode and sentiment and their correlations. Section 7 concludes the report with the key findings and implications of using LLMs for travel mode from tweets. Finally, Section 8 is Acknowledgments.

2. Literature review

- **In-Context Learning:** [1] explores how variations in the demonstration format rather than precise input-label matches can significantly enhance model performance. This study emphasizes that LLMs can efficiently process unlabelled data by focusing on the format of demonstrations, thus facilitating more effective learning of travel patterns without explicit input-label mapping.
- **Chain-of-Thought(CoT):** Detailed in [2], this approach breaks down complex tasks into manageable components using natural language rationales. This method not only improves interpretability and accuracy in tasks involving intricate reasoning but also supports the comprehension of nuanced travel behaviors, enhancing the LLMs' ability to analyze unstructured data for travel mode insights.
- **Sentiment Analysis:** [3] utilizes Twitter data to assess public sentiment towards various travel modes during the pandemic. By employing NLP techniques and sentiment analysis, the research highlights the power of social media as a tool for understanding changes in travel behaviors and public attitudes, proving essential for informing transportation policy.
- **Tree-of-Thought (ToT):** [4] introduces a structured problem-solving approach that outperforms traditional methods by organizing intermediate steps into a tree structure. This method allows for exploring multiple reasoning pathways simultaneously, which is crucial for identifying subtle shifts in travel mode preferences and understanding public sentiments towards different travel modes.

3. Summary

The weekly code, literature review, and CSV file can be found in Google Drive. The weekly progress can be found in the Weekly Progress Report. The final code and related documents can be found on GitHub. The experiment code can be found in the GitHub folder code experiment.

4. Methodology

This section outlines the methodology employed to analyze unlabelled travel data from Twitter using various LLMs, specifically focusing on extracting travel modes, sentiments, and underlying reasons from tweets.

(1) Step 1: Extraction

The initial phase involves using GPT-3.5-turbo to identify travel modes mentioned in tweets. The model is prompted to answer three key questions:

- What is the corresponding travel mode in this tweet? Possible answers include Irrelevant (NA), subway, bus, bike, taxi, or car.
- What is the user sentiment regarding the travel mode's service? The response can be positive, negative, or neutral.

- If the tweet is related to a specific travel mode, what are the reasons for the sentiment expressed? This inquiry requires a detailed explanation of the factors influencing the sentiment, whether positive or negative.

The methodology initially integrates chain-of-thought reasoning to structure the model's output systematically, followed by in-context learning with demonstrations to refine the answers in multiple prompts style. This approach ensures that the output is precise and relevant. Prompts are designed to elicit close-ended responses in a JSON format with lowercase outputs, enhancing the processability and consistency of the data extracted.

(2) Step 2: Verification and Refinement

The extracted data undergoes a three-part verification process:

- **Self-Verification:** GPT-3.5-turbo checks its outputs for accuracy and consistency.
- **Cross-Verification:** GPT-4-turbo and LLAMA-2-7b are used to assess the correctness of the initial outputs, providing a broader evaluation scope and identifying discrepancies missed by the first model.
- **Human Verification:** Humans also review outputs to ensure the output makes sense contextually and factually.

Feedback from these verification steps is used to refine and improve the prompt designs in Step 1. This iterative process continues until the outputs meet the desired standards of accuracy and relevance.

(3) Step 3: Output Analysis

The final step involves analyzing the verified data to explore travel modes and sentiments expressed in tweets. This analysis includes comparing results from different prompt styles and approaches to identify the most effective methods for extracting and understanding travel patterns and user sentiments by metrics or visualizations. This comprehensive analysis helps in understanding how travel preferences and public sentiment towards various modes of transportation are articulated on social media.

5. Experiment Process

(1) Exploration

The exploration phase used prompts to inquire directly about the travel mode, sentiment, and reasons behind the sentiment from tweets. The initial results indicated that outputs were in sentence form rather than discrete words, showing a need for more structured responses.

```
[ ]  response = openai.chat.completions.create(
        model="gpt-3.5-turbo",
        messages=[
            {"role": "system", "content": "You are a helpful assistant."},
            {"role": "user", "content": f"Given the tweet: \"{prepared_tweet}\""
                " 1. What is the corresponding travel mode? (Train, subway, bus, bike, taxi, Uber, private car)"
                " 2. Is the user satisfied with the corresponding travel mode's service? (Yes or No)"
                " 3. If not satisfied, what are the reasons behind?"}
        ],
        temperature=0.5,
        max_tokens=100
    )
```

```
[ ] print(response.choices[0].message.content)

1. The corresponding travel mode is a shuttle bus.
2. No, the user is not satisfied with the shuttle bus service.
3. The user is dissatisfied because they have been waiting for almost an hour, and no shuttle bus has shown up despite being just one stop
```

(2) Baseline Model

Feedback from the Exploration phase led to the structuring of outputs in JSON format to facilitate smoother extraction into a data framework. The revised prompt was designed to produce outputs like {"travel_mode": "<inferred mode>", "sentiment": "<inferred sentiment>", "reasons": "<inferred reasons>"}. Despite these changes, the output included 61 possible travel modes and 19 different sentiments, indicating a lack of structure in categorizing travel modes and sentiments.

```
[ ] def generate_base(tweet):
        tweet = tweet.lower()  # Ensure the tweet is in lowercase
        response = openai.chat.completions.create(
            model="gpt-3.5-turbo",
            messages=[
                {
                    "role": "system",
                    "content": "You are an AI trained to extract travel experiences from tweets."
                },
                {
                    "role": "user",
                    "content": f"Given the tweet: \"{tweet}\""
                               " 1. What is the corresponding travel mode? (Train, subway, bus, bike, taxi, Uber, private car)"
                               " 2. Is the user satisfied with the corresponding travel mode's service? (Yes or No)"
                               " 3. If not satisfied, what are the reasons behind?"
                               " Conclude with a concise, structured summary (keep this format for all your response) in JSON format,"
                               " like this: "
                               "{\"travel_mode\": \"<inferred mode>\", \"sentiment\": \"<inferred sentiment>\", \"reasons\": \"<inferred reasons>\"}
                }
            ],
            temperature=0.5,
            max_tokens=250
        )

        return response.choices[0].message.content if response.choices else "No response"
```

```
Number of distinct outcomes for Travel Mode: 61
Counts for each distinct outcome in Travel Mode: private car    758
car                                 60
taxi                                23
train                               22
subway                              19
                                   ...
vehicle                              1
Rental car                           1
horses                               1
ride-sharing service (Lyft/Uber)     1
plane                                1
```

```
Number of distinct outcomes for Sentiment: 19
Counts for each distinct outcome in Sentiment: No    534
Neutral             175
Yes                  80
positive             22
Satisfied            14
Positive             10
neutral               9
Negative              7
Not satisfied         4
satisfied             3
Not applicable        2
Mixed                 2
unknown               2
Not Applicable        2
unsatisfied           1
Not mentioned         1
Unknown               1
Not specified         1
negative              1
```

(3) Guide Reasoning

Following the feedback from the Baseline Model phase, the prompts were enhanced to better guide the model's reasoning process. Specific scenarios, such as "multiple modes of transport," were included to refine the scope of responses. The model was also instructed to categorize sentiments more distinctly as "positive," "negative," or "neutral." Despite these enhancements, the results still manifested in an open format, which did not strictly conform to the discrete options provided. Moreover, an overfitting issue became apparent, as indicated by an extraction rate of 1 (excluding cases marked as NA), suggesting that the model was overly confident in its responses, often without sufficient justification from the tweet's content. To address this, further refinements focused on directing the model to adhere strictly

to the information available in the tweets, thereby avoiding over-extrapolation and enhancing the precision of the output.

```
messages=[
    {
        "role": "system",
        "content":
            "You're an assistant skilled in understanding public transport experiences from tweets. "
            "Provide an analysis based on the tweet."
    },
    {
        "role": "user",
        "content": (
            f"Given the tweet: '{tweet}', carefully analyze the content and context to determine the travel experience "
            "shared by the tweet's author. Considering that multiple modes of transportation might be mentioned, "
            "focus on identifying the primary mode the author is describing as their own experience. "
            "Follow these steps in your analysis:\n\n"
            "1. **Identify the Primary Mode of Transportation:** Examine the tweet for mentions of travel modes "
            "(car, subway, bus, bike, taxi). Determine which mode the author specifically refers to as their own "
            "means of transportation, if directly mentioned or implied through context.\n"
            "2. **Assess Sentiment:** Once the primary mode is identified, evaluate the sentiment the author "
            "expresses towards this mode of transportation—positive, negative, or neutral.\n"
            "3. **Extract Reasons:** Delve into the reasons behind the author's sentiment towards the primary "
            "travel mode. Identify any specific experiences, aspects, or observations mentioned that influence "
            "their view.\n\n"
            "Conclude your analysis with a structured summary of your findings in the following format: "
            "{travel_mode: '<mode>', sentiment: '<sentiment>', reasons: '<reasons>'}."
        )
    }
],
```

```
Number of distinct outcomes for Travel Mode: 259
Counts for each distinct outcome in Travel Mode: car                                    218
taxi                                        119
Uber                                         29
Lyft                                         27
Taxi                                         25
                                           ...
driving (as an independent contractor)        1
Taxi/Ride-sharing Service                     1
Water-based transportation                    1
Tesla (electric car)                          1
public transportation (subway or bus)         1
```

```
{'Extraction Rate': 1.0, 'Satisfaction Rate': 0.147}
```

(4) Overfitting & Underfitting

Responding to the guidance from the Guide Reasoning phase, a command was added to the prompt directing the model to avoid assumptions unsupported by the tweet's content, targeting the overfitting issue. Subsequent experiments revealed an underfitting issue when focusing solely on avoiding overfitting. To address this, the command was refined to include "ensuring to provide as much detail as possible without defaulting to 'na' unless absolutely necessary." Additionally, the temperature was set to 0.7 to foster more nuanced judgments in cases involving multiple travel modes. A directive, "Consistency and evidence adherence are key," was introduced to enhance the reliability of outputs across all tweets. Despite these adjustments, while the number of recognized travel modes decreased to a structured 19 and sentiment types to 7, the extraction rate stood at 0.952, showing a significant improvement though indicating potential bias in self-verification by the same LLM. This necessitated further human verification.

```
tweet = tweet.lower()  # Convert tweet to lowercase
system_message = (
    "You are an AI specifically trained to extract travel experiences from tweets, focusing "
    "on the mode of transportation, the sentiment towards it, and the reasons behind the sentiment. "
    "It's crucial to maintain consistency in your analysis, avoiding assumptions not supported by "
    "the tweet's content (overfitting), yet also ensuring to provide as much detail as possible "
    "without defaulting to 'na' unless absolutely necessary (underfitting). Your responses should "
    "be in lowercase and adhere to the specified categories."
)
user_message = (
    f"Analyze the tweet: '{tweet}'. Your goal is to clearly identify the travel experience it describes, "
    "focusing on these aspects:\n\n"
    f"Given the tweet: \"{tweet}\" "
    "1. What is the corresponding travel mode? (Train, subway, bus, bike, taxi, Uber, private car) "
    "2. Is the user satisfied with the corresponding travel mode's service? (Yes or No) "
    "3. If not satisfied, what are the reasons behind? "
    "Conclude with a concise, structured summary in JSON format, like this: "
    "{\"travel_mode\": \"<inferred mode>\", \"sentiment\": \"<inferred sentiment>\", \"reasons\": \"<inferred reasons>\"}. "
    "Remember, consistency and adherence to the evidence presented in the tweet are key."
)
response = openai.chat.completions.create(
    model="gpt-3.5-turbo",
    messages=[
        {"role": "system", "content": system_message},
        {"role": "user", "content": user_message}
    ],
    temperature=0.7,
    max_tokens=250
)
```

```
Number of distinct outcomes for Travel Mode: 19
Counts for each distinct outcome in Travel Mode: car          867
na              48
bike            18
subway          18
train           11
walking          8
scooter          6
taxi             4
plane            3
bus              2
helicopter       2
flying cars      1
boat             1
pedestrians      1
flight           1
skates           1
motorcycle       1
transit          1
horse            1
```

```
Number of distinct outcomes for Sentiment: 7
Counts for each distinct outcome in Sentiment: negative       557
positive              260
neutral               158
na                     20
mixed                   3
surprise                1
potentially negative    1
Name: Sentiment, dtype: int64

{'Extraction Rate': 0.952,
 'Satisfaction Rate': 0.26,
 'Self Verification Rate': 1.0}
```

(5) Feedback from Human Verification

The prompt structure developed in the Overfitting & Underfitting phase was retained due to its enhanced precision. Human verification, however, revealed discrepancies, notably in tweets with multiple travel modes. For instance, a tweet mentioning "the best driver t769489c blocked the bike lane near 274 dean st on July 6" was inaccurately classified under 'bike', whereas it pertained more to 'car'. To rectify such issues, the prompt was adjusted to "If multiple modes focus on the most emphasized," guiding the model to prioritize the dominant travel mode in ambiguous cases. This refinement successfully narrowed down sentiment options to 6 but resulted in a proliferation of travel modes to 35 due to more specific categorizations like 'taxi', 'Uber', and 'Lyft', increasing the complexity. This required further prompt adjustments, and the extraction rate decreased to 0.743, indicating more accurate but diverse outcomes.

```
    system_message = (
        "You are an AI trained to extract travel experiences from tweets, focusing on "
        "the mode of transportation, sentiment, and reasons behind the sentiment. Ensure consistency "
        "in your analysis, avoid unsupported assumptions, and provide detail without defaulting to 'na'. "
        "All responses should be in lowercase and adhere to the specified categories."
    )
    user_message = (
        f"Analyze the tweet: '{tweet}'. Focus on:\n"
        "1. **Mode of Transportation:** Identify the primary mode (car, subway, bus, bike, taxi, uber, lyft, or 'na'). "
        "Use context clues for indirect mentions. If multiple modes, focus on the most emphasized.\n"
        "2. **Sentiment Analysis:** Determine the sentiment (positive, negative, neutral). "
        "Use context if unclear or default to 'na' if not possible to infer.\n"
        "3. **Reasons Behind Sentiment:** Extract reasons or factors mentioned, summarize contextual hints, "
        "or 'na' if no clear reasons are stated.\n"
        "Conclude with a JSON summary: "
        "{\"travel_mode\": \"<mode>\", \"sentiment\": \"<sentiment>\", \"reasons\": \"<reasons>\"}. "
        "Consistency and evidence adherence are key."
    )
```

```
Number of distinct outcomes for Travel Mode: 35
Counts for each distinct outcome in Travel Mode: car                          300
na                          257
taxi                        209
uber                         60
bike                         42
lyft                         42
bus                          16
walking                      13
subway                        9
plane                         8
train                         5
```

```
Number of distinct outcomes for Sentiment: 6
Counts for each distinct outcome in Sentiment: negative     564
positive     249
na           109
neutral       75
mixed          2
confusion      1
```

```
{'Extraction Rate': 0.743,
 'Satisfaction Rate': 0.249,
 'Self Verification Rate': 0.998}
```

(6) Add Demonstration to Extraction

After a thorough analysis of outputs and Human Verification Feedback, demonstrations were introduced to guide the model more effectively. The new prompts incorporated examples of incorrect outputs identified during human verification, addressing key issues like the misidentification of travel modes in tweets concerning multiple travel modes or on-hailing services like Uber, often misconstrued as travel modes rather than delivery services. Valid modes and sentiments were predefined to streamline the model's choices to a structured format. As a result, the output was refined to 8 travel modes and 4 sentiments, closely aligning with the desired structured responses, demonstrating the effectiveness of this approach to refining prompts.

```
system_message = (
    "You are an AI trained to extract travel experiences from tweets. Use these demonstrators for guidance:\n"
    "- Example 1: 'The best driver t664893c blocked the bike lane near 341 vanderbilt ave on june 30...' "
    "Expected: {'travel_mode': 'car', 'sentiment': 'negative', 'reasons': 'driver blocks the bike lane'}\n"
    "- Example 2: 'Hi. Someone ordered food through you and it was delivered to my door by mistake...' "
    "Expected: {'travel_mode': 'na', 'sentiment': 'na', 'reasons': 'na'}\n"
    "- Example 3: '...he is driving so slow...it is a 12 min driving he is making it a 20 min drive.' "
    "Expected: {'travel_mode': 'taxi', 'sentiment': 'negative', 'reasons': 'long time to drive'}\n"
    "Avoid assumptions not supported by the tweet's content and do not default to 'na' unless absolutely necessary. "
    "Focus on consistency and evidence-based analysis."
)
user_message = (
    f"Analyze the tweet: '{tweet}'. Focus on:\n"
    "1. **Mode of Transportation:** Identify the primary mode (choices: 'car', 'subway', 'bus', 'bike', 'uber', 'lyft', 'taxi', 'na').\n"
    "2. **Sentiment Analysis:** Determine the sentiment ('positive', 'negative', 'neutral').\n"
    "3. **Reasoning Behind Sentiment:** Extract reasons influencing the sentiment or use 'na'.\n"
    "Provide a structured summary in JSON format."
)
```

```
# Validate responses to ensure they are within acceptable categories
valid_modes = ['car', 'subway', 'bus', 'bike', 'uber', 'lyft', 'taxi', 'na']
valid_sentiments = ['positive', 'negative', 'neutral']
result['travel_mode'] = result['travel_mode'] if result['travel_mode'] in valid_modes else 'na'
result['sentiment'] = result['sentiment'] if result['sentiment'] in valid_sentiments else 'na'
```

```
Distinct outcomes for Travel Mode: ['car' 'na' 'uber' 'taxi' 'lyft' 'subway' 'bus' 'bike']
Number of distinct outcomes for Travel Mode: 8
Counts for each distinct outcome in Travel Mode: na        396
car       260
taxi      189
uber       80
lyft       41
bike       15
bus        11
subway      8
```

```
Distinct outcomes for Sentiment: ['positive' nan 'negative' 'neutral' 'na']
Number of distinct outcomes for Sentiment: 4
Counts for each distinct outcome in Sentiment: negative    378
positive    179
neutral      46
na            1
```

(7) Add Demonstration to Verification

Leveraging the successes from the Extraction phase, demonstration-based prompting was extended to the verification process to mitigate bias and enhance the generalizability of the findings. Different demonstrations were used for this phase to ensure a broad evaluation scope. Three LLMs were engaged for verification: GPT-3.5-turbo conducted self-verification, while GPT-4-turbo and LLAMA 2 performed cross-verification. Initial testing indicated that GPT-4-turbo was more time-efficient than LLAMA 2, leading to its selection as the primary tool for cross-verification moving forward.

- **Partial Prompts of GPT-3.5-turbo and GPT-4-turbo:**

```
verification_prompt = (
    "System: You are a highly intelligent assistant trained to verify the accuracy of travel mode analysis. "
    "Provide your verification based on the analysis and instructions below:\n\n"
    "Instructions: Assess the analysis for correctness based on relevance and accuracy of the travel mode, "
    "sentiment, and reasons given. Respond with 'correct', 'incorrect', or 'undetermined'. All responses must be in lowercase.\n\n"
    "Examples:\n"
    "- Example 1: Tweet: 'the best eats and doordash drivers abuse tf out this button https: and and t.co and xefkdiaef6' "
    "Analysis: Travel Mode: 'car', Sentiment: 'negative', Reasons: 'drivers abuse the button'. "
    "Verification: 'correct'\n"
    "- Example 2: Tweet: 'i do not give a fuck about my the best rating the best drivers make it so hard to sit still' "
    "Analysis: Travel Mode: 'na', Sentiment: '', Reasons: ''. "
    "Verification: 'incorrect'\n\n"
    f"Given Analysis:\n"
    f"- Tweet: {tweet}\n"
    f"- Travel Mode: {travel_mode}\n"
    f"- Sentiment: {sentiment}\n"
    f"- Reasons: {reasons}\n\n"
    "Is this analysis 'correct' or 'incorrect' or 'undetermined'?\n"
)
```

- **Partial Prompts of LLAMA 2:**

```
# Example prompt demonstrating the output we are looking for
example1_prompt = """
Based on the following tweet analysis:
    Tweet: {tweet}
    - Travel Mode: {travel_mode}
    - Sentiment: {sentiment}
    - Reasons: {reasons}

    Is this analysis 'correct' or 'incorrect' or 'undetermined'? Consider the relevance and accuracy of the travel mode, sentiment, and reasons

Here is the example I can provide with you.
The tweet is:
"the best eats and doordash drivers abuse tf out this button https: and and t.co and xefkdiaef6"
The travel_mode is:"car".The sentiment is: "negative". The reasons is: "drivers abuse the button".

[/INST]
{{"Cross Verification": "correct, "Cross Verification Reason":"since the drivers have been mentioned in the tweet that have abuse button." }}
</s>
"""
```

6. Result & Comparison

(1) Travel Mode & Sentiment Distribution

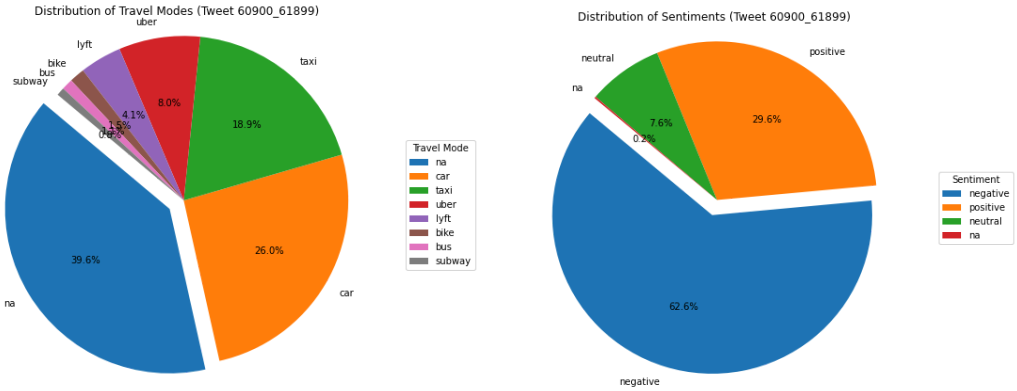

Figure Set 1: Distribution of Travel Mode & Sentiment

The first pie chart showcases the distribution of travel modes within the dataset. From the chart, it's evident that 'car' is the most popular travel mode, accounting for a significant 26.0% of the entries, followed by 'taxi' at 18.9%. This indicates a substantial preference for private modes of transportation over public ones like 'bus' and 'subway', which appear less frequently. The second pie chart focuses on the distribution of sentiments associated with the travel modes. Here, 'negative' sentiments are the most prevalent, making up 62.6% of the travel mode, closely followed by 'positive' sentiments at 29.6%.

(2) Correlation: Travel Mode and Sentiment

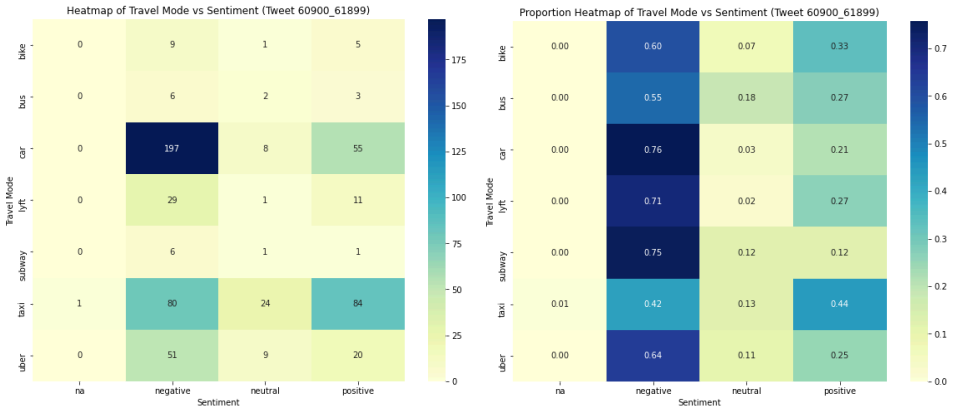

Figure Set 2: HeatMap of Travel Mode & Sentiment

The provided Figure Set 2 displays the correlation between different travel modes and sentiments. The first heatmap shows absolute counts, where 'car' exhibits the highest negative sentiment (197), indicating significant dissatisfaction related to car. Conversely, 'taxi' shows a relatively balanced sentiment distribution with 84 positive sentiments, highlighting a more favorable view among users. The second heatmap, which represents proportions, reveals that negative sentiments predominantly occur with most travel modes, especially 'car' and 'subway', which show high negative proportions (0.76 and 0.75). This suggests underlying challenges or negative perceptions associated with these modes. Interestingly, taxis have the most balanced sentiment distribution, with a substantial proportion of positive feedback (0.44), suggesting a more satisfactory user experience. These are also revealed in Figure 3.

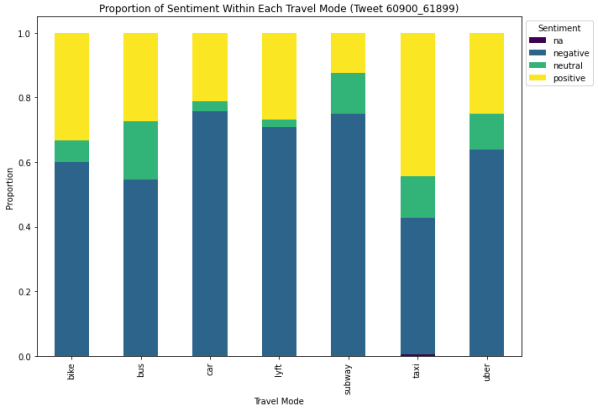

Figure 3: Proportion of Sentiment in Travel Mode

(3) Tweet Length and Sentiment

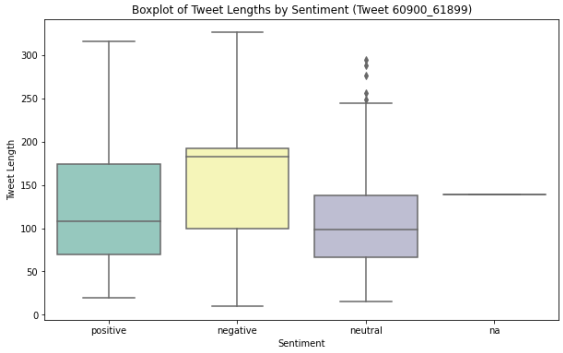

Figure 4: Boxplot of Tweets Length & Sentiment

The boxplot illustrates tweet lengths by sentiment, showing that negative sentiments generally feature longer tweets, suggesting detailed expressions of dissatisfaction. Neutral sentiments exhibit a narrower range, indicating more concise expressions, while positive sentiments, though similar in range to negative, have a lower median, implying that positive feelings are often expressed more succinctly.

(4) Baseline Model & Final Experiment

The bar chart compares performance metrics across baseline and final experiment models applied to two distinct datasets.

- **Extraction Rate:** Notably, there's a decrease in the extraction rate in the final models compared to the baseline. This reduction likely results from addressing overfitting issues and more stringent criteria for classifying non-'na' travel modes, leading to a cleaner, more accurate dataset.
- **Satisfaction Rate:** There is a marked increase in satisfaction rates in the final models. This improvement stems from refining sentiment queries into a closed format, where responses align more consistently with positive sentiments, enhancing the clarity and reliability of sentiment analysis.
- **Verification Rate:** Both self-verification and cross-verification rates remain high across the models. These high rates suggest potential model biases and underscore the need for using diverse LLMs for verification to bolster result reliability.

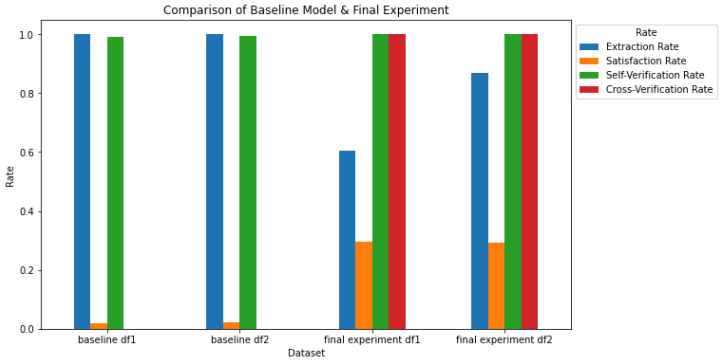

Figure 5: Comparison of Baseline & Final Experiment

The baseline model did not incorporate cross-verification, reflecting its initial state. To increase reliability, exploring further advanced LLMs for verification might be beneficial, potentially offering a broader perspective and mitigating bias. These findings underline the importance of refining model prompts and verification methods to enhance the accuracy and reliability of sentiment analysis in travel-related tweets.

7. Conclusion

Throughout this project, we explored the capabilities of LLMs to extract and analyze travel modes, sentiment, and underlying reasons from unstructured tweet data. Our methodology integrated advanced techniques such as in-context learning, CoT reasoning, and the use of demonstrations, significantly enhancing the precision and relevance of our outputs.

During the experimental process, we encountered challenges including managing structured response formats, handling tweets with multiple travel modes, and addressing issues of overfitting and underfitting. To address these challenges, we continuously refined our prompts, utilized reasoning guides, and incorporated demonstrations to direct the LLMs toward more accurate and contextually relevant outputs after receiving feedback from verification processes, particularly human verification.

The results demonstrated improvements in satisfaction rates, attributed to the transition from open to closed-form responses in sentiment analysis. This adjustment facilitated a more straightforward categorization of sentiments, boosting the rate of positive (satisfaction). However, we also observed a reduction in extraction rates, likely due to the more stringent criteria established to combat overfitting, effectively reducing the incidence of erroneous travel mode categorizations.

Both self-verification and cross-verification rates were high, suggesting a potential internal bias within the LLMs used. This finding underscores the necessity for further diversification in verification methods. Future improvements should include the deployment of additional LLMs for cross-verification purposes to mitigate bias and enhance the reliability and generalizability of the findings. These steps are crucial for advancing the robustness and applicability of our methodology in real-world scenarios.

8. Acknowledgments

I would like to thank Darren Ruan for his clear framework instructions and unwavering support throughout this semester. His guidance in suggesting code improvements and sharing valuable advice has been instrumental in the accomplishments of this project.

Reference

[1] Min, S., Lyu, X., Holtzman, A., Artetxe, M., Lewis, M., Hajishirzi, H., & Zettlemoyer, L. (2022). Rethinking the role of demonstrations: What makes in-context learning work? Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing. https://doi.org/10.18653/v1/2022.emnlp-main.759

[2] Wei, J., Wang, X., Schuurmans, D., Bosma, M., Ichter, B., Xia, F., Chi, E., Le, Q., & Zhou, D. (2023, January 10). Chain-of-thought prompting elicits reasoning in large language models. arXiv.org. https://arxiv.org/abs/2201.11903

[3] Chen, X., Wang, Z., & Di, X. (2023, February 10). Sentiment Analysis on multimodal transportation during the COVID-19 using social media data. MDPI. https://www.mdpi.com/2078-2489/14/2/113

[4] Yao, S., Yu, D., Zhao, J., Shafran, I., Griffiths, T. L., Cao, Y., & Narasimhan, K. (2023, December 3). Tree of thoughts: Deliberate problem solving with large language models. arXiv.org. https://arxiv.org/abs/2305.10601