SI 630 Project Proposal

Shaoze Yang
shaozey@umich.edu

1 Project Goal

As electric vehicles (EVs) continue to gain popularity, the importance of charging stations has become increasingly paramount. However, when it comes to selecting the best charging station, individuals often find themselves relying on popular mapping applications that provide only an overall score and numerous user comments. Sifting through these comments to determine the most suitable charging station can be both time-consuming and demanding. To address this challenge, I propose to develop a multi-class classifier based on the user comments associated with charging stations to perform a quantitative analysis. The goal is to identify specific aspects that users may find appealing or unappealing, such as charging speed, cost, and the quality of infrastructure at a particular charging station. Ultimately, this approach will enable individuals to make efficient choices by simply reviewing these categorized insights.

2 NLP Task Definition

This project aims to develop a classifier that maps comments to a predefined set of labels, encapsulating various aspects of charging stations. These aspects include accessibility & availability, amenities & location, compatibility & connectivity, among 16 distinct categories. Additionally, the system is designed to generate a corpus for each label. I may integrate the extracted keywords to a search engine for users to efficiently search for charging stations.

To illustrate how this system operates, consider the following example:

Input: Had trouble connecting initially. I called customer service. They rebooted the machine and it worked just fine after that. It is a paid service. Close to the Papa Murphys Pizza store. Uncovered. Would be accessible at any time day or night. Well lit parking lot. This is one of the better charging stations. Seems to be cheaper sometimes n less crowded, maybe? They're all pretty crowded.

Outputs (Aspect / (Sentiment) / Detected phrases):

1. compatibility and connectivity / Negative / Had trouble connecting initially
2. customer service / Positive / They rebooted the machine and it worked just fine after that
3. payment Options / Positive / It is a paid service
4. amenities and location / Positive / Close to the Papa Murphys Pizza store, walking distance to Target
5. accessibility and availability / Positive / Would be accessible at any time day or night
6. ...

Extracted Keywords: Compatibility issues, Customer service, Payment, Amenities, Location, Accessibility, Lighting, Pricing, Crowding

3 Data

The data comes from reviews about the stations in the former dataset from Google Maps. It would contain fields such as name, business status, and most important, detailed review content. There is an API provided by SerpApi that helps us to grab data. The data would come in the form of text documents. The collection contains 16767 records in total. The basic statistics of the data are illustrated in Figure 1.

The ChatGPT API automatically generates the training and validation sets, and I perform manual filtering in two stages. However, due to API speed constraints, I have managed to generate data for 6,000 charging stations and 38,484 sentences. An example provided by GPT is discussed in Section

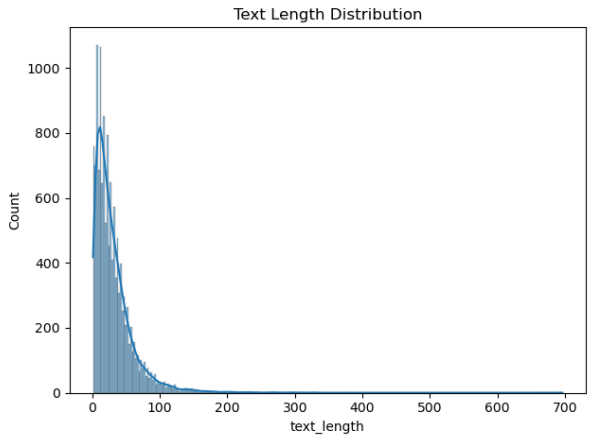

Figure 1: Distribution of Comment Length

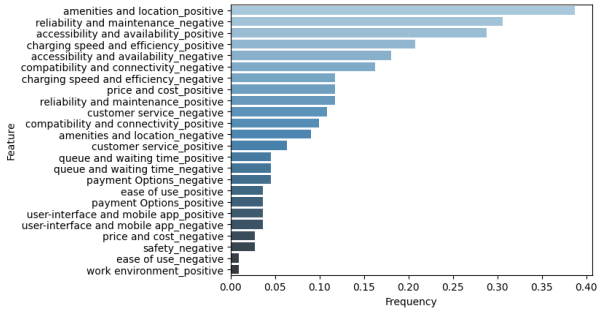

Figure 2: Counts of Extracted Aspects

2. **My goal is to develop a model that is significantly lighter than GPT yet maintains comparable accuracy.** Statistics for the extracted aspects are depicted in Figure 2.

4 Related Work

I am particularly interested in leveraging the attention mechanism, as it aligns with my objectives. Through observation, I've noted that GPT can pinpoint aspects from just a sentence or even a phrase within a lengthy comment. This is a significant advancement over previous models that relied solely on keywords. The attention mechanism, as outlined by Vaswani et al. in their seminal paper [1], allows for more resources to be concentrated on relevant areas to extract detailed information about the target while minimizing the impact of irrelevant data. Furthermore, Yang et al. introduced an attention-based document classifier [2], adding to the body of research supporting this approach. I've also explored various methods, including LSTM-based RNN classifiers [3] and CNN-based techniques. Should time allow, I plan to incorporate these methodologies into my project.

5 Evaluation

To evaluate the performance of my classifier, I would compare the result of the generated labels with the ground truth. Since we have 16 different aspects, then we can describe a comment in a vector in 16 dimensions:

$$C = [A_1, A_2, ..., A_{16}]$$

We can calculate the distance of this vector to the vector of ground truth in L2 norm. Notably, this may also be used as loss function. Moreover, I can generate the confusion matrix of this result to further analyze which 2 or aspects are hard for model to distinguish.

5.1 Baseline Model

I opt to use 2 baselines in this project. Firstly, I can test the accuracy of just choosing the most common label, the amenities and location. Also, since I would probably encode the comment to a vector, then I can also use other basic classifiers in scikit-learn such as decision tree, lightgbm, neural networks to see the corresponding performances.

6 Work Plan

1. Data scratching and Exploratory Data Analysis. (1-2 Weeks, already done)
2. Literature review and methodology development (1-2 weeks)
3. Implementation and analysis (2 weeks)
4. Evaluation and drafting reports (1 week)

References

[1] Ashish Vaswani et al. “Attention is All you Need”. In: *Advances in Neural Information Processing Systems*. Ed. by I. Guyon et al. Vol. 30. Curran Associates, Inc., 2017. URL: https://proceedings.neurips.cc/paper_files/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf.

[2] Zichao Yang et al. “Hierarchical Attention Networks for Document Classification”. In: *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Ed. by Kevin Knight, Ani Nenkova, and Owen Rambow. San Diego, California: Association for Computational Linguistics, June 2016, pp. 1480–1489. DOI: 10.18653/v1/N16-1174. URL: https://aclanthology.org/N16-1174.

[3] Chunting Zhou et al. *A C-LSTM Neural Network for Text Classification*. 2015. arXiv: 1511.08630 [cs.CL].