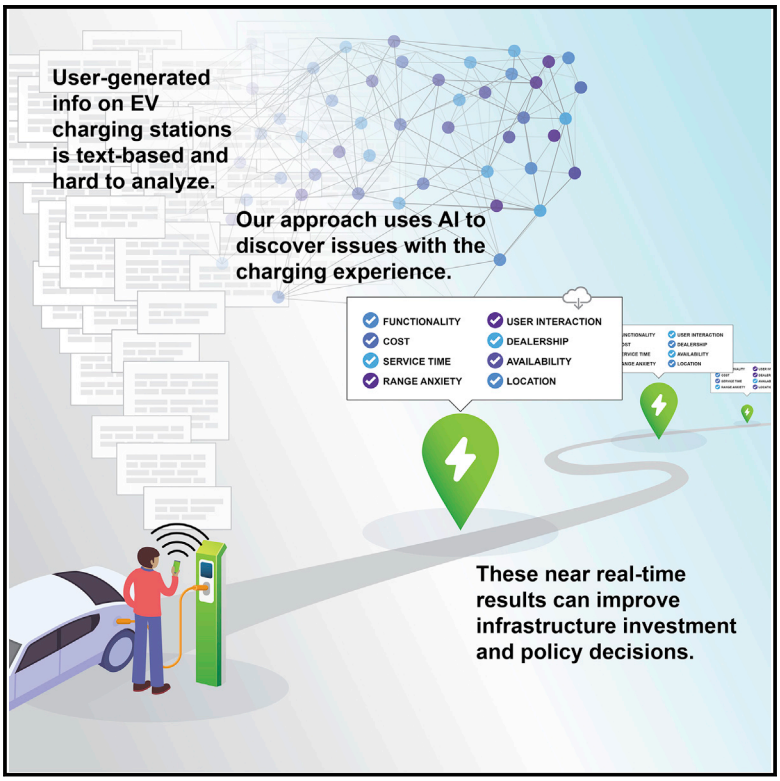
Article

Patterns

# Topic classification of electric vehicle consumer experiences with transformer-based deep learning

## Graphical Abstract




## Highlights

- Consumer data on EV charging behavior are unstructured and remain largely dormant
- We provide proof of concept for automated topic classification with transformer models
- We achieve 91% accuracy (F1 0.83), outperforming previously leading algorithms
- Applications for local and regional policy analysis of EV behavior are described

## Authors

Sooji Ha, Daniel J. Marchetto, Sameer Dharur, Omar I. Asensio

## Correspondence

asensio@gatech.edu

## In Brief

Government analysts and policy makers have failed to fully utilize consumer behavior data in decisions related to EV charging infrastructure. This is because a large share of EV data is unstructured text, which presents challenges for data discovery. In this article, we deploy advances in transformer-based deep learning to discover issues in a nationally representative sample of EV user reviews. We describe applications for public policy analysis and find evidence that less populated areas could be underserved in station availability.

Ha et al., 2021, Patterns 2, 100195
February 12, 2021 
https://doi.org/10.1016/j.patter.2020.100195

CellPress

Patterns

CellPress
OPEN ACCESS

Article

# Topic classification of electric vehicle consumer experiences with transformer-based deep learning

Sooji Ha,[1,2] Daniel J. Marchetto,[3] Sameer Dharur,[4] and Omar I. Asensio[3,5,6,*]
[1]School of Civil and Environmental Engineering, Georgia Institute of Technology, Atlanta, GA 30332, USA
[2]School of Computational Science and Engineering, Georgia Institute of Technology, Atlanta, GA 30308, USA
[3]School of Public Policy, Georgia Institute of Technology, Atlanta, GA 30332, USA
[4]School of Computer Science, Georgia Institute of Technology, Atlanta, GA 30332, USA
[5]Institute for Data Engineering and Science (IDEaS), Georgia Institute of Technology, Atlanta, GA 30308, USA
[6]Lead contact
*Correspondence: asensio@gatech.edu
https://doi.org/10.1016/j.patter.2020.100195

**THE BIGGER PICTURE** Transformer neural networks have emerged as the preeminent models for natural language processing, seeing production-level use with Google search and translation algorithms. These models have had a major impact on context learning from text in many fields, e.g., health care, finance, manufacturing; however, there have been no empirical advances to date in electric mobility.

Given the digital transformations in energy and transportation, there are growing opportunities for real-time analysis of critical energy infrastructure. A large, untapped source of EV mobility data is unstructured text generated by mobile app users reviewing charging stations. Using transformer-based deep learning, we present multilabel classification of charging station reviews with performance exceeding human experts in some cases. This paves the way for automatic discovery and real-time tracking of EV user experiences, which can inform local and regional policies to address climate change.


**Development/Pre-production:** Data science output has been rolled out/validated across multiple domains/problems

## SUMMARY

The transportation sector is a major contributor to greenhouse gas (GHG) emissions and is a driver of adverse health effects globally. Increasingly, government policies have promoted the adoption of electric vehicles (EVs) as a solution to mitigate GHG emissions. However, government analysts have failed to fully utilize consumer data in decisions related to charging infrastructure. This is because a large share of EV data is unstructured text, which presents challenges for data discovery. In this article, we deploy advances in transformer-based deep learning to discover topics of attention in a nationally representative sample of user reviews. We report classification accuracies greater than 91% (F1 scores of 0.83), outperforming previously leading algorithms in this domain. We describe applications of these deep learning models for public policy analysis and large-scale implementation. This capability can boost intelligence for the EV charging market, which is expected to grow to US$27.6 billion by 2027.

## INTRODUCTION

In recent years, there has been a growing emphasis on vehicle electrification as a means to mitigate the effects of greenhouse gas emissions[1] and related health impacts from the transportation sector.[2] For example, typical calculations suggest that electric vehicles (EVs) reduce emissions from 244 to 98 g/km, and this number could further decrease to 10 g/km with renewable energy integration.[3] The environmental benefits range by fuel type, with reported carbon intensities of 8,887 g $CO_2$ per gallon of gasoline and 10,180 g $CO_2$ per gallon of diesel.[4] Government-driven incentives for switching to EVs, including utility rebates, tax credits, exemptions, and other policies, have been rolled out in many US states.[5–7] In this effort, public charging infrastructure remains a critical complementary asset for consumers in building range confidence for trip planning and in EV purchase decisions.[8–10] Prior behavioral research has shown that policies designed to enhance EV adoption have largely focused on

increasing the quantity of cars and connected infrastructure as opposed to the quality of the charging experience.[11] However, a fundamental challenge to deploying large-scale EV infrastructure is regular assessments of quality.

Private digital platforms such as mobility apps for locating charging stations and other services have become increasingly popular. Reports by third-party platform owners suggest there are already over 3 million user reviews of EV charging stations in the public domain.[12–15] In this paper, we evaluate whether transformer-based deep learning models can automatically discover experiences about EV charging behavior from unstructured data and whether supervised deep learning models perform better than human benchmarks, particularly in complex technology areas. Because mobile apps facilitate exchanges of user texts on the platform, multiple topics of discussion exist in EV charging reviews. For example, a review states: "Fast charger working fine. Don't mind the $7 to charge, do mind the over-the-phone 10 min credit card transaction." A multilabel classification algorithm may be able to discover that the station is functional, that a user reports an acceptable cost, and that a user reports issues with customer service. Consequently, text classification algorithms that can automatically perform multilabel classification are needed to interpret the data.

Being able to do multilabel classification on these reviews is important for three principal reasons. First, these algorithms can enable analysis of massive digital data. This is important because behavioral evidence about charging experiences has primarily been inferred through data from government surveys or simulations. These survey-based approaches have major limitations, as they are often slow and costly to collect, are limited to regional sampling, and are often subject to self-report or recency bias. Second, multilabel algorithms with digital data can characterize phenomena across different EV networks and regions. Some industry analysts have criticized EV mobility data for poor network interoperability, which prevents data from easily being accessed, shared, and collected.[16] This type of multilabeled output is also important for application programming interface (API) standardization across the industry, such as with emerging but not yet widely accepted technology standards, including the Open Charge Point Protocol[17] that would help with real-time data sharing across regions. Third, this capability may be critical for standardizing software and mobile app development in future stages of data science maturity (see https://www.cell.com/patterns/dsml) to detect behavioral failures in near real-time from user-generated data.

Modern computational algorithms from natural language processing (NLP) could uniquely address the need for fast, real-time consumer intelligence related to electric mobility, but these algorithms need to be appropriately tailored to domains to be useful. Large-scale analysis of unstructured EV user data remains difficult to carry out, especially when there are multiple topics discussed in each review and the datasets are imbalanced. Imbalanced data create challenges for models to learn important but less frequently occurring labels and often lead to algorithmic bias. In this paper, we demonstrate the use of deep neural networks to automatically discover insights for topic analysis. We use supervised learning to overcome prior challenges with unsupervised methods that could produce clusters with very little theoretical or social meaning. We provide a proof of concept for the complex task of multilabel topic classification in this domain, which builds on an earlier demonstration of binary sentiment classification with NLP.[11] We apply transformer neural networks, a recent class of pre-trained contextual language models, to accurately detect long-tail discussion topics with imbalanced data, a capability that has been elusive with prior approaches.

Prior research demonstrated the efficacy of convolutional neural networks (CNNs)[18–21] and long short-term memory (LSTM), a commonly used variant of recurrent neural networks[21,22] for NLP. These models have been recently applied to sentiment classification and single-label topic classification tasks in this domain. As a result, the use of NLP has increased our understanding of potential EV charging infrastructure issues, such as the prevalence of negative consumer experiences in urban locations compared with non-urban locations.[11,23,24] Although these models showed promise for binary classification of short texts, generalizing these models to reliably identify multiple discussion topics automatically from text presents researchers with an unsolved challenge of underdetection, particularly in corpora with wide-ranging topics and imbalances in the training data. Prior research using sentiment analysis indicates negative user experiences in EV charging station reviews, but it has not been able to extract the specific causes.[11] As a result, multilabel topic classification is needed to understand behavioral foundations of user interactions in electric mobility.

In this paper, we achieve state-of-the-art multilabel topic classification in this domain using the transformer-based[25] deep neural networks BERT, which stands for bidirectional encoder representations,[26] and XLNet, which integrates ideas from Transformer-XL[27] architectures. We benchmark the performance of these transformer models against classification results obtained from adapted CNNs and LSTMs. We also evaluate the potential for super-human performance of the classifiers by comparing human benchmarks from crowd-annotated training data with expert-annotated training data and transformer models. The extent of this improvement could significantly accelerate automated research evaluation using large-scale consumer data for performance assessment and regional policy analysis. We discuss implications for scalable deployment, real-time detection of failures, and management of infrastructure in sustainable transportation systems.

## RESULTS AND DISCUSSION

### Discovering topics

Charging station reviews can be considered asynchronous social interactions within a community of EV drivers. To characterize user experiences, we introduce 8 main topics and 32 sub-topics that make up a typology of charging behavior. This typology allows for easier identification of behavioral issues with the charging process (Table 1). The definitions we use for supervised learning are as follows: Functionality refers to comments describing whether particular features or services are working properly at a charging station. Range anxiety refers to comments regarding EV drivers' fear of running out of fuel mid-trip and comments concerning tactics to avoid running out of fuel. Availability refers to comments concerning whether charging stations are available for use at a given location. Cost refers to comments about the amount of money required to park and/or charge at particular locations. User interaction refers to comments in

**Table 1. EV mobile app typology of user reviews**

| Topic | Sub-topic examples |
|---|---|
| Functionality | general functionality, charger, screen, power level, connector type, card, reader, connection, time, error message, station, mobile application, customer service |
| Range anxiety | trip, range, location accessibility |
| Availability | number of stations available, ICE, general congestion |
| Cost | parking, charging, payment |
| User interactions | charger etiquette, anticipated time available, user tips |
| Location | general location, directions, staff, amenities, points of interest, user activity, signage |
| Service time | charging rate |
| Dealership | dealership charging experience, competing brand quality, relationship with dealers |
| Other | general experiences |

ICE refers to situations where a charging station is blocked by an internal combustion engine vehicle.

**Table 2. Overall model performance**

| | Accuracy % (SD) | F1 score (SD) |
|---|---|---|
| BERT | 91.6 (0.13) | 0.83 (0.0037) |
| XLNet | 91.6 (0.07) | 0.84 (0.0015) |
| Majority classifier | 81.1 (0.00) | 0.45 (0.0000) |
| LSTM | 90.3 (0.17) | 0.80 (0.0036) |
| CNN | 90.9 (0.12) | 0.81 (0.0032) |

Models were trained and tested on expert annotated data.

which users are directly interacting with other EV drivers in the community. Location refers to comments about various features or amenities specific to a charging station location. The Service time topic refers to comments reporting charging rates (e.g., 10 miles of range per hour charged) experienced in a charging session. The Dealership topic refers to comments concerning specific dealerships and user's associated charging experiences. Reviews that do not fall into the previous eight topics refer to the Other topic, and are relatively rare. For more information on the robustness of typology, see Supplemental experimental procedures and Tables S5–S7 in the Supplemental information.

In preliminary experiments, we investigated several unsupervised topic modeling techniques that did not provide theoretically meaningful clusters. By contrast, our empirically driven typology is ideally suited to hypothesis testing, spatial analysis, benchmarking with other corpora in this domain, and real-time tracking of station failures, all of which are not identifiable with current information systems. For additional details on how the typology and coding scheme were developed from prior work and theory, see Developing the coding scheme for supervised learning.

## Transformers beat other deep neural networks

### *Overall performance*

We evaluated the accuracy of BERT and XLNet transformer models against other leading models, CNN and LSTM, which were previously dominant architectures in this domain.[11,24] Given that we have imbalanced data for machine classification, we also report the F1 score, which is the harmonic average of precision and recall and is considered a measure of detection efficiency. As shown in Table 2, we achieved high overall accuracy scores for BERT and XLNet of 91.6% (0.13 SD) and 91.6% (0.07 SD), and F1 scores of 0.83 (0.0037 SD) and 0.84 (0.0015 SD), respectively. The standard deviations were generated from 10 cross-validation runs. While CNN and LSTM models had slightly lower accuracy, we find that both transformer models outperform the CNN and LSTM models considering both accuracy and F1 score. We report 2 to 4 percentage point improvements in the F1 scores for both transformer models. For implementation details, see Supplemental experimental procedures and Figure S1. For reference, we provide the hyperparameters used for the transformer models in Table S1. We also open sourced the model weights (see Resource availability).

The F1 scores for the transformer models are also a substantial 40 percentage points higher compared with the majority classifier (Table 2). This means the models learned to detect minority classes effectively. Briefly, the majority classifier provides a measure of the level of imbalance. For a given category, the majority classifier simply predicts the most prevalent label. For example, if 90% of training data has not been selected for a topic, then the classifier predicts all data as not selected, giving a high accuracy of 90%. Thus, for highly imbalanced data, a majority classifier can provide arbitrarily high accuracy without significant learning.[28] Because it is possible that misclassification errors may not distribute equally across the topics, in the next section, we also evaluate the performance by topics.

### *Increasing detection of imbalanced labels*

A key challenge was to evaluate whether we could improve multi-label classifications even in the presence of imbalanced data. Figure 1A shows a large percentage point increase in accuracy for all the deep learning models tested, compared with the majority classifier. This evidence of learning is especially notable for the most balanced topics (i.e., Functionality, Location, and Availability). As shown in Figure 1B, we report improvements in the F1 scores for BERT and XLNet across most topics versus the benchmark models. In particular, this result holds for the relatively imbalanced topics (i.e., Range anxiety, Service time, and Cost), which have presented technical hurdles in prior implementations.[24] In comparison with the previously leading CNN algorithm, BERT and XLNet produce F1 score increases of 1–3 percentage points on Functionality, Availability, Cost, Location, and Dealership topics and 5–7 percentage points on User interaction and Service time topics. For Range anxiety, BERT is within the statistical uncertainty of the CNN performance, while XLNet produces an increase in the F1 score of 4 percentage points. These numbers represent considerable improvements in topic level detection. For detailed point estimates, see Tables S2 and S3.

Given these promising results, next we consider some requirements for possible large-scale implementation such as computation time and scalability related to the sourcing of the training data.

## Computation time

An important metric to consider while running deep learning models for large-scale deployment is the computation time.

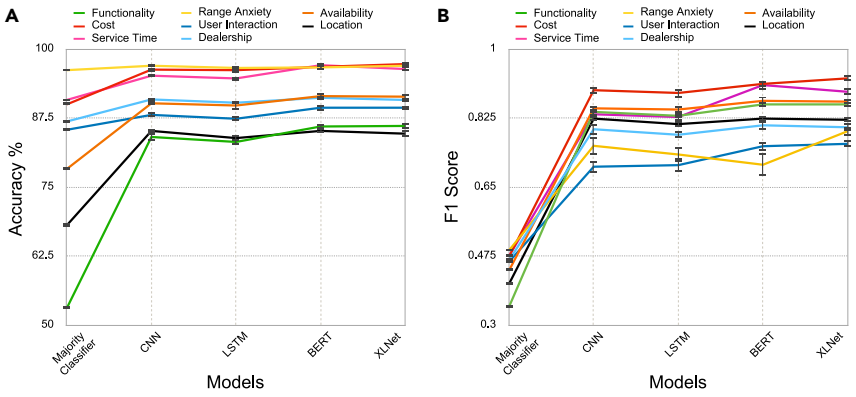



**Figure 1. Topic level classification performance**

(A) For the baseline model we use the majority classifier, which predicts the simple majority for a given topic. For higher values in accuracy, the majority classifier reflects more imbalance in the training and testing data. We find that the deep learning models outperform the majority classifier in model accuracy, particularly for more frequently occurring labels, the Functionality, Location, and Availability topics.

(B) We also compare the relative performance of the transformer models with CNN and LSTM classifiers. High F1 scores for imbalanced topics indicate strong detection of true positives. Our results indicate that transformer models, BERT and XLNet, which achieve similar performance, improve upon the CNN and LSTM benchmarks in the F1 score across all topics. The error bars represent upper and lower 95% confidence intervals.

See also Tables S2 and S3.

Deep neural networks have recently been criticized for the large amount of resources needed, such as graphics processing units (GPUs) and distributed computing clusters, frequently leading to higher costs of deployment.[29] Further, NLP researchers have also considered the environmental costs of the power consumption and $CO_2$ emissions for computing,[30] which necessarily involve trade-offs. In our application, we report the training times per epoch for BERT and XLNet as 196 and 346 s, respectively. These results were generated using four widely available NVIDIA Tesla P100 GPUs with 16 GB of memory.

We find that the training and testing times are considerably longer for the transformer models compared with CNN and LSTM. For transformers, total computing times vary from 1 to 4 h, and for CNN and LSTM, computing times vary from 1 to 90 min, depending on the number of GPUs (see Table S4 for details). We argue that the model performance improvements in the transformer models may be justified for large-scale deployment. This is because the increase in computational cost is offset by substantial gains in accuracy and F1 score. When comparing BERT and XLNet within the class of transformers, we also show BERT to be considerably faster in total computing time for a comparable level of performance. Therefore, we argue that, as further enhancements to BERT and its optimized variants are rapidly advancing in the literature,[31–33] BERT could be a preferred text classification algorithm for this domain. In the next section, we consider scalability of the models by evaluating potential sources of training data.

## Trained experts beat the crowd

In Table 3, we compare the machine classification results based on training data from a crowd of non-experts versus a group of trained expert annotators. For performance comparison of models trained with expert- and crowd-annotated data, we created a ground truth dataset by conducting researcher audits to ensure 100% agreement on the ground truth labels. See Human annotation of training data for further details. Not surprisingly, we find that human experts are closer to the ground truth (random holdout sample; $n = 100$) in both accuracy and F1 score, as shown in Table 3. This is consistent with related literature on limitations to wise crowds.[34] In fact, prior research has found gaps in general public knowledge about EVs and consumer misperceptions.[35–38] In the next section, we quantify the performance of crowd-trained versus expert-trained transformer models, using the two experimentally curated sources of training data.

### *Crowd-trained models perform poorly*

The transformer models trained with crowd-annotated data produced accuracies of 73.2% (3.85 SD) and 74.2% (4.15 SD) and F1 scores of 0.53 (0.06 SD) and 0.54 (0.07 SD) for BERT and XLNet, respectively (see Table 3). By contrast, we see a remarkable improvement in these results with the expert-trained BERT and XLNet models, which produced model accuracies of 89.1% (4.09 SD) and 91.0% (4.70 SD) and F1 scores of 0.82 (0.06 SD) and 0.85 (0.06 SD), respectively. We discovered that the enhancement in the F1 score is largely due to gains in the inter-rater reliability, which is the result of improvements in the quality of the training data between crowds and experts (see the Fleiss $\kappa$ score increase from 0.007 to 0.538 in Table 3). We argue that inter-rater agreement is critical when working with annotated data from complex domains such as EV mobility. For reference, at the sub-topic level, values for Fleiss' $\kappa$ range from $-0.001$ to 0.019 for the crowd, and from 0.30 to 0.72 for the experts, which indicates considerable disagreement on the labeling task within a sample of adults, 18 years and over, representative of the US population. See Experimental procedures for details on human annotation experiments.

Although sourcing strategies with online labor pools may be inexpensive, we find that the cost advantage does not justify the poor performance (F1 score 0.61, 0.09 SD). These results indicate that the use of low-cost crowd-sourcing approaches to build massive training sets is likely not feasible for large-scale implementation in this domain. This is in stark contrast to other deep learning domains, such as computer vision, where cheap, crowd-sourced training data can be easily acquired. For example, identifying sections of a road or public bus in an image is an easy task for the average person, but the average person cannot easily categorize the topics of EV user reviews. To

**Table 3. Ground truth evaluation of human performance versus transformer models**

| Classifier | Training set | Accuracy % (SD) | F1 score (SD) |
|---|---|---|---|
| BERT | Expert annotated | 89.1 (4.09) | 0.82 (0.06) |
| BERT | Crowd annotated | 73.2 (3.85) | 0.53 (0.06) |
| XLNet | Expert annotated | 91.0 (4.70) | 0.85 (0.06) |
| XLNet | Crowd annotated | 74.2 (4.15) | 0.54 (0.07) |
| Crowd ($\kappa$ = 0.007) | – | 73.9 (6.06) | 0.61 (0.09) |
| Human experts ($\kappa$ = 0.538) | – | 86.0 (4.40) | 0.79 (0.07) |

Cross validation was for 10 runs.

provide an example of this, in our experiments, the review, "… What an inconvenience when I need to drive to Glendale and I have a very low charge …," was cognitively difficult for general crowd annotators to correctly classify as Range anxiety, even when annotators had unrestricted access to definitions and related examples. This was not the case for most experts. As a result, for these complex domains, expert-curated training data will be required for large-scale implementation. In the next section, we compare the performance of our best classifiers, using artificial intelligence versus human intelligence.

## Possibility of super-human classification

During hand validations of the transformer-based experiments, we noticed that some test data that were not correctly labeled by the human experts were being correctly labeled by the transformer models. This caught our attention, as it indicated the possibility that BERT and XLNet could in some cases exceed the human experts in multilabel classification. In Table 3, we see that expert-trained transformer models performed about 3–5 percentage points higher in accuracy and 0.03–0.06 points higher in the F1 score compared with our human experts. In Table 4, we provide six specific examples of this phenomenon where the expert-trained transformers do better than human experts. For example, exceeding human expert benchmarks could happen in multiple ways. It could be that the algorithm correctly detects a topic that the human experts did not detect (i.e., reviews 1 and 2 in Table 4), or that it does not detect a topic that has been incorrectly labeled by an expert (i.e., reviews 4–6 in Table 4), or that the sum of misclassification errors is smaller than that of human experts (i.e., reviews 3–6 in Table 4). We also provide quantitative measures in accuracy for these examples in Table 4.

Although a full investigation of super-human performance for these transformer neural networks is outside the scope of the current study, we suggest this as an important future work. Evidence that artificial intelligence can outperform human benchmarks on multilabel classification tasks can have practical benefits for station managers and investors to be able to accurately predict system problems or examine customer needs at high resolution in ways not previously possible.

## Applications to local and regional policy

As EV consumer data expands, we comment on the possibility to apply this computational approach widely to local and regional policy analysis. We note that, previously, this type of extracted consumer intelligence has not been easily accessible to policy makers or governments due to the nature of unstructured data and issues with data access. For example, the US Department of Energy's Alternative Fuels Data Center maintains a list of all publicly accessible stations in the United States and Canada. This includes location information, such as station name, address, phone number, charging level (e.g., L1, L2, or L3), number of connectors, and operating hours with a developer-friendly API. However, these aggregated data sources do not typically include real-time usage or station availability, due to challenges with network interoperability.[16] This means that due to the presence of different charging standards of manufacturers in regional EV networks, there remain structural issues with sharing and receiving EV usage data between regions.

Recently, there has been a movement by a global consortium of public and private EV infrastructure leaders to promote open standards such as the Open Charge Point Protocol[17] and the Open Smart Charging Protocol.[39] As these technology standards become more widely adopted, there will be a rapid increase in the amount of real-time data that can be shared with researchers and analysts. For instance, a growing number of digital platform providers have begun moving toward open data. These include platforms such as Open Charge Map, Recharge, and Google Maps. In the future, it should be possible to easily merge consumer review data with other spatial features and information. This could provide a wealth of commonly used features for analysis such as socioeconomic indicators, including population, income levels, educational attainment, age, poverty rates, unemployment, and affordability of nearby housing. Other important features could include transportation economic indicators, air pollution, health data, mobile phone tracking data, point-of-interest information, and local and regional incentives.

To provide an example of possible data insights for urban policy, we conducted a spatial analysis of metropolitan and micropolitan statistical areas (MSAs and μSAs). One of the dominant topics is Availability, which is predicted when a user reports whether a given charging station is available for use. In Figure 2, we visualize the spatial distribution of predicted station availability by US census regions. To create this map, we merged the predicted review topics with counties based on shape files from the Office of Management and Budget's (OMB) 2013 specification of MSAs and μSAs. In the United States, there are 1,167 MSAs (population larger than 50,000) and 641 μSAs (population greater than 10,000), and 1,335 non-core-based statistical areas (population less than 10,000). To visualize model predictions, we standardized the predicted frequency of the Availability topic into quantiles for each census region (West, Midwest, Northeast, and South), with 0%–44%, rarely; 45%–69%, sometimes; 70%–90%, a moderate amount; and over 90%, a great deal (see Figure 2). The map reveals areas with high and low predicted Availability issues in in all core-based statistical areas.

Using this approach, we find that predicted station availability issues are not necessarily concentrated in the large central metro counties (MSAs over 1 million population), but rather

**Table 4. Examples where expert-trained transformers exceed human benchmarks**

| | Ground truth | Human expert | | Expert-trained transformers | | | |
|---|---|---|---|---|---|---|---|
| | | | | BERT | | XLNet | |
| | Labels | Labels | Acc. (%) | Labels | Acc. (%) | Labels | Acc. (%) |
| 1. "… unit says decommissioned but it will still release the charger after a long pause." | Functionality | User interaction | 75 | Functionality | 100 | Functionality | 100 |
| 2. "Thanks very busy dealership but happy to allow use of qcdc." | Functionality, Availability, Dealership | Functionality, Dealership | 87.5 | Functionality, Availability, Dealership | 100 | Functionality, Availability, Dealership | 100 |
| 3. "Charging on the quick charger - will be done by 12:15." | Functionality, User interaction | Functionality, Location | 75 | User interaction | 87.5 | User interaction | 87.5 |
| 4. "Went from 18-82% in 27 min! First time DC charging and met another nice Leaf owner who showed me how to use the machine. Thanks for the charge!" | Functionality, Service time | Functionality, Availability, Location, User interaction, Dealership | 62.5 | Service time | 87.5 | Functionality, Service time, Dealership | 87.5 |
| 5. "The CHAdeMO charger does work …. Nissan Hill had to move an ICE for me to gain access, but did so quickly. The CHAdeMO did not cost me any $ Charged quick! Don't hesitate to use." | Functionality, Availability, Cost, Dealership | Functionality, Availability, Cost, User interaction, Location, Service time, Dealership | 62.5 | Functionality, Cost, Dealership | 87.5 | Functionality, Cost, Service time, Dealership | 75 |
| 6. "So the dealer had all of their cars being serviced parked in every spot including the quick charger. I called and asked them for at least access to the quick charger and they agreed but never did anything so I left and drove to Larry h nissan. I was willing to pay because I was in a hurry and obviously the Toyota dealer doesn't want my business." | Availability, Cost, Dealership | Functionality, Availability, User interaction, Location, Dealership | 50 | Availability, Dealership | 87.5 | Availability, Location, Dealership | 75 |

away from the city centers, such as smaller μSAs of population less than 50,000. This is particularly true in the West (e.g., Oregon, Utah, Colorado, Wyoming, New Mexico) and Midwest (e.g., South Dakota and Nebraska) and Hawaii. By contrast, for the South (e.g., Texas, Alabama, Florida, North Carolina, South Carolina, Tennessee) and Northeast regions (e.g., New York, New Jersey, Massachusetts, Maryland, Pennsylvania), we find the highest frequency of availability issues in the major MSAs for the period of analysis. One primary insight from this analysis is that μSAs could be underserved with regard to station availability. In additional analyses, we also used our methodology to detect whether a specific station is functioning. Based on the rate of consumers leaving reviews at charging stations across the United States, we find that the deep learning algorithms can detect the functioning of a certain station, daily. For further details of these estimates, see Supplemental experimental procedures. This type of detection could also be done with any of our introduced topics and with expanded sample datasets from network providers.

Given the proliferation of EV policies worldwide, this spatial analysis could be expanded globally, for example, in the European Union, policies such as Alternative Fuels Infrastructure

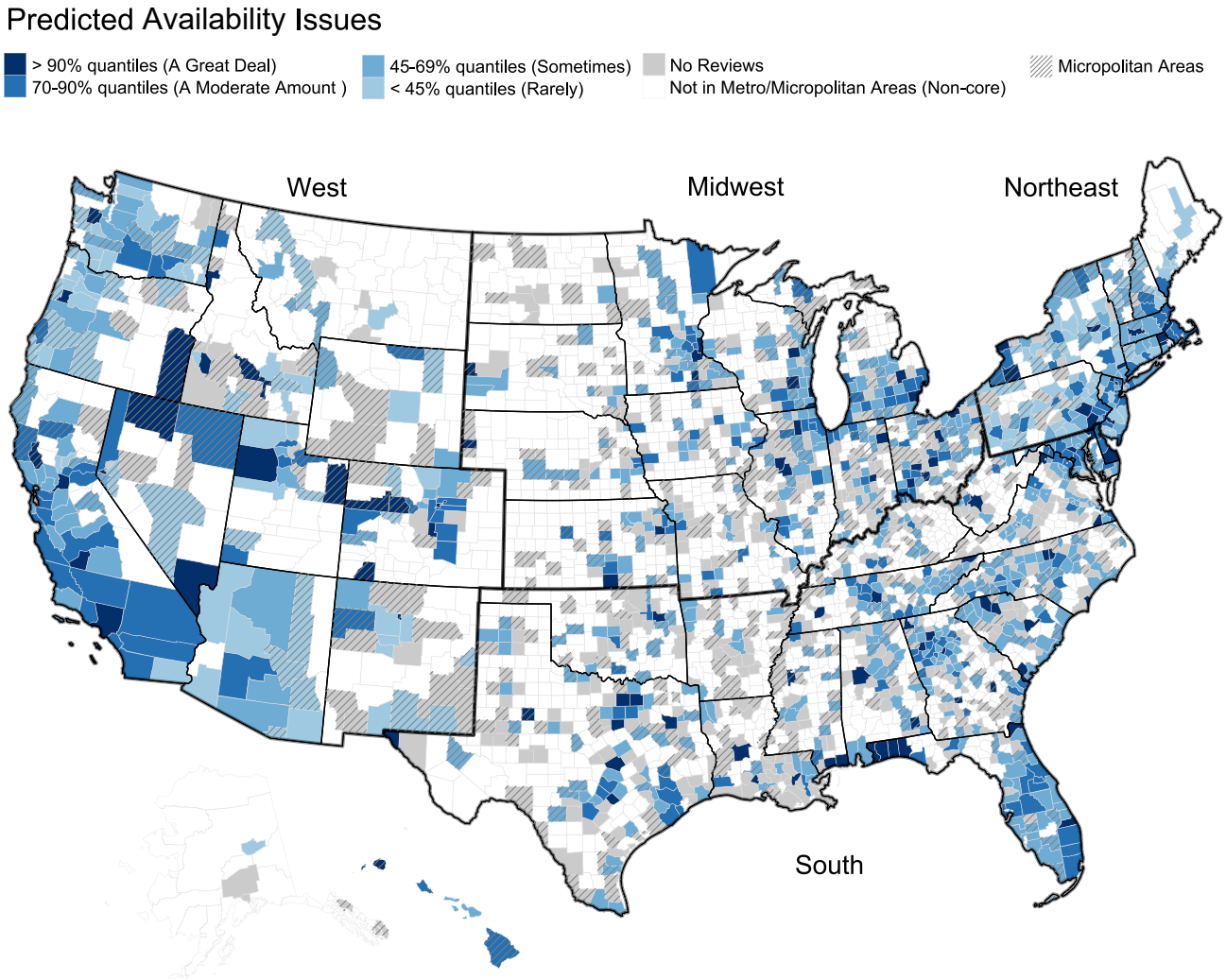



**Figure 2. Predicted discussion frequency of station availability for US metropolitan and micropolitan statistical areas**

The map reveals areas with high and low discussion frequency for predicted Availability issues in all metropolitan statistical areas (e.g., population greater than 50,000). Micropolitan statistical areas (e.g., population 10,000–49,999) have higher Availability discussions in some states in the West and Midwest regions. The algorithms predict that many micropolitan statistical areas could be underserved with regard to station availability.

Directives (previously known as the Directive on Alternative Fuels Infrastructure).[40] In addition, the European Commission has supported implementation of fast charging infrastructure through the Trans-European Network for Transport and Connecting Europe Facility Transport programs.[40,41] This type of national-scale infrastructure expansion in the European Union is part of an overall strategy by the European Union to reduce $CO_2$ emissions from the transportation sector by 60% by 2050.[42]

This capability to deploy accurate and more efficient deep learning models can be applied to evaluate other charging infrastructure rollout policies that aim to increase the number of charge points, reduce charging congestion, promote vehicle-to-grid and overnight charging, as well as solar adoption.[43] For recent reviews on how charging behavior can guide charging infrastructure implementation policy, see van der Kam et al.[43] and McCollum et al.[44] Other applications that use artificial intelligence and NLP to discover hard-to-reveal patterns in unstructured data, especially those that merge spatial information, should generate fruitful areas of future inquiry.

## Concluding remarks

In this study, we report state-of-the-art results for multilabel topic classification of consumer reviews in EV infrastructure. This represents a potential step change in our ability to aggregate data and insights for EV business model development and public policy advisory. Implementing automated topic modeling solutions has been challenging because of the technical nature of the corpus and training data imbalances. Our experimental protocols highlight the importance of the quality of training data annotations in the data processing pipeline. First, human expert annotators outperform the general crowd in both accuracy and F1 score metrics. This is due to improvements in the interrater reliability that is critical while working with data from complex domains. Second, improvements in training data

quality also produce more accurate and reliable detection. This is seen in the approximate increase of 15 percentage points in accuracy and 50% improvement in the F1 score in the expert-trained transformer models compared with the crowd-trained models (Table 3). Third, when the models are trained on top of high-quality expert curated training data, surprisingly, the transformer neural networks can outperform even human experts. This indicates evidence of super-human classification on imbalanced corpora. As deep learning models have often been criticized for their black-box nature, we suggest technical enhancements that focus on model interpretability as future work, such as through the use of rationales,[45] influence functions,[46] or sequence tagging approaches[47] that can offer deeper insights on the models and the reasons for their predictions. This is an area of active research.

Further applications of methods that we propose, particularly those that integrate artificial intelligence with real-time data and spatial analysis, can greatly enhance new ways of thinking about infrastructure management as well as economic and policy analysis. Other opportunities abound.

## EXPERIMENTAL PROCEDURES

### Resource availability

#### *Lead contact*

Further information and requests for resources and materials should be directed to and will be fulfilled by the lead contact, Dr. Omar I. Asensio (asensio@gatech.edu).

#### *Materials availability*

The trained model weights for BERT and XLNet generated in this study have been deposited in Figshare: https://doi.org/10.6084/m9.figshare.12612092.v1.

#### *Data and code availability*

The anonymized datasets and code generated during this study have been deposited in the Zenodo: https://doi.org/10.5281/zenodo.4276350 The raw data may not be posted publicly due to privacy restrictions.

### Data

We reanalyzed data derived from a nationally representative collection of unstructured consumer reviews from 12,720 charging station locations across the United States. It comprised 127,257 reviews, all written in English, by 29,532 registered and unregistered EV drivers across a 4-year duration from 2011 to 2015.[11,23,48]

The spatial coverage of the dataset includes reviews from 750 MSAs (309 large MSAs of population 1 million or more; 228 medium MSAs of population 250,000–999,999; 213 small MSAs of population 50,000–249,999). This also includes 294 μSAs (i.e., population 10,000–49,999) and 232 non-core-based statistical areas (i.e., population less than 10,000). This spatial coverage is based on the 2013 OMB delineation of MSAs and μSAs.

The data are statistically representative of the entire US EV market, which includes all major EV networks and a mix of both public and private stations, urban and rural stations, and both low and highly rated stations. The data include the text of consumer reviews and contains other useful indicators such as the timestamp of the reviews and the car make and model. We also geo-coded the station location and related points of interest using the Google Places API. However, the dataset does not contain EV transactions data, such as how many kilowatt hours were transferred. The data are also observable only on condition of a user checking in and posting a review.

This type of data is expanding globally and we estimate that there are already over 3.2 million reviews through 2020 across more than 15 charge station locator apps.[12–16] This includes English-language reviews as well as reviews in over 42 languages on all continents, such as Ukrainian, Russian, Spanish, French, German, Finnish, Italian, Croatian, Icelandic, Haitian-Creole, Ganda, Sudanese, Kinyarwanda, Afrikaans, Nyanja, Korean, Mandarin, Japanese, Indonesian, and Cebuano.

### Developing the coding scheme for supervised learning

We developed the coding scheme for our typology from prior work and theory using three strategies. First, we reviewed the extant literature to capture the most important potential behavioral issues for EV drivers. This led to identification of Range anxiety,[6,49–52] Dealership practices,[53–55] Cost,[6,52,56–58] Service time,[6,52,56,58] Availability issues,[59,60] User interaction,[61–63] station Functionality,[11,58,64] and Location.[11] Second, to find evidence of the importance of these topics from the data, we hand-coded 8,953 randomly selected reviews to validate the 8 topics from prior literature and used these to generate 34 subtopics for classification. We found that only 1% of the reviews were unclassifiable according to our 8 main categories (i.e., Other). Third, to validate the coding scheme, we also interviewed industry experts and practitioners, which allowed us to further refine our main topics and sub-topics shown in Table 1. This included informal communications with representatives from firms such as General Motors, ChargePoint, ReCharge Technologies, Electrada, Electrify America, and charging station managers (e.g., representatives from Ford and Georgia Tech Parking and Transportation Services) who were not directly involved in the research.

### Human annotation of training data

A common criticism with deep neural networks is the high cost and annotator skill requirements for implementations in specialized corpora. We evaluated possible methods to lower implementation costs, such as crowd sourcing by using online labor pools for human annotation. This led us to conduct human annotator experiments with two training sets, each labeled by a crowd of non-experts and a small group of trained experts. Given the known possible biases with historical data, we investigated whether protocols related to the labeling of the training data could have an impact on performance.[65,66]

The crowd and expert annotators each labeled a random sample of 10,652 reviews. We used an 80:10:10 split for training, validation, and testing, which met our objective of having equal amounts of training data for both annotator groups. We conducted statistical tests to determine whether the sampled training dataset was representative of the full dataset in key observable station characteristics. We confirmed that the training dataset was statistically representative in the mix of urban and non-urban stations (t test, $p = 0.426$) and public and private stations (t test, $p = 0.709$), as well as by station points of interest (t test, $p = 0.802$), e.g., retail, shopping, workplace, transit centers, etc. We also found that the training data were not statistically different in topic distribution from the predictions of the full dataset (Kolmogorov-Smirnov test, $p = 0.9801$).

#### *Crowd annotators*

For the crowd-sourced training data sample, 1,000 US adults (age 18+) were pre-recruited via a Qualtrics online panel using their popular online survey platform. The crowd was statistically sampled on the basis of age, income, education, and sex, representative of the US population. This is important to mitigate possible human rater biases that could arise when discussing environmental topics. To enhance understanding of the domain-specific terminology for the general crowd, definitions and examples for the topics and sub-topics as shown in Table 1 were provided for annotation along with a supporting diagram containing typical components of an EV charging station (see Figures S2 and S3). We report the Fleiss κ for crowd annotators as 0.007.

#### *Expert annotators*

For the expert-sourced training data sample, five student annotators with technical backgrounds were recruited and trained in a facilitated focus group. They were instructed to recognize the domain-specific topics using a detailed training manual for the annotation. To support scientific replication and to document the protocols, we have open sourced this training manual.[67] These protocols were developed in consultation with EV industry experts who have been in contact with the researchers. Although our expert annotators have been trained to recognize domain-specific terminology, we acknowledge that we were not able to compare the performance of our expert annotators with that of EV industry professionals due to cost reasons. Despite this limitation, we find that our human experts were two orders of magnitude more reliable in the annotation (76-fold increase in our reliability measure) than the crowd annotators ($\kappa = 0.538$ and $\kappa = 0.007$, respectively). See Model metrics under Performance measures for additional details on computing Fleiss' κ.

To provide a greater control over the labeling task, we developed a custom web application used by the expert annotators as shown in Figure S3. The web app provides efficient database support for random sampling from a large dataset and overcomes latency and scaling challenges that we encountered during crowd annotation in popular survey software.

### *Ground truth labels*

To generate the ground truth labels, we followed the same training protocols used by the expert annotators. Then, we randomly sampled 100 overlapping reviews that were annotated by both annotator groups to enable performance comparisons. On this sample, we conducted an additional round of researcher audits that validated 100% agreement on the annotations. Given that the human experts exhibited some level of disagreement (Fleiss' $\kappa = 0.538$, Table 3), this sample was used to benchmark the performance of the US crowd and the human experts. The results of these comparisons as well as their statistical uncertainty are reported in Table 3. To generate the uncertainty, we performed a cross validation using block randomization with 10 equal-sized blocks of ground truth data.

## Performance measures

### *Model metrics*

To assess model performance, we report the micro-averaging F1 score, which is a standard metric for classifier performance on detection of false positives and false negatives. We used standard measures for multilabel accuracy, where annotators could choose multiple labels per review. Our overall accuracy metric accounts for partially correct matches. By convention, this is equivalent to 1 − Hamming loss, where the Hamming loss is an *xor* calculation of the dissimilarity (i.e., a fraction of wrong labels compared with the total number of labels). For $L$ categories classified on a sample of size $N$, the accuracy can be calculated as:

$$\begin{aligned}\text{Overall Accuracy} &= 1 - \text{Hamming Loss} \\ &= 1 - \frac{1}{|N| \cdot |L|} \sum_{i=1}^{|N|} \sum_{j=1}^{|L|} xor\left(y_{i,j}, z_{i,j}\right)\end{aligned} \quad \text{(Equation 1)}$$

For example, if a multilabel prediction [1, 1, 1, 0] had a true label [1, 1, 1, 1], the accuracy is 3/4 or 75%.

### *Interrater Reliability*

To measure the interrater agreement level among the annotators, we used Fleiss' $\kappa$, which allows for the measurement of agreement between multiple annotators (i.e., more than 2). It is calculated as shown below:

$$\kappa = \frac{\overline{P} - \overline{P_e}}{1 - \overline{P_e}}, \quad \text{(Equation 2)}$$

where $\overline{P}$ is the average number of agreements on all annotations between rater pairs for the reviews, and $\overline{P_e}$ is the sum of squares of the probability share for the assignment to a topic. As $\kappa$ is bounded between −1 and 1, when $\kappa$ is less than 0, agreement between raters is occurring below what would be expected at random, while a $\kappa$ above 0 means that agreement between raters is occurring at more than what would be expected by random chance.[68] For more information, see Fleiss.[69]

## Ethics statement

Human subjects research was conducted under the approved Institutional Review Board Protocol H18250.

## SUPPLEMENTAL INFORMATION

Supplemental information can be found online at https://doi.org/10.1016/j.patter.2020.100195.

## ACKNOWLEDGMENTS

We gratefully acknowledge funding support by the National Science Foundation (awards 1945332 and 1931980), a Microsoft Azure Sponsorship, and the Ivan Allen College Dean's SGR-C Award. For high-performance computing infrastructure support, we thank Srinivas Aluru and IDEaS. For assistance in constructing the typology of charging station reviews, we thank Mary Elizabeth Burke and Soobin Oh. For other valuable research assistance, we thank M. Cade Lawson, Christopher J. Blanton, Duncan Hemauer, Dunsin Awodele, Kira O'Hare, Matteo Zullo, Michael Weiner, Semir Sarajlic, Steven Leone, and Taylor Sparacello. This research was also supported in part through the research cyberinfrastructure resources and services provided by the Partnership for an Advanced Computing Environment (PACE) at the Georgia Institute of Technology, Atlanta, Georgia, USA.

## AUTHOR CONTRIBUTIONS

Conceptualization, O.I.A. and S.D.; Methodology, S.H., D.J.M., S.D., and O.I.A.; Investigation, S.D., S.H., and O.I.A.; Validation, S.D., S.H., and D.J.M.; Data Curation, S.H. and D.J.M.; Writing – Original Draft, O.I.A., S.D., S.H., and D.J.M.; Writing – Review & Editing, O.I.A. and S.H.; Visualization, S.H. and S.D.; Software, S.H. and S.D.; Resources, O.I.A.; Funding Acquisition, O.I.A.

## DECLARATION OF INTERESTS

The authors declare no competing interests.

Received: July 15, 2020
Revised: November 16, 2020
Accepted: December 17, 2020
Published: January 22, 2021

## REFERENCES

1. Environmental Protection Agency (2018). Inventory of U.S. Greenhouse Gas Emissions and Sinks: 1990-2016. document No. 430-R-18-003. https://www.epa.gov/ghgemissions/inventory-us-greenhouse-gas-emissions-and-sinks-1990-2016.
2. National Research Council (2010). Hidden Costs of Energy: Unpriced Consequences of Energy Production and Use (National Academies Press).
3. Hoekstra, A. (2019). The underestimated potential of battery electric vehicles to reduce emissions. Joule *3*, 1412–1414.
4. Environmental Protection Agency. (2018). Greenhouse Gas Emissions from a Typical Passenger Vehicle. document No. 420-F-18-008. https://www.epa.gov/greenvehicles/greenhouse-gas-emissions-typical-passenger-vehicle.
5. Department of Energy. (2019). Electric vehicles: tax credits and other incentives database. https://www.energy.gov/eere/electricvehicles/electric-vehicles-tax-credits-and-other-incentives.
6. Carley, S., Krause, R.M., Lane, B.W., and Graham, J.D. (2013). Intent to purchase a plug-in electric vehicle: a survey of early impressions in large us cites. Transp. Res. D Transport Environ. *18*, 39–45.
7. Sheldon, L.T., DeShazo, J.R., and Carson, R.T. (2017). Electric and plug-in hybrid vehicle demand: lessons for an emerging market. Econ. Inq. *55*, 695–713.
8. Hardman, S., Jenn, A., Tal, G., Axsen, J., Beard, G., Daina, N., Figenbaum, E., Jakobsson, N., Jochem, P., Kinnear, N., et al. (2018). A review of consumer preferences of and interactions with electric vehicle charging infrastructure. Transp Res. D Transport Environ. *62*, 508–523.
9. Anderson, J.E., Lehne, M., and Hardinghaus, M. (2018). What electric vehicle users want: real-world preferences for public charging infrastructure. Int. J. Sustain. Transp. *12*, 341–352.
10. Brückmann, G.M., and Bernauer, T. (2020). What drives public support for policies to enhance electric vehicle adoption? Environ. Res. Lett. *15*, 094002.
11. Asensio, O.I., Alvarez, K., Dror, A., Wenzel, E., Hollauer, C., and Ha, S. (2020). Real-time data from mobile platforms to evaluate sustainable transportation infrastructure. Nat. Sustain. *3*, 463–471.
12. Recargo (2020). Plugshare key features and benefits. https://recargo.com/plugshare.html.
13. Chargemap (2020). Chargemap's community. https://chargemap.com/community.

14. Open Charge Map (2020). Open charge map community. https://community.openchargemap.org/.

15. ChargePoint; Chargepoint map (2020). https://na.chargepoint.com/charge_point.

16. Recharge (2020). United States EV charging network interoperability is a lie. https://www.evpassport.com/post/us-ev-charging-networkinteroperability-is-a-lie.

17. Open Charge Alliance (2020). Open charge point protocol 2.0.1 specification. https://www.openchargealliance.org/protocols/ocpp-201/.

18. LeCun, Y., and Bengio, Y. (1998). Convolutional networks for images, speech, and time series. In The Handbook of Brain Theory and Neural Networks, M.A. Arbib, ed. (MIT Press), pp. 255–258.

19. Kim, Y. (2014). Convolutional neural networks for sentence classification. In Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP) (Association for Computational Linguistics), pp. 1746–1751, https://doi.org/10.3115/v1/D14-1181.

20. Zhang, Y., and Wallace, B.C. (2017). A sensitivity analysis of (and practitioners' guide to) convolutional neural networks for sentence classification. In Proceedings of the Eighth International Joint Conference on Natural Language Processing, *1*, arXiv:1510.03820.

21. Yin, W., Kann, K., Yu, M., and Schütze, H. (2017). Comparative study of cnn and rnn for natural language processing. arXiv, arXiv:1702.01923.

22. Hochreiter, S., and Schmidhuber, J. (1997). Long short-term memory. Neural Comput. *9*, 1735–1780.

23. Alvarez, K., Dror, A., Wenzel, E., and Asensio, O.I. (2019). Evaluating electric vehicle user mobility data using neural network based language models. In Proceedings of the 98th Annual Meeting of the Transportation Research Board https://trid.trb.org/view/1573203%20.

24. Ha, S., Marchetto, D.J., Burke, M.E., and Asensio, O.I. (2020). Detecting behavioral failures in emerging electric vehicle infrastructure using supervised text classification algorithms. In Proceedings of the 99th Annual Meeting of the Transportation Research Board https://par.nsf.gov/biblio/10165854%20.

25. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L., and Polosukhin, I. (2017). Attention is all you need. In Advances in Neural Information Processing Systems 30 https://papers.nips.cc/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf.

26. Devlin, J., Chang, M., Lee, K., and Toutanova, K. (2019). Bert: pretraining 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*, (Long and Short Papers). arXiv:1810.04805.

27. Yang, Z., Dai, Z., Yang, Y., Carbonell, J., Salakhutdinov, R.R., and Le, Q.V. (2019). XLNet: generalized autoregressive pretraining for language understanding. In Advances in Neural Information Processing Systems https://papers.nips.cc/paper/2019/file/dc6a7e655d7e5840e66733e9ee67cc69-Paper.pdf.

28. Schapire, R.E. (1990). The strength of weak learnability. Machine Learn. *5* (2), 197–227.

29. Yan, F., Ruwase, O., He, Y., and Chilimbi, T. (2015). Performance modeling and scalability optimization of distributed deep learning systems. In Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining https://www.cse.unr.edu/~fyan/Paper/Feng-KDD15.pdf.

30. Strubell, E., Ganesh, A., and McCallum, A. (2019). Energy and policy considerations for deep learning in NLP. In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. https://doi.org/10.18653/v1/P19-1355.

31. Lan, Z., Chen, M., Goodman, S., Gimpel, K., Sharma, P., and Soricut, R. (2020). Albert: a lite bert for self-supervised learning of language representations. In International Conference on Learning Representations, arXiv:1909.11942.

32. Liu, W., Zhou, P., Wang, Z., Zhao, Z., Deng, H., and Ju, Q. (2020). FastBERT: a self-distilling BERT with adaptive inference time. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics https://www.aclweb.org/anthology/2020.acl-main.537.pdf.

33. Sanh, V., Debut, L., Chaumond, J., and Wolf, T. (2019). Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. arXiv, arXiv:1910.01108.

34. Surowiecki, J. (2005). The Wisdom of Crowds (New York: Anchor).

35. Roberson, L.A., and Helveston, J.P. (2020). Electric vehicle adoption: can short experiences lead to big change? Environ. Res. Lett. *15*, 0940c3.

36. Krause, R.M., Carley, S.R., Lane, B.W., and Graham, J.D. (2013). Perception and reality: public knowledge of plug-in electric vehicles in 21 us cities. Energy Policy *63*, 433–440.

37. Axsen, J., Langman, B., and Goldberg, S. (2017). Confusion of innovations: mainstream consumer perceptions and misperceptions of electric-drive vehicles and charging programs in Canada. Energy Res. Soc. Sci. *27*, 163–173.

38. Wang, S., Wang, J., Li, J., Wang, J., and Liang, L. (2018). Policy implications for promoting the adoption of electric vehicles: do consumer's knowledge, perceived risk and financial incentive policy matter? Transp. Res. A Policy Pract. *117*, 58–69.

39. Open Charge Alliance. (2020). Open smart charging protocol 2.0 specification. https://www.openchargealliance.org/protocols/oscp-20/.

40. European Parliament and Council of the European Union. (2014). Directive 2014/94/eu of the European parliament and of the Council of 22 October 2014 on the deployment of alternative fuels infrastructure text with EEA relevance. Official J. Eur. Union *57* (L307), 1–20.

41. TEN-T. (2015). EU-funded fast-charge network opens up pan-European travel for EV drivers. https://www.cegc-project.eu/images/TEN-T_Closing_event_press_release.pdf%20.

42. European Commission, Directorate-General for Mobility and Transport (2011). White Paper on Transport: Roadmap to a Single European Transport Area: Towards a Competitive and Resource-Efficient Transport System (Office of the European Union).

43. Kam, M.V.D., Sark, W.V., and Alkemade, F. (2020). Multiple roads ahead: how charging behavior can guide charging infrastructure roll-out policy. Transp. Res. D Transport Environ. *85*, 102452.

44. McCollum, D.L., Wilson, C., Bevione, M., Carrara, S., Edelenbosch, O.Y., Emmerling, J., Guivarch, C., Karkatsoulis, P., Keppo, I., Krey, V., et al. (2018). Interaction of consumer preferences and climate policies in the global transition to low-carbon vehicles. Nat. Energy *3*, 664–673.

45. Zaidan, O.F., Eisner, J., and Piatko, C. (2008). Machine learning with annotator rationales to reduce annotation cost. In Proceedings of the NIPS* 2008 Workshop on Cost Sensitive Learning https://cs.jhu.edu/~jason/papers/zaidan+al.nipsw08.pdf.

46. Serrano, S., and Smith, N.A. (2019). Is attention interpretable? In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics https://www.aclweb.org/anthology/P19-1282.pdf.

47. Nguyen, A.T., Wallace, B.C., Li, J.J., Nenkova, A., and Lease, M. (2017). Aggregating and predicting sequence labels from crowd annotations. In Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics. https://doi.org/10.18653/v1/P17-1028.

48. Asensio, O.I., Mi, X., and Dharur, S. (2020). Using machine learning techniques to aid environmental policy analysis: a teaching case in big data and electric vehicle infrastructure. Case Studies in the Environment, 961302, https://doi.org/10.1525/cse.2020.961302.

49. Rauh, N., Franke, T., and Krems, J.F. (2015). Understanding the impact of electric vehicle driving experience on range anxiety. Hum. Factors *57*, 177–187.

50. Jung, M.F., Sirkin, D., Gür, T.M., and Steinert, M. (2015). Displayed uncertainty improves driving experience and behavior: the case of range anxiety in an electric car. In Proceedings of the 33rd Annual ACM Conference on

Human Factors in Computing Systems. https://doi.org/10.1145/2702123.2702479.

51. Noel, L., and Sovacool, B.K. (2016). Why did better place fail?: range anxiety, interpretive flexibility, and electric vehicle promotion in Denmark and Israel. Energy Policy *94*, 377–386.
52. Egbue, O., and Long, S. (2012). Barriers to widespread adoption of electric vehicles: an analysis of consumer attitudes and perceptions. Energy Policy *48*, 717–729.
53. Rubens, G.Z., Noel, L., and Sovacool, B.K. (2018). Dismissive and deceptive car dealerships create barriers to electric vehicle adoption at the point of sale. Nat. Energy *3*, 501–507.
54. Matthews, L., Lynes, J., Riemer, M., Matto, T.D., and Cloet, N. (2017). Do we have a car for you? Encouraging the uptake of electric vehicles at point of sale. Energy Policy *100*, 79–88.
55. Lynes, J. (2018). Dealerships are a tipping point. Nat. Energy *3*, 457–458.
56. Hidrue, M.K., Parsons, G.R., Kempton, W., and Gardner, M.P. (2011). Willingness to pay for electric vehicles and their attributes. Resour. Energy Econ. *33*, 686–705.
57. Nicolson, M., Huebner, G.M., Shipworth, D., and Elam, S. (2017). Tailored emails prompt electric vehicle owners to engage with tariff switching information. Nat. Energy *2*, 1–6.
58. Kühl, N., Goutier, M., Ensslen, A., and Jochem, P. (2019). Literature vs. Twitter: empirical insights on customer needs in e-mobility. J. Clean. Prod. *213*, 508–520.
59. Kempton, W., Tomic, J., Letendre, S., Brooks, A., and Lipman, T. (2001). Vehicle-to-grid power: battery, hybrid, and fuel cell vehicles as resources for distributed electric power in California (UC Davis Institute of Transportation Studies). https://escholarship.org/uc/item/5cc9g0jp.
60. Liao, F., Molin, E., and Wee, B.V. (2017). Consumer preferences for electric vehicles: a literature review. Transport Rev. *37*, 252–275.
61. Burgess, M., King, N., Harris, M., and Lewis, E. (2013). Electric vehicle drivers' reported interactions with the public: driving stereotype change? Transp. Res. F Traffic Psychol. Behav. *17*, 33–44.
62. Morstyn, T., Farrell, N., Darby, S.J., and McCulloch, M.D. (2018). Using peer-to-peer energy-trading platforms to incentivize prosumers to form federated power plants. Nat. Energy *3*, 94–101.
63. Lee, Z.J., Pang, J.Z.F., and Low, S.H. (2020). Pricing EV charging service with demand charge. Electric Power Syst. Res. *189*, 106694.
64. National Research Council (2015). Overcoming Barriers to Deployment of Plug-In Electric Vehicles (National Academies Press).
65. Rambachan, A., Kleinberg, J., Ludwig, J., and Mullainathan, S. (2020). An economic perspective on algorithmic fairness. AEA Pap. Proc. *110*, 91–95.
66. Cowgill, B., and Tucker, C. (2017). Algorithmic bias: a counterfactual perspective. In Workshop on Trustworthy Algorithmic Decision-Making (Arlington, VA: NSF Trustworthy Algorithms) http://www.columbia.edu/~bc2656/papers/NSF-Workshop-Cowgill.pdf%20.
67. Ha, S., and Marchetto, D.J. (2020). Labeling sentiment and topics of user generated reviews on electric vehicle charging experience for supervised machine learning. https://github.com/asensio-lab/transformer-EV-topic-classification/blob/master/training-manual/training-manual.pdf.
68. Landis, J.R., and Koch, G.G. (1977). The measurement of observer agreement for categorical data. Biometrics *33*, 159–174.
69. Fleiss, J.L. (1971). Measuring nominal scale agreement among many raters. Psychol. Bull. *76*, 378.