# Extracting User Behavior at Electric Vehicle Charging Stations with Transformer Deep Learning Models

**Daniel J. Marchetto [1], Sooji Ha[2,3], Sameer Dharur[4], Omar Isaac Asensio[1,5*]**
[1]School of Public Policy, Georgia Institute of Technology, [2]School of Civil and Environmental Engineering, Georgia Institute of Technology, [3]School of Computational Science and Engineering, Georgia Institute of Technology, [4]School of Computer Science, Georgia Institute of Technology, [5]Institute for Data Engineering and Science, Georgia Institute of Technology

***Abstract***

*Mobile applications have become widely popular for their ability to access real-time information. In electric vehicle (EV) mobility, these applications are used by drivers to locate charging stations in public spaces, pay for charging transactions, and engage with other users. This activity generates a rich source of data about charging infrastructure and behavior. However, an increasing share of this data is stored as unstructured text—inhibiting our ability to interpret behavior in real-time. In this article, we implement recent transformer-based deep learning algorithms, BERT and XLnet, that have been tailored to automatically classify short user reviews about EV charging experiences. We achieve classification results with a mean accuracy of over 91% and a mean F1 score of over 0.81 allowing for more precise detection of topic categories, even in the presence of highly imbalanced data. Using these classification algorithms as a pre-processing step, we analyze a U.S. national dataset with econometric methods to discover the dominant topics of discourse in charging infrastructure. After adjusting for station characteristics and other factors, we find that the functionality of a charging station is the dominant topic among EV drivers and is more likely to be discussed at points-of-interest with negative user experiences.*

***Keywords:*** *Electric Vehicles; Mobility; Mobile Data; Natural Language Processing; Transformer Models*

## 1. Introduction

The transportation sector is undergoing rapid transformation such as vehicle electrification and increased usage of mobile apps. These two developments offer the possibility to do real-time monitoring of large-scale infrastructure with streaming data. Electric vehicles have also become a dominant strategy to reduce emissions that includes health co-benefits from displacing internal combustion engines (Carley et al., 2013; Sheldon et al., 2017). The growth of the EV market has brought an increase in complementary digital infrastructure—including charging stations and locator apps intended for use in public spaces. This digital infrastructure has created an ecosystem for users to engage with each other and share information about their EV experiences. The resulting user-generated data can be useful for policy analysis and real-time infrastructure management; however, large portions of this data is in the form of unstructured text, which require computational methods to extract insights (Asensio et al., 2020; Kühl et al., 2019).

In this article, we deploy recent advances in neural-net-based classification algorithms in order to learn the dominant topics of discourse within the EV community. This task of natural language processing (NLP) has been challenging because, although neural-net-algorithms have been shown to perform well in sentiment classification tasks, there are still computational issues related to underdetection particularly with highly imbalanced data (Asensio et al., 2020; Ha et al., 2020). Our approach here is to implement transformer based deep neural networks such as Bidirectional Encoder Representations from Text (BERT) and Transformer-XL (XLnet), which have both yielded promising results in a number of NLP tasks (LeCun & Hinton, 2015; Vaswani et al, 2017; Devlin et al., 2018; Yang et al., 2019). In order to understand user behavior in this domain, we begin with predefined behavioral topics as identified in Ha et al. (2020) and build multi-label classification models that assign one or more relevant topic labels to a given user text as demonstrated in Dharur et al. (2020). Using the output of these supervised classification architectures to find the dominant topics of discussion, we then implement econometric techniques to adjust the algorithmic predictions for observable station characteristics and other factors. We analyzed the multi-labeled topic classifications by points of interest (POI) to find evidence of charging station quality. We comment on implications for the use of transformer deep learning models in policy analysis and infrastructure management.

## 2. Data and Methodology

We have a nationally representative sample of unstructured consumer reviews at 12,720 U.S. charging station locations as provided by a popular EV charging station locator app. The text data consists of 127,257 reviews written in English from 29,532 registered and unregistered EV drivers during the period from 2011 to 2015, which are the early growth years of the EV infrastructure. These represent charging station usage for the entire U.S. market during the period of study. In the sample, contemporaneous information from the mobile app was used to geocode the data with checks against Google Places API data. Example POI categories

include Parking Garage/Lot, Government, Healthcare, Hotel/Lodging, Restaurant, School/University, Store/Retail, Workplace, etc. during the given year a review was made. For descriptive statistics of the review data by station, user and year, see Table 1.

**Table 1. Review Count Descriptive Statistics**

| Reviews per | Mean | SD | Min | Max |
|---|---|---|---|---|
| Station | 9.26 | 21.22 | 1 | 578 |
| User | 5.17 | 15.96 | 1 | 728 |
| Year (2011 – 2015) | 28,034 | 22,024 | 1,331 | 50,217 |

### *2.1. Data Collection*

As in many supervised machine learning tasks, we built a pre-labeled training set using the typology identified in Ha et al. (2020). There are 8 topics—Functionality, Range Anxiety, Availability, Cost, User Interaction, Location, Service Time, and Dealership—selected for human annotation of the review text. Reviews outside of these topics were labeled as Other. We recruited 5 annotators and provided a series of guidelines including a codebook with label definitions, examples from actual reviews for each topic, followed by a 1-hour guided training using a web application developed for annotation (Ha & Marchetto, 2020). After annotator training, inter-rater agreement (Fleiss' kappa) on a holdout sample for all topics ranged from 0.30 to 0.72. This calculation indicated substantial agreement for Service Time (0.72), Availability (0.61), Cost (0.65); moderate agreement for Range Anxiety (0.56), Functionality (0.52), Dealership (0.51), Location (0.45); and fair agreement for User Interaction (0.30). A total of 10,133 randomly selected, unique reviews were labeled by the 5 trained annotators. This selection is intended to be representative of the full dataset and includes reviews across all 8 main topics. Table 2 shows the counts and percentages of each labeled topic in the training data. The most frequently selected labels were Functionality, Location and Availability. We see the highest imbalance in Range Anxiety and Service Time, which were the least frequently selected labeled topics. Other represented only 1.1% of the training data.

**Table 2. Counts and Percentages of Labeled Topics in Training Data**

| Functionality | Location | Availability | Dealership | Cost | Service Time | Range Anxiety | Other |
|---|---|---|---|---|---|---|---|
| 5,399 | 3,377 | 2,197 | 1,391 | 1,072 | 982 | 513 | 116 |
| 52.7% | 33.0% | 21.5% | 13.6% | 10.5% | 9.6% | 5.0% | 1.1% |

### *2.2. Classification through BERT and XLnet*

Our classification task is to assign a predefined topic labels to a given text sequence. The advent of transformer-based deep neural network models has set new benchmarks on a wide variety of NLP tasks such as sentiment analysis, topic classification, question answering, machine translation, among others (Vaswani et al., 2017). A key reason for this algorithmic advancement was the use of the attention mechanism (Lin et al., 2017; Yang et al., 2016). This mechanism is a novel architecture that draws global dependencies between input and

output and eliminates the need for other recurrent and convolution mechanisms. For more detailed expositions of BERT, XLnet and transformer models, see Vaswani et al. (2017), Devlin et al. (2018) and Yang et al. (2019). For the implementations described here, we followed the replication protocols outlined in Trivedi et al. (2019) and Dharur et al. (2020).

### *2.3. Fractional Response Models*

Many observable station, location, and time factors can impact predictions of the topic classifications. Given possibilities for algorithmic bias from historical training data, we would also like to statistically adjust the algorithmic predictions to account for variations in use by POI, networks, and connector technologies available. Our unit of analysis is at the station level. The dependent variable $0 \leq y_{i,j,t} \leq 1$ is a standardized fractional response outcome (Equation 1) where a measure near 0 indicates a low incidence of a particular topic at that station location, while a score near 1 indicates a high incidence of a topic at that station location. This allows us to adjust our dependent variable for the usage frequency at a given station location. Given that Functionality is the dominant label in the training set, we focus our econometric analysis on that label. From Ha et al. (2020), Functionality refers to comments describing whether particular features or services are working properly at a charging station. Because we have a bounded outcome variable, we implement fractional response models (FRM) that use a quasi-maximum likelihood estimator (QMLE) to generate estimates of the likelihood of predicting a topic conditional on observable station characteristics. For additional details about FRM models, see Papke and Woolridge (1996) and Ramalho et al. (2011). The standardized topic score is,

$$y_{i,j,t} = Standardized\ Topic\ Score = \frac{Count\ of\ Topic\ Reviews_{i,j,t}}{Total\ Count\ of\ Reviews_{i,j,t}} \tag{1}$$

where *i* is a given review at *j* station location, in *t* year. Our main model specification is:

$$\begin{aligned} E(y_{i,j,t}|\boldsymbol{X_{i,j,t}}) \quad &= G(\boldsymbol{X_{i,j,t}\beta}) \\ &= G(\beta_0 + POI\ Dummies_{i,j}\boldsymbol{\beta_1} + Station\ Characteristics_{i,j}\boldsymbol{\beta_2} \\ &+ Negativity\ Score_{i,j,t}\boldsymbol{\beta_3} + Interaction\ Effects_{i,j,t}\boldsymbol{\beta_4} \\ &+ Year\ Fixed\ Effects_i\boldsymbol{\beta_5}) \end{aligned} \tag{2}$$

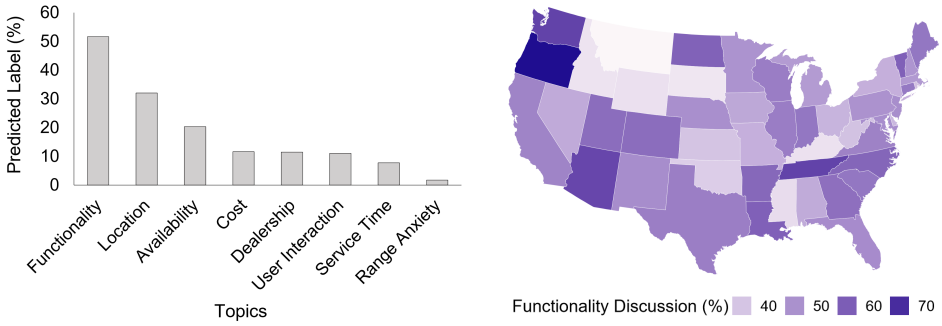
where $\boldsymbol{X_{i,j,t}}$ is a vector of 1 by *k*, of explanatory variables, *E* is the expected value, and $G(\cdot)$ is a unit-bound, nonlinear transform function of a distribution defined as $\frac{e^{x_i\beta}}{1+e^{x_i\beta}}$, which is the logit function. The control variables, given in Equation (2) include POIs, Station Characteristics that include Number of Networks (e.g. 1, 2, 3+) and Number of Connector Types (e.g. 1, 2, 3) at a given station location. To mitigate the possibility for unobserved confounding variables, we also include a proprietary Station Quality Rating provided by the platform provider that ranges between 1 and 5, where 5 indicates a high-quality station location. The Negativity Score is a standardized measure of negative sentiment derived from Asensio et al. (2020). It is used to test the conjecture that negative experiences at different POIs differentially affect the likelihood that a review will be classified as Functionality through interaction effects. To calculate the partial effects and to assist with interpretation of

the coefficients, the average effect on $y$ of a unit change in $X_{i,j,t}$ estimated at the conditional mean is given by $\frac{\partial E(y_{i,j,t}|X_{i,j,t})}{\partial X_{i,j,t}} = \beta_i g(X_{i,j,t}\beta_i)$, where $g(X_{i,j,t}\beta_i)$ is $G(X_{i,j,t}\boldsymbol{\beta})[1 - G(X_{i,j,t}\boldsymbol{\beta})]$, estimated by the QMLE.

## 3. Results and Discussion

### *3.1. Classification*

Across all topics of interest, we find that Functionality is the dominant topic among EV users (Figure 1). This is surprising because issues such as the cost of charging and range have typically received the most attention in public discourse on EV use. From Figure 1a, we see that Cost and Range Anxiety are not dominant topics of discussion. Surprisingly, in Figure 1b, we see that majority of the continental U.S. states discuss Functionality in over 50% of the reviews. Next, we report the classification results for accuracy (measured as partial accuracy) and F1 score (harmonious average of precision and recall). Table 3 contains the accuracy and F1 score of the overall multi-label topic classification. Accounting for uncertainty across 25 runs, we report a mean accuracy of 91.30% on BERT, 91.29% on XLnet, and a mean F1 score of 0.82 on both models. These results provide evidence that the use of these transformer-based models helped overcome the technical challenges of learning from imbalanced data (Table 2).




(a) (b)

*Figure 1. (a) Frequency distribution of predicted topics across the entire dataset. (b) Percent of reviews discussing Functionality at the state-level*

Next, we evaluate the classification results using BERT and XLnet on each of the 8 topics of interest. In Figure 2, we report the accuracies achieved for each topic after 15 model replications. We compared this with a majority classifier, which predicts the most commonly occurring label by a simple majority. Improvements in accuracy over the majority classifier can be interpreted as a measure of the classifier's learning ability. From Figure 2, we see impressive accuracy improvement in the Functionality topic (23.3−23.7 percentage points), followed by Location (18.2−18.6), Availability (12.8−13.1), Cost (8.3−8.4), Dealership (6.2−7.6), Service Time (7.0−7.3) and user Interaction (3.9−4.1) topics. This result overcomes a common criticism of many machine learning algorithms with imbalanced data.

**Table 3. Transformer model cross-validation results.**

| Architecture | Mean Accuracy % (s.d.) | Mean F1 Score (s.d) |
|---|---|---|
| BERT | 91.30 (0.23) | 0.82 (0.0071) |
| XLnet | 91.29 (0.22) | 0.82 (0.0046) |

We find that the Range Anxiety topic gives the lowest accuracy improvement versus a majority classifier (0.4-0.6 percentage points). This could be due to this label being the least selected topic in our dataset, which suggests further scope for improvement to increase the size of the training data, or further tuning of the hyperparameters.

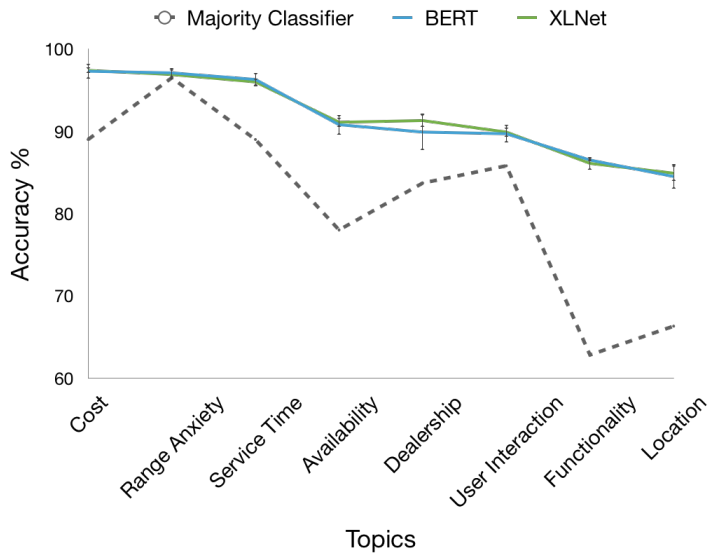



*Figure 2. Topic-level accuracy comparisons with 95% C.I*

### *3.2. Fractional Response Models*

In this section, we present the results for the fractional response models, which statistically adjust for station location and timing factors on the likelihood of selecting Functionality as the predicted label. Unlike many of other applications of machine learning, we argue that this statistical adjustment is needed to mitigate observational biases prevalent in training data. From Model 1 in Table 4, we find evidence of significant heterogeneity on the likelihood of discussing Functionality in public spaces. For example, compared with Street Parking and Parking Lots as the baseline counterfactual, consumers using stations located at POIs such as Shopping, Gas Stations, Supermarkets, Restaurants, and Hotels, are more likely to discuss the functionality of stations. These reviews typically discuss subtopics related to issues such as chargers, screens, connector types, connection, time, error messages, and customer service. However, POI locations such as Government, Healthcare, and Transit Stations were not statistically different from Street Parking and Parking Lots. In order to further understand if these discussions were related to negative user experiences, we evaluated the

subpopulation of reviews by utilizing an algorithmically-generated sentiment score interacted with the POIs. In this analysis, we find that our most important and prominent POI, Shopping, was 7.4% more likely to discuss Functionality issues as compared to our reference case (Model 3 in Table 3). Similarly, reviews at Hotels, although not significantly different from Street Parking and Parking Lots, were 6.4% more likely to discuss Functionality in the presence of negative sentiment. These results from consumer data could suggest that charging station operators at many public spaces or POIs may not have sufficient incentives to ensure the proper maintenance and upkeep of publicly accessible EV infrastructure. Future work could further evaluate mechanisms of dissatisfaction.

**Table 4. Partial Effect Results from FRMs of Standardized Functionality Score.**

| | Model 1 | | Model 2 | | Model 3 | |
|---|---|---|---|---|---|---|
| | Coef | SE | Coef | SE | Coef | SE |
| **POIs** | | | | | | |
| Shopping | 0.076** | 0.022 | 0.048** | 0.018 | 0.013 | 0.023 |
| Car Dealership | 0.060** | 0.019 | 0.021 | 0.016 | 0.021 | 0.016 |
| Workplace | 0.081** | 0.027 | 0.038 | 0.020 | 0.03 | 0.020 |
| Gas Station | 0.165** | 0.021 | 0.167** | 0.017 | 0.154** | 0.019 |
| Government | 0.028 | 0.025 | 0.036 | 0.023 | 0.036 | 0.023 |
| Supermarket | 0.196** | 0.028 | 0.149** | 0.023 | 0.115** | 0.037 |
| Hotel | -0.006 | 0.022 | 0.000 | 0.019 | -0.023 | 0.019 |
| Restaurant | 0.077** | 0.028 | 0.084** | 0.022 | 0.056* | 0.027 |
| Education | 0.066* | 0.027 | 0.049* | 0.020 | 0.047* | 0.022 |
| Transit Station | 0.035 | 0.034 | 0.038 | 0.027 | 0.038 | 0.027 |
| Healthcare | -0.073 | 0.043 | -0.001 | 0.028 | -0.001 | 0.028 |
| Entertainment | -0.031 | 0.034 | -0.016 | 0.034 | -0.015 | 0.034 |
| Airport | 0.040** | 0.033 | -0.114* | 0.040 | -0.129 | 0.077 |
| Library | -0.007* | 0.034 | 0.076* | 0.034 | 0.077* | 0.034 |
| Residential | -0.023 | 0.034 | 0.098 | 0.067 | 0.099 | 0.067 |
| Quality Rating | | | -0.018** | 0.003 | -0.018** | 0.003 |
| Negativity Score | | | 0.130** | 0.009 | 0.107** | 0.011 |
| **Interactions of Negativity with:** | | | | | | |
| Shopping | | | | | 0.074* | 0.038 |
| Gas Station | | | | | 0.029 | 0.027 |
| Supermarket | | | | | 0.068 | 0.057 |
| Hotel | | | | | 0.064* | 0.040 |
| Restaurant | | | | | 0.033 | 0.119 |
| Education | | | | | 0.058 | 0.028 |
| Airport | | | | | 0.006 | 0.033 |
| Station Characteristics | Yes | | Yes | | Yes | |
| Year FE | Yes | | Yes | | Yes | |
| Clustered SE | Yes | | Yes | | Yes | |
| No. Observations | 127,257 | | 127,257 | | 127,257 | |

*p < 0.05, **p < 0.01, ***p < 0.001

## 4. Conclusion

In this article, we have demonstrated the use of neural-net-based classification algorithms to automatically discover topics of EV discourse among members of the EV community. We

provide evidence that transformer-based models overcome prior challenges of training models with highly imbalanced data. In the context of EV reviews, we then use these classification results to identify major issues that users experience in public charging infrastructure. We find that Functionality is the dominant topic of discussion with significant heterogeneity by POIs. This is counter to the public discourse that focuses on cost and range anxiety as dominant themes. This research also provides a proof-of-concept for large-scale practical implementation that can enable real-time processing of mobility behavior patterns.

## Acknowledgments

We gratefully acknowledge funding by the National Science Foundation (NSF Award No. 1931980 and Award No. 1945332) and Microsoft Azure. We also thank IDEaS, and the PACE high performance computing team at Georgia Tech.

## References

Asensio, O. I., Alvarez, K., Dror, A., Wenzel, E., Hollauer, C., & Ha, S. (2020). Real-time data from mobile platforms to evaluate sustainable transportation infrastructure. *Nature Sustainability.* DOI 10.1038/s41893-020-0533-6

Carley, S., Krause, R. M., Lane, B. W., & Graham, J. D. (2013). Intent to purchase a plug-in electric vehicle: A survey of early impressions in large US cities. *Transportation Research Part D: Transport and Environment*, 18, 39-45.

Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2019, June). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1* (pp. 4171-4186).

Dharur, S., Ha, S., Marchetto, D. J., & Asensio, O. I. (2020) Topic classification of electric vehicle consumer experiences with transformer-based deep learning. Working paper.

Ha, S., Marchetto, D. J., Burke, M. E., & Asensio, O. I. (2020) Detecting behavioral failures in emerging electric vehicle infrastructure using supervised text classification algorithms. In *Proceedings of the Transportation Research Board Annual Meeting.* 20-03461.

Ha, S. & Marchetto, D. J. (2020). Codebook, Available online at https://github.com/asensio-lab/transformer-EV-topic-classification/tree/master/training-manual.

Kühl, N., Goutier, M., Ensslen, A., & Jochem, P. (2019). Literature vs. Twitter: Empirical insights on customer needs in e-mobility. *Journal of Cleaner Production*, 213, 508-520.

LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *Nature*, 521(7553), 436-444.

Lin, Z., Feng, M., Santos, C. N. D., Yu, M., Xiang, B., Zhou, B., & Bengio, Y. (2017). A structured self-attentive sentence embedding. arXiv preprint arXiv:1703.03130.

Papke, L. E., & Wooldridge, J. M. (1996). Econometric methods for fractional response variables with an application to 401 (k) plan participation rates. *Journal of Applied Econometrics*, 11(6), 619-632.

Ramalho, E. A., Ramalho, J. J., & Murteira, J. M. (2011). Alternative estimating and testing empirical strategies for fractional regression models. *Journal of Economic Surveys*, 25(1), 19-68.

Sheldon, T. L., DeShazo, J. R., & Carson, R. T. (2017). Electric and plug-in hybrid vehicle demand: lessons for an emerging market. *Economic Inquiry*, 55(2), 695-713.

Trivedi, K. (2019). Fast-Bert, Accessed online on 02/21/2020 at https://github.com/ kaushaltrivedi/fast-bert.

Yang, Z., Yang, D., Dyer, C., He, X., Smola, A., & Hovy, E. (2016, June). Hierarchical attention networks for document classification. In *Proceedings of the 2016 Association for Computational Linguistics: Human Language Technologies conference* (pp. 1480-1489).

Yang, Z., Dai, Z., Yang, Y., Carbonell, J., Salakhutdinov, R. R., & Le, Q. V. (2019). Xlnet: Generalized autoregressive pretraining for language understanding. *In Advances in Neural Information Processing Systems* (pp. 5754-5764).