Aspect-Based Sentiment Classification on Electric Vehicle Users Reviews of Charging Infrastructure

Anonymous ACL-IJCNLP submission

Abstract

With California's mission of fully phasing out the sales of gasoline cars by 2035 (CSG), more and more attention has been drawn to Electric Vehicles. In 2020, electric vehicles made up 2.2% of the U.S auto market (Harder) and this number will continue to grow with more states announcing goals of mass adoption. However, the often negative charging experience of charging infrastructures (Ha et al., 2021) has been a major challenge to boosting the adoption of electric vehicles for the broadest spectrum of drivers. In this study, we use natural language processing to classify the sentiments (positive, negative and neutral) regarding each of the four aspects (functionality, accessibility, cost and service time) (Ha et al., 2021) expressed in the text reviews of charging stations. We annotated 5,000 electric vehicle drivers reviews from the study by Ha et al. (2021) and used neural network models to automatically classify these reviews into positive, negative and neutral experiences. By doing so, we hope to provide more granularity to understanding electric vehicle owners' charging experiences. Our machine learning models can potentially be deployed on charging station tracker and map service applications, showing sentiment polarities by topics based on users' text reviews of each charging station.

1 Related Work

To fight against climate change and mitigate GHG emissions, Biden's administration is planning a new EV era by calling for the construction of 500,000 new public charging stations by the end of 2030 and a future without gas-powered vehicles (Room). Making this great transition is not easy and the current state of EV charging infrastructure is still limited in numerous aspects. Thus, it's essential to understand the difficulties that EV owners face in charging in order to systematically integrate users' demand in EV infrastructure planning. Besides conventional data sources such as EV consumer surveys (McKinsey), we believe text-based reviews and posts on online forums are an underutilized resource for decision makers to understand EV owners' charging experiences. These EV forums are mostly open to the public and have a good potential to capture more open-ended feedback from EV users than conventional surveys. However, going through all the feedback in the forums can be an extremely labor-intensive task. Fortunately, several NLP techniques can be well positioned to analyze users' sentiments and perceptions on specific facets of EV charging.

Past NLP studies on EV consumers' charging experiences largely focus on Topic Classification, a supervised learning method to classify their texts by topics. For example, in Ha et al. (2021), researchers pre-identified and crowd-labeled 8 main topics and associated 32 sub-topics. Although those identified class labels may look comprehensive for now, they are likely to expand quickly as new problems arise with greater EV adoption in the near future. Of course, knowing what EV owners like to talk about charging is just the first step. Identifying their sentiment on their sub-experiences is not less important for policy makers and infrastructure planners to evaluate strengths and weaknesses.

Further along the way, we explored some previous research on aspect-based sentiment analysis on reviews. Ganu et al. (2009) used support vector machine to predict the sentiment expressed in each aspect of the restaurant reviews. With the dataset created by Ganu et al. (2009), the 2014 ACL conference proposed an aspect-based sentiment classification task and over 30 teams participated. Kiritchenko et al. (2014) were able to stand out as the top-performing team because of their novel way of leveraging an additional lexicon to enrich their features.

A study by Marchetto et al. combines topic classification and sentiment analysis using BERT and XLnet. Some other studies focus on understanding user behaviors, including Ha et al. (2021) that uses CNN and RNN to characterize the health of emerging EV charging station infrastructure and offer insights into the behavior of EV drivers; and Asensio et al. (2020) extracted behavioral insights from text reviews generated by EV users in 651 core-based statistical areas in the United States. Asensio et al. (2020) used CNN to classify the general sentiment expressed through EV drivers' review on charging stations, by their type, location and point of interest. Past efforts have also been made on predicting sentiment for short texts, such as twitter messages and online short reviews (Dos Santos and Gatti de Bayser, 2014). Also, Cliche (2017) ensembled CNNs and LSTMs to improve the performance of sentiment analysis task. While binary sentiment analysis tasks have been the mainstream, more and more studies start to focus on multimodal sentiment analysis (Kumar and Vepa, 2020; Tsai et al., 2019; Ghosal et al., 2018). Jiang et al. (2019) addresses the common challenges in existing Aspect-Based Sentiment Analysis datasets, where most sentences contain only one aspect or multiple aspects with the same sentiment polarity.

This study marks a step forward in understanding electric vehicle drivers' charging experience. However, it does not address this sentiment classification problem at a more granular level. For example, in Asensio et al. (2020), the authors only classified each review into one of the two categories: positive and negative. This overall sentiment might not be representative of the sentiment the driver was trying to convey. Consider this review from an electric vehicle driver : *"... chargepoint 2 units were working. No exclusive signs for EV so spots taken by ICEs."* In this review, the driver was complaining about the accessibility of the charging station because his/her spot was taken by gasoline powered cars. Moreover, the driver mentioned the functionality of the charging station as well, stating that the chargepoint units were working properly. So in this case, there are two aspects expressed in this consumer review, each with a different sentiment: functionality: positive, accessibility: negative. It would be hard if we try to give reviews like this an overall sentiment when there are more than one aspect with different sentiments expressed.

More importantly, simply categorizing reviews into positive and negative experiences might not be the optimal way. During our annotation process, we found that a lot of the electric vehicle drivers were using these platforms to share information, in other words, to inform other drivers about their charging experience. For example, in this short review *"5:10. 2 available."*,the driver mentioned the time and availability. However, there is no clear sentiment associated with each aspect. During our discussion we thought it would be more reasonable to categorize reviews into three categories: positive, negative and neutral. By doing so, we can account for the fact that some drivers were doing information sharing without expressing their positive or negative sentiments.

To conclude, in our study, we leveraged the pretrained CNN and LSTM model from Asensio et al. (2020) and deployed them on the topic classification dataset developed by Ha et al (2021). We used the topics generated and classified in the study by Ha et al. (2021) as the ground truth and conducted sentiment classification for four aspects: functionality, accessibility, cost and service time. There are three sentiment labels we are trying to predict: positive, negative and neutral.

2 Data

We leveraged the datasets from Ha et al. (2021) study, and the GitHub Repository (https://github.com/asensio-lab/transformer-EV-topic-classification). It is a publicly available dataset that contains 10652 reviews on electric vehicle charging stations in the US. Each user review was labeled as one (or more) of eight main aspect categories by expert annotators and crowdsourcing. For the scope of our study, we will only focus on aspect-based sentiment classification and will directly utilize those pre-labelled aspects as the ground truth. Also, in the Ha et al. (2021) study, the authors predefined eight aspect categories as Functionality, Range anxiety, Availability, Cost, User interactions, Location, Service time and Dealership.

After thoughtful consideration, we decided to narrow down our scope to only four categories, excluding range anxiety user interactions, dealership and location. The reasons behind are as follows: firstly, these four aspects are not so aligned with the scope for our capstone project. Secondly, reviews containing these three categories are extremely sparse, which means very few electric ve-

Topic	Subtopic Examples
Service time	charging rate
Functionality	general functionality, charger, screen, power level, connector type, card, reader, connection, time, error message, station, mobile application, customer service
Availability	number of stations available, ICE, general congestion
Cost	parking, charging, payment

Table 1: EV mobile app typology of user reviews (Ha et al., 2021)

hicle owners mentioned user interactions, dealership, range anxiety or location in their reviews. Lastly, during our annotation process, we found that most of the sentiment expressed in these three categories(user interactions, dealership and location) are neutral (86%, 84%,73% respectively), meaning that the drivers were merely sharing information on these three categories. For the range anxiety aspect, over 88% of the reviews contain negative sentiment. Looking into these four categories might not add much to our existing knowledge compared to the other aspects such as functionality and availability. Moreover, it might be just sufficient enough to use a majority vote model to predict the sentiment in these four categories.

3 Data Annotation

During our annotation process, We adopted the annotation guidelines from the Ha et al. (2021) study as shown below.

In the first phase of annotation, our team of three independently annotated the sentiment polarities (Positive/Negative/Neutral) with respect to the four extracted aspect terms (Functionality, Availability, Cost and Service Time). Each review is annotated by two different annotators. Specifically, only extracted aspects for each review are annotated with sentiments, and each review can have different sentiments in different aspects. Phase one resembles a pilot study where we calculated our Cohen's Kappa and handpicked the reviews we disagreed on. The purpose of doing this was to discuss and establish a gold standard for later annotation. During phase two, each of the three annotators annotated up to 1,000 reviews independently for each aspect for later training and testing purposes.

Aspect	Cohen's Kappa
functionality	0.68
availability	0.77
cost	0.69
service time	0.71

Table 2: Inter-annotator Agreement

From the annotation process, we have several interesting observations. The first observation is that a great portion of reviews of service time category only contains a simple short mention of electric power or current, without any further expressions or judgement (whether the charging rate is satisfying or not). For example, *"30 amps, 240 volts."* The large presence of such reviews populating the service time aspect makes it difficult to learn any emotion-related expressions. The second observation is the imbalance proportion of different review aspects; functionality seems to take the majority of reviews, and user interaction and dealership aspects are relatively less mentioned. So in our calculation of inter-annotator agreement, we will calculate Cohen's Kappa for each one of four aspects separately to better understand across-aspect differences. Consistent with the findings from Ha et al. (2021), there are a lot of reviews containing the word *"ICE"* or *"ICE'd"*, which is a word specific to the electric vehicle domain. It means the occasion where the spot designated to an electric vehicle is taken by a gasoline powered car. A lot of the electric vehicle drivers use this word to express their frustration and negative charging experience when they found out that their cars were "ICEd".

4 Method and Analysis

We use pre-trained GloVe word embeddings (Jeffrey Pennington) to featurize the words expressed in each review. To be more specific, we adopted the code from Asensio et al. (2020) and made some adjustments and changes based on our project goal. For each word, we take the corresponding 300 dimensional word embedding from GloVe (Jeffrey Pennington) . For words that are not in the GloVe embedding (Jeffrey Pennington) corpus, we randomly initialize the word embeddings and update them during the training process. We chose three window sizes: 3,4,5 respectively, 50 filters for each region size and 1-max pooling. We also used dropout during the training process to prevent overfitting. The output space contains three sentiment

Label	Precision	Recall	Fscore
Negative	80%	65%	0.72
Positive	56%	79%	0.65
Neutral	56%	32%	0.41

Table 3: Functionality, LSTM

Label	Precision	Recall	F Score
Negative	64%	43%	0.52
Positive	65%	57%	0.61
Neutral	67%	29%	0.40

Table 4: Functionality, CNN

labels: positive, negative and neutral.

The structure of this CNN model is illustrated as below:

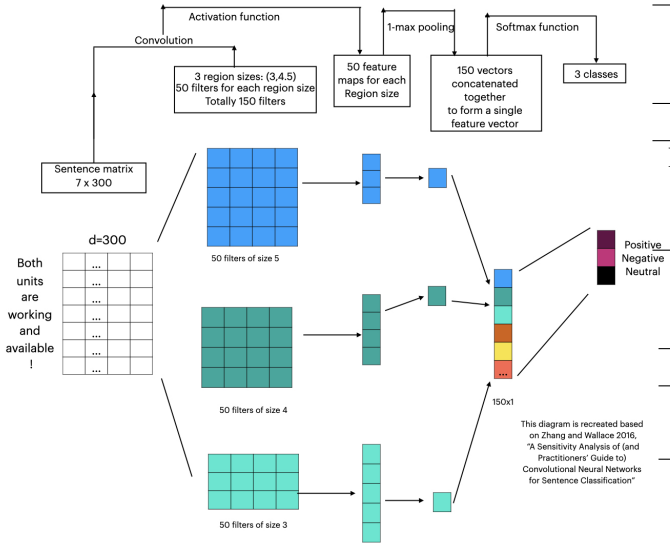

Figure 1: CNN Structure

In terms of evaluation, we calculated the precision and recall score for our CNN and LSTM model output. Since we have three labels(positive, negative and neutral) in our output space, we calculated the precision and recall score metrics for each label.

For the logistic regression model, we used a simple bag-of-words transformation as our model input and compared the model performance to that of the other models. We later used this logistic regression model as our baseline model for comparison.

4.1 Service Time

Generally all three models achieved the highest F1 score (greater than 0.8) in predicting the neutral sentiment and negative sentiment was better

Label	Precision	Recall	F Score
Negative	60%	44%	0.51
Positive	63%	75%	0.68
Neutral	46%	45%	0.45

Table 5: Functionality, Logistic Regression

Label	Precision	Recall	F Score
Negative	83%	61%	0.70
Positive	82%	34%	0.48
Neutral	73%	94%	0.82

Table 6: Service Time, LSTM

Label	Precision	Recall	F Score
Negative	88%	77%	0.82
Positive	87%	70%	0.77
Neutral	87%	90%	0.89

Table 7: Service Time, CNN

Label	Precision	Recall	F Score
Negative	95%	67%	0.79
Positive	80%	56%	0.67
Neutral	81%	98%	0.89

Table 8: Service Time, Logistic Regression

Label	Precision	Recall	Fscore
Negative	77%	74%	0.70
Positive	61%	34%	0.44

Table 9: Availability, LSTM

Label	Precision	Recall	F Score
Negative	74%	52%	0.61
Positive	90%	19%	0.32

Table 10: Availability, CNN

Label	Precision	Recall	F Score
Negative	60%	76%	0.67
Positive	59%	51%	0.55
Neutral	40%	15%	0.22

Table 11: Availability, Logistic Regression

Label	Precision	Recall	F Score
Negative	75%	8%	0.14
Positive	72%	70%	0.71
Neutral	59%	38%	0.46

Table 12: Cost, LSTM

Label	Precision	Recall	F Score
Negative	100%	3%	0.06
Positive	82%	73%	0.77
Neutral	62%	18%	0.28

Table 13: Cost, CNN

Label	Precision	Recall	F Score
Negative	58%	28%	0.38
Positive	70%	94%	0.8
Neutral	62%	44%	0.52

Table 14: Cost, Logistic Regression

predicted than positive sentiment. The reason is neutral reviews take the majority of human annotations. Drivers are more likely to only leave a charging rate as a fact than express any attitude towards the rate in the review (eg. *" charged at 22 mph"*). Though annotators can indeed categorize the reported charger rate to Level1, Level2 or DC fast charger, those review texts are in nature neutral and it's hard to capture any emotional messages. Second, for both positive and negative sentiment, precision was consistently higher than recall, meaning our prediction results are better in quality than in quantity. Since negative sentiment is of great interest to our project and the low recall implied a great many false negatives in negative sentiment prediction, it is worth further investigation if the models mis-predicted both positive/negative as neutral.

4.2 Functionality

Compared with negative and neutral sentiment, positive sentiment is more consistently predicted, with a F1 score around 0.63 across all models. The confusion between negative and neutral was also demonstrated in the manual annotation process. Many reviews would start with a series of complaints about malfunctioning equipment and only ended the review with *"it finally worked after so many hours"*. There was also greater dispute between annotators if such a hardship in charging experience should be considered negative or if we should only focus on whether the driver finally managed to charge. By comparison, positive reviews usually contain phrases typical of expressing positive emotions such as *"it worked amazing"* or *"very successful charging experience!"* and are thus more distinguishable to the model. Neutral sentiment was overall very poorly predicted and LSTM scored the best in negative sentiment prediction (F1=0.72). This might result from LSTM's capability of taking account the sequence of texts, which is especially helpful for predicting long reviews where the driver experience a series charging difficulties and finally got the charger to work.

4.3 Availability

Extremely imbalanced labels is an issue with the availability aspect. Most drivers have a clear sentiment to express regarding the availability chargers, rather than staying neutral about this topic. Only very few of the labels in the availability topic are neutral. That is the reason why our LSTM and CNN model did not output the performance metrics for the neutral label. In general, the precision is higher than recall, meaning that the quality is better than quantity. Also, the models were better at capturing the negative labels than positive ones. It is worth further investigating why the model did a better job on the positive labels.

4.4 Cost

The precision of both LSTM and CNN models is way higher than the recall score for all three labels. In general, the CNN model did the best job in terms of the three metrics(precision, recall and F Score). For the negative labels, it even reached a precision score of 100%, while the recall score for the negative label is extremely low, 3%. This indicates that our model might have outputted a lot of false negatives. A perfect precision rate means there were no false positives. The models were able to achieve a harmony between the precision and recall for the positive sentiments.

5 Limitations

5.1 Label Imbalance

The three sentiment polarities are not balanced in each aspect. For example, For the availability aspect, neutral sentiments only take up less than 19% of the reviews. The majority of reviews are electric vehicle users complaining about their cars being ICEd. Few of the drivers chose to share information on availability issues. Therefore, neutral sentiments in the availability aspect are very rare. This label imbalance issue might hinder the performance of our machine learning models. In future work, we can use oversampling or downsampling technique to address this imbalance issue.

5.2 Annotation

Due to limited time and human resources, we were only able to annotate approximately 1,000 reviews for each aspect. More input data is always the better for model training purposes. Also, since neither of us is an expert in the electric vehicle industry, our annotation might not be the most accurate. Our cohen's kappas for each aspect range from 0.68 0.75, while a cohen's kappa above 0.8 would be ideal. In the future, if we can consult the advice from experts we might be able to improve our model performance. On the other hand, during our annotation process, we found that most of the reviews for charging stations are relatively vague. In other words, the sentiments were not clearly conveyed. Thus, it was hard for human annotators to assign a label to each review regarding each aspect in some cases. Human annotators had to use their best judgement to annotate.

5.3 Data Quality

During our annotation process, we have found that some of the topic classification labels derived from the Ha et al. (2021) study were incorrect. For example, consider this review *"free. Charged for one hour"*, two aspects, cost and service time were mentioned in this context, but they only identified the service time aspect in their data set. Such issues in our source data might cause our model to under-perform as well. In our future research we can further examine the quality of our source data for better performance.

6 Discussion and Future Application

6.1 Cost of Charging

During our annotation process, we noticed some drivers were complaining about the expensive charging rates. Without a universal guideline for charging rate, a lot of the charging rates are set by private owners. The cost of charging is one of the main concerns expressed by some reviews, for example, *"Powered and appears operational... Still more pricey than petrol."*;*"And...this station is OVERPRICED! Come on California - incentivize the use of electric vehicles, if you want a million on the road!"* Although literature suggested EV charging prices should have advantage over gasoline fueling prices (INL; NRDC), the expensive charging rates as expressed in the reviews can become a concern for drivers to opt for electric vehicles. There are two possible reasons for the higher charging costs: 1) Due to the lack of charging stations in some areas, charger operators may overprice their rates for charging, but the price will be automatically corrected after the supply side increases over time. 2) The actual energy cost in dollars per mile is usually lower for EV charging compared to gasoline fueling, but not all consumers are aware of how to calculate the cost difference. There needs to some educational programs for would-be EV users to understand the benefits of lower operational cost so that they wouldn't be discouraged by the seemingly high charging rates.

6.2 Availability

Availability is another pressing issue a lot of drivers face in charging. "ICE'd" is a commonly expressed word in charging station reviews, indicating the electric vehicle charging spot is taken by a gas car. The issue is annoying to EV drivers who need the spot to charge and also impacts revenues and reputation for the stations. It is very important for policy makers, dealerships and other parties involved to come up with innovative ways to develop some mechanism to counteract this situation and improve charger etiquette.

6.3 Functionality

A lot of drivers also complained about the functionality of charging stations. It's very common in the reviews that some chargers are not functional or have not been fixed, or sometimes drivers find their charging connector is not compatible with the available charging ports. The issue of charger compatibility is another roadblock to the nationwide adoption of electric vehicles – for example, non-Tesla cars will need an adapter in order to charge at a Tesla station (Cline). Standardizing charging ports are much needed to ensure higher utilization of charging infrastructure for different EV models entering the market in the years to come.

6.4 Future Application

The models used in this study demonstrate the potential of deploying such natural language processing techniques onto relevant applications for locating charging stations. With the power of machine learning, we can analyze real-time text reviews and generate insights on charging experiences simultaneously. For example, for each specific charging station, aggregated sentiment scores for overall charging experience and aspect-specific charging

experiences can assist users in making informed decisions of which charging stations to use, especially when planning for a road trip or stopping by a place that is not on their regular commute routes. The cross-station comparison with such aspect-based granularity can help charging station operator identify the areas to improve and build reputation in order to be more competitive in the charging station market.

One existing problem for such application is there are too many apps for locating a charging station. (e.g. Greenlots, EVgo, ChargePoint, PlugShare, ChargeHub, Blink and the Chevy app.) *"Because of competing business interests,"* (Schaal) not all charging stations can be found on one integrated app. Drivers often have too download five to six apps beforehand to ensure charging. Integrating all the information and providing better coverage is one of the prerequisites of the deployment of sentiment analysis techniques.

References

Omar Asensio, Kevin Alvarez, Arielle Dror, Emerson Wenzel, Catharina Hollauer, and Sooji Ha. 2020. Real-time data from mobile platforms to evaluate sustainable transportation infrastructure. *Nature Sustainability*, 3.

Mathieu Cliche. 2017. BB_twtr at SemEval-2017 task 4: Twitter sentiment analysis with CNNs and LSTMs. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pages 573–580, Vancouver, Canada. Association for Computational Linguistics.

Amanda Cline. Can non-tesla electric vehicles charge at tesla charging stations.

California State Government CSG. Governor newsom announces california will phase out gasoline-powered cars drastically reduce demand for fossil fuel in california's fight against climate change..

Cicero Dos Santos and Maira Gatti de Bayser. 2014. Deep convolutional neural networks for sentiment analysis of short texts.

Gayatree Ganu, Noemie Elhadad, and Amélie Marian. 2009. Beyond the stars: Improving rating predictions using review text content.

Deepanway Ghosal, Md Shad Akhtar, Dushyant Chauhan, Soujanya Poria, Asif Ekbal, and Pushpak Bhattacharyya. 2018. Contextual inter-modal attention for multi-modal sentiment analysis. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3454–3466, Brussels, Belgium. Association for Computational Linguistics.

Sooji Ha, Daniel J. Marchetto, Sameer Dharur, and Omar I. Asensio. 2021. Topic classification of electric vehicle consumer experiences with transformer-based deep learning. *Patterns*, 2(2):100195.

Amy Harder. To combat climate change, electric cars have to be cheaper.

INL. Comparing energy costs per mile for electric and gasoline-fueled vehicles.

Christopher D. Manning. Jeffrey Pennington, Richard Socher. Glove: Global vectors for word representation.

Qingnan Jiang, Lei Chen, Ruifeng Xu, Xiang Ao, and Min Yang. 2019. A challenge dataset and effective models for aspect-based sentiment analysis. pages 6281–6286.

Svetlana Kiritchenko, Xiaodan Zhu, Colin Cherry, and Saif Mohammad. 2014. Nrc-canada-2014: Detecting aspects and sentiment in customer reviews. pages 437–442.

Ayush Kumar and Jithendra Vepa. 2020. Gated mechanism for attention based multimodal sentiment analysis. *CoRR*, abs/2003.01043.

Daniel J Marchetto, Sooji Ha, Sameer Dharur, and Omar Isaac Asensio. Extracting user behavior at electric vehicle charging stations with transformer deep learning models.

McKinsey. Mckinsey electric vehicle index.

NRDC. Electric vs. gas: Is it cheaper to drive an ev?

Briefing Room. Fact sheet: Biden administration advances electric vehicle charging infrastructure.

Eric Schaal. 5 biggest problems with electric vehicle charging.

Yao Hung Hubert Tsai, Shaojie Bai, Paul Pu Liang, J. Zico Kolter, Louis-Philippe Morency, and Ruslan Salakhutdinov. 2019. Multimodal transformer for unaligned multimodal language sequences. *CoRR*, abs/1906.00295.