Cite paper as:
Ha, S., Marchetto, D. J., Dharur, S., & Asensio, O. I. (2021). Topic classification of electric vehicle consumer experiences with transformer-based deep learning. Patterns, 2(2), 100195.
https://doi.org/10.1016/j.patter.2020.100195

Topic Classification of Electric Vehicle Consumer Experiences with Transformer-Based Deep Learning

Sooji Ha[1,2], Daniel J. Marchetto[3], Sameer Dharur[4], 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

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

[*] 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.

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.

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 $27.6 billion USD 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 reduce emissions from 244 to 98g/km, and this number could further decrease to 10g/km with renewable energy integration.[3] The environmental benefits range by fuel type with reported carbon intensities of 8,887 grams CO_2 per gallon of gasoline, and 10,180 grams CO_2 per gallon of diesel.[4] Government-driven incentives for switching to electric vehicles, including utility rebates, tax credits, exemptions and other policies, have been rolled out in many U.S. states.[5-7] In this effort, public charging infrastructure remains a critical complementary asset to 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 minute credit card transaction."* A multi-label 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. Therefore, text classification algorithms that can automatically perform multi-label classification are needed to interpret the data. Being able to do multi-label 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, multi-label 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 multi-labeled 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. Unbalanced data creates challenges for models to learn important but less frequently occurring labels 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 to the complex task of multi-label 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 (RNNs)[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, this has increased our understanding of potential EV charging infrastructure issues such as the prevalence of negative consumer experiences in urban locations as compared to non-urban locations.[11,23,24] While 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 under-detection, particularly in corpora with wide-ranging topics and possible 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, multi-label topic classification is needed to understand behavioral foundations of user interactions in electric mobility.

In this paper, we achieve state-of-the-art multi-label topic classification in this domain using 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, versus 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 & 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 to 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 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 *Dealerships* topic refers to comments concerning specific dealerships and user's associated charging experiences. Reviews that do not fall into the previous 8 topics refer to the *Other* topic, which are relatively rare. For more information on the robustness of typology, see Supplemental Experimental Procedures and Table S5-S7 in 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 for 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 section.

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 s.d.) and 91.6% (0.07 s.d.), and F1 scores of 0.83 (0.0037 s.d.) and 0.84 (0.0015 s.d.), 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 in the Supplemental Information. For reference, we provide the hyper-parameters 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 evaluated 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, as compared with the majority classifier. This evidence of learning is especially notable for the most balanced topics (e.g. *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 (e.g. *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, 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 Table S2 and S3 in the Supplemental Information.

Given these promising results, next we consider some requirements for possible large-scale implementation related to 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. Deep neural networks have 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 seconds, respectively. These results were generated using 4 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 hours and for CNN and LSTM, computing times vary from 1 to 90 minutes, 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 note that as further enhancements to BERT and its optimized variants are rapidly advancing in the literature,[31-33] we argue that 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 section 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 s.d.) and 74.2% (4.15 s.d.) and F1 scores of 0.53 (0.06 s.d.) and 0.54 (0.07 s.d.) 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 s.d.) and 91.0% (4.70 s.d.) and F1 scores of 0.82 (0.06 s.d.) and 0.85 (0.06 s.d.), 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 Fleiss' κ 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' κ range from -0.001 to 0.019 for the crowd, and 0.30 to 0.72 for the experts, which indicate considerable disagreement on the labeling task within a sample of 18+ adults representative of the U.S. population. See Experimental Procedures for details on human annotation experiments.

While 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 s.d.). These results indicate that the use of low-cost crowd-sourcing approaches to build massive training sets are 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 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 implementations. In the next section, we compared the performance of our best classifiers using artificial intelligence versus human intelligence.

Possibility of Super-Human Classification

During hand validations of the transformers-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 multi-label 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 as compared to our human experts. In Table 4, we provide 6 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 superhuman 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 multi-label classification tasks can benefit 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 for Local and Regional Policy

As EV consumer reviews 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 U.S. Department of Energy's (DOE) Alternative Fuels Data Center maintains a list of all publicly accessible stations in the U.S. 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 by manufacturers and 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 (OCPP)[17] and the Open Smart Charging Protocol (OSCP).[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 towards 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 reviews data with other spatial features and information. This could provide a wealth of commonly used features for analysis such as socio-economic 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 U.S. 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 *Availability* topic into quantiles for each census region (West, Midwest, Northeast, and South), where 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* consumer discussions 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 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 under-served 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 U.S., we find that the deep learning algorithms can detect 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 Directives, or AFID (previously known as the Directive on Alternative Fuels Infrastructure, or DAFI).[40] In addition, the European Commission has supported implementation of fast charging infrastructure through the Trans-European Network for Transport (TEN-T) and Connecting Europe Facility Transport (CEF-T) programs.[40,41] This type of national scale infrastructure expansion in the EU 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 roll-out 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 multi-label 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 both in accuracy and F1 score metrics. This is due to improvements in the inter-rater 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 as compared to 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 been 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 to Figshare DOI: 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 to the Zenodo repository at: https://doi.org/10.5281/zenodo.4276350. The raw data may not be posted publicly due to privacy restrictions.

Data

We reanalyze data derived from a nationally representative collection of unstructured consumer reviews from 12,720 charging station locations across the United States. It comprises 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 metropolitan statistical areas (309 large MSAs of population 1 million or more; 228 medium MSAs population of 250,000-999,999; 213 small MSAs population of 50,000-249,999). This also includes 294 micropolitan statistical areas (e.g. μSA population 10,000-49,999), and 232 non-core-based statistical areas (e.g. population less than 10,000). This spatial coverage is based on the 2013 OMB delineation of metropolitan statistical areas (MSA) and micropolitan statistical areas.

The data is statistically representative of the entire U.S. 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 includes the text of consumer reviews and contains other useful indicators such as the timestamp of the reviews, 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 has how many kWh were transferred. The data is also only observable conditional on 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 in 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 sub-topics for classification. We found that only 1% of the reviews were unclassifiable according to our 8 main categories (e.g. *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 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).

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 an equal number of training data for both annotator groups. We conducted statistical tests to determine whether the sampled training dataset is representative of the full dataset in key observable station characteristics. We confirmed that the training dataset is statistically representative in the mix of urban and non-urban stations (t-test p-value 0.426), public and private stations (t-test p-value 0.709), as well as by station points of interest (t-test p-value 0.802), e.g. retail, shopping, workplace, and transit centers, etc.). We also found that the training data was not statistically different in topic distribution from the predictions of the full dataset (Kolmogorov-Smirnov test p-value 0.9801).

Crowd Annotators. For the crowd-sourced training data sample, 1,000 U.S. 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 U.S. 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-topic as shown in Table 1 were provided for annotation along with a supporting diagram containing typical components of an EV charging station (See Figure S2 and Figure S3 in the Supplemental Information). We report the Fleiss' Kappa 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 are not able to compare the performance of our expert annotators to EV industry professionals due to cost reasons. Despite this limitation however, we find that our human experts are two orders of magnitude more reliable in the annotation (76-fold increase in our reliability measure) versus the crowd annotators (κ= 0.538 and κ= 0.007, respectively). See the Model Metrics section under Performance Measures for additional details on computing Fleiss' Kappa.

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 U.S. 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. In order 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 use standard measures for multi-label 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 to the total number of labels). For L categories classified on a sample of size N, the accuracy can be calculated as:

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

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

Inter-Rater Reliability. To measure the inter-rater agreement level among the annotators, we used Fleiss' Kappa (κ), which allows for the measurement of agreement between multiple annotators (e.g., more than 2). It is calculated as below:

$$\kappa = \frac{\bar{P} - \bar{P}_e}{1 - \bar{P}_e}, \tag{2}$$

where $\bar{P}$ is the average number of agreements on all annotations between rater pairs for the reviews, and $\bar{P}_e$ is the sum of squares of the probability share for the assignment to a topic. As κ is bounded between -1 and 1, when κ is less than 0, agreement between raters is occurring below what would be expected at random, while a κ above 0 means that agreement between raters is occurring 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 (IRB) Protocol No. H18250.

SUPPLEMENTAL INFORMATION

Additional details on model implementations can be found in the Supplemental Information.

ACKNOWLEDGMENTS

We gratefully acknowledge funding support by the National Science Foundation (Award No. 1945332 and Award No. 1931980), 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-Reviewing & 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.

REFERENCES

1. Environmental Protection Agency. (2018). Inventory of U.S. Greenhouse Gas Emissions and Sinks: 1990-2016. document No. 430-R-18-003.
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(6):1412 – 1414.
4. Environmental Protection Agency. (2018). Greenhouse gas emissions from a typical passenger vehicle. document No. 420-F-18-008.
5. Department of Energy. Electric vehicles: Tax credits and other incentives database. (2019). Access date: 07/31/2019, https://www.energy.gov/eere/electricvehicles/ele ctric-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. *Transportation Research Part D: Transport and Environment*, 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. *Economic Inquiry*, 55(2):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. *Transportation Research Part D: Transport and Environment*, 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. *International Journal of Sustainable Transportation*, 12(5):341–352.

10. Brückmann, G. M., and Bernauer, T. (2020). What drives public support for policies to enhance electric vehicle adoption? *Environmental Research Letters*.

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. *Nature Sustainability*, 3:463–471.

12. Recargo. Plugshare key features and benefits. (2020). Access date: 02/13/2020, https://recargo.com/plugshare.html.

13. Chargemap. Chargemap’s community. (2020). Access date: 10/27/2020, https: //chargemap.com/community.

14. Open Charge Map. Open charge map community. (2020). Access date: 10/27/2020, https://community.openchargemap.org/.

15. ChargePoint. Chargepoint map. (2020). Access date: 10/27/2020, https://na.chargepoint.com/charge_point.

16. Recharge. United states EV charging network interoperability is a lie. (2020). Access date: 08/31/2020, https://www.evpassport.com/post/us-ev-charging-networkinteroperability-is-a-lie.

17. Open Charge Alliance. Open charge point protocol 2.0.1 specification. (2020). Released: 03/31/2020.

18. LeCun, Y., and Bengio, Y. (1998). Convolutional networks for images, speech, and time series. *The handbook of brain theory and neural networks*, 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)*.

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 (Volume 1: Long Papers)*.

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

22. Hochreiter, S., and Schmidhuber, J. (1997). Long short-term memory. *Neural computation*, 9(8):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*.

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*.

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*.

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)*.

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*.

28. Schapire, R. E. (1990). The strength of weak learnability. *Machine Learning*, 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*.

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*.

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*.

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*.

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

34. Surowiecki, J. (2005), *The wisdom of crowds*. Anchor.

35. Roberson, L. A., and Helveston, J. P. (2020). Electric vehicle adoption: can short experiences lead to big change? *Environmental Research Letters*, 15(9):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 Research & Social Science*, 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? *Transportation Research Part A: Policy and Practice*, 117:58–69.

39. Open Charge Alliance. Open smart charge protocol 2.0 specification. (2020).

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 Journal of the European Union*, 57(L307):1–20.

41. TEN-T. Eu-funded fast-charge network opens up pan-european travel for EV drivers. (2015).

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. *Transportation Research Part D: Transport and Environment*, 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. *Nature Energy*, 3(8):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*.

46. Serrano, S., and Smith, N. A. (2019). Is attention interpretable? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*.

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*.

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.

49. Rauh, N., Franke, T., and Krems, J. F. (2015). Understanding the impact of electric vehicle driving experience on range anxiety. *Human Factors*, 57(1):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*.

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. *Nature Energy*, 3(6):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. *Nature Energy*, 3(6):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. *Resource and Energy Economics*, 33(3):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. *Nature Energy*, 2(6):1–6.

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

60. Liao, F., Molin, E., and Wee, B. V. (2017). Consumer preferences for electric vehicles: a literature review. *Transport Reviews*, 37(3):252–275.

61. Burgess, M., King, N., Harris, M., and Lewis, E. (2013). Electric vehicle drivers' reported interactions with the public: Driving stereotype change? *Transportation Research Part F: Traffic Psychology and Behavior*, 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. *Nature Energy*, 3(2):94–101.

63. Lee, Z. J., Pang, J. Z. F., and Low, S. H. (2020). Pricing EV charging service with demand charge. *Electric Power Systems Research*, 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. In *AEA Papers and Proceedings*, 110:91–95.

66. Cowgill, B., and Tucker, C. (2017). Algorithmic bias: A counterfactual perspective. In *Workshop on Trustworthy Algorithmic Decision-Making*.

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(1):159–174.

69. Fleiss, J. L. (1971). Measuring nominal scale agreement among many raters. *Psychological bulletin*, 76(5):378.

FIGURE TITLES AND LEGENDS

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, *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 Table S2 and S3.

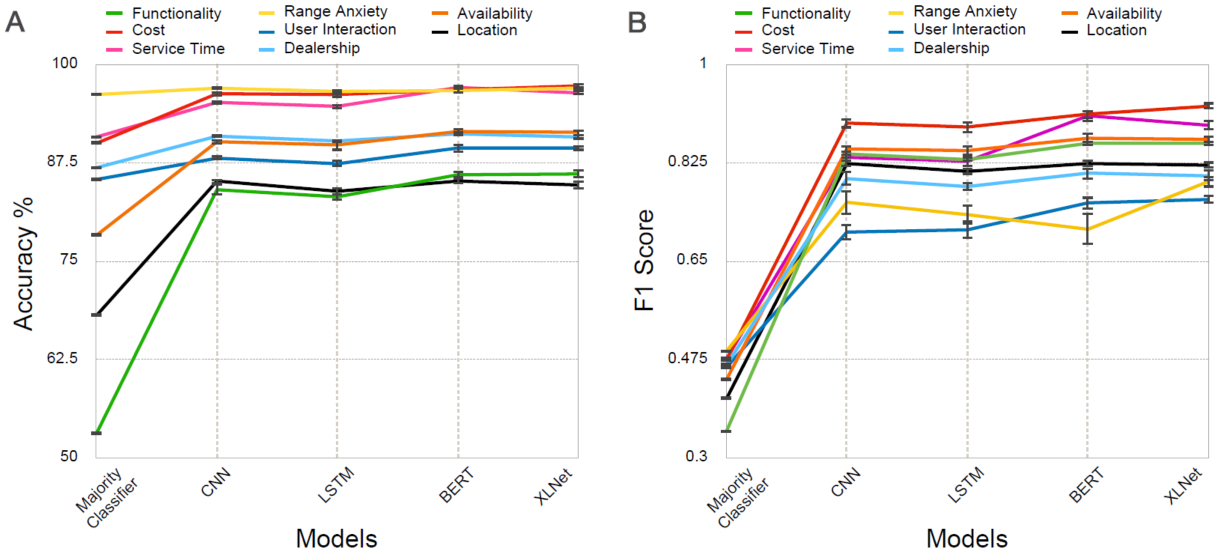
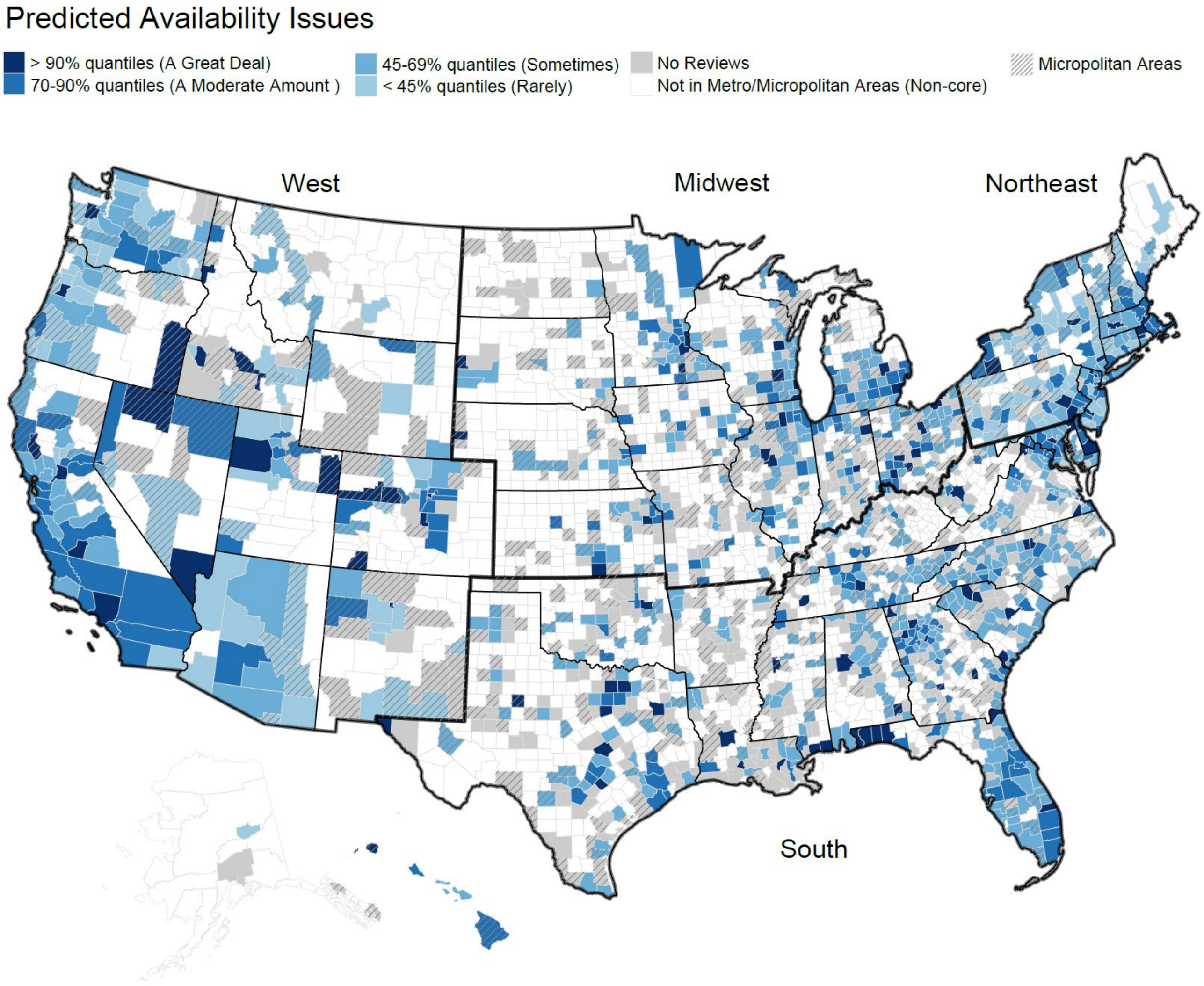


Figure 2. Predicted discussion frequency of station availability for U.S. metropolitan and micropolitan statistical areas.

The map reveals areas with high and low predicted *Availability* consumer discussions 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. Thus, algorithms predict that many micropolitan statistical areas could be under-served with regards to station availability.


Predicted Availability Issues
> 90% quantiles (A Great Deal)
70-90% quantiles (A Moderate Amount)
45-69% quantiles (Sometimes)
< 45% quantiles (Rarely)
No Reviews
Not in Metro/Micropolitan Areas (Non-core)
Micropolitan Areas
West
Midwest
Northeast
South

Tables

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

Table 2. Overall model performance

	Accuracy % (s.d.)	F1 score (s.d.)
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)

Note: Models are trained and tested on expert annotated data

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

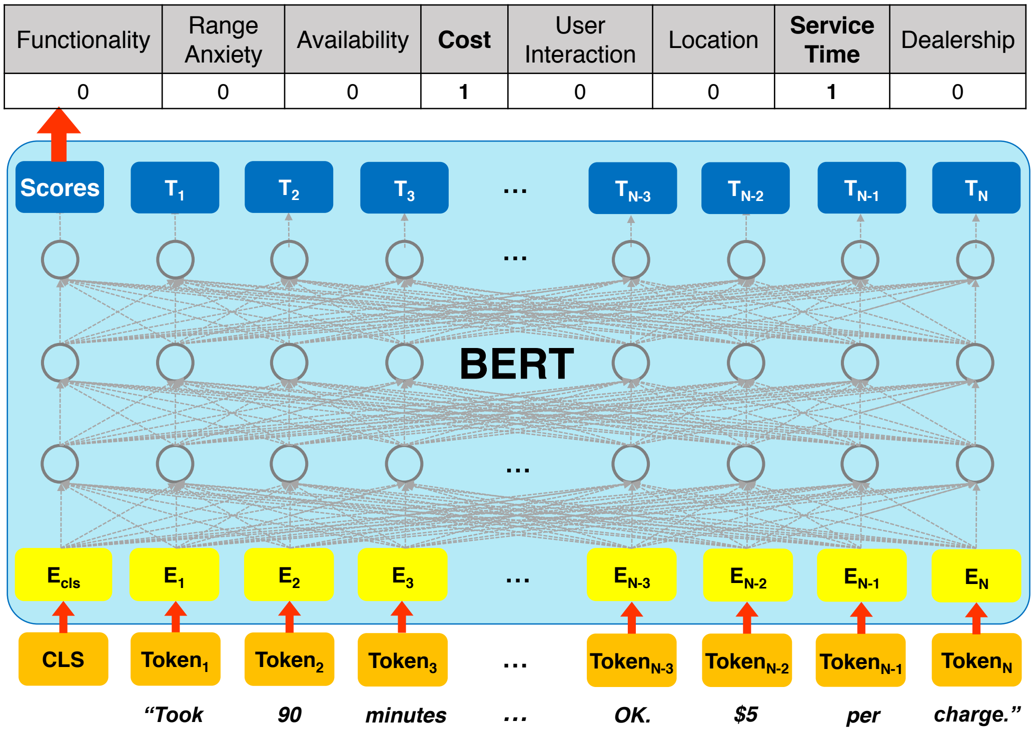
Classifier	Training set	Accuracy % (s.d.)	F1 score (s.d.)
BERT	Expert-annotated	89.1 (4.09)	0.82 (0.06)
BERT	Crowd-annotated	73.2 (3.85)	0.53 (0.06)
XLNet	Expert-annotated	91.0 (4.70)	0.85 (0.06)
XLNet	Crowd annotated	74.2 (4.15)	0.54 (0.07)
Crowd ($\kappa = 0.007$)	-	73.9 (6.06)	0.61 (0.09)
Human Experts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

Note: Cross validation = 10 runs

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

				Expert-trained Transformers			
	Ground Truth	Human Expert		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 minutes! 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

Supplemental Figures



Functionality	Range Anxiety	Availability	**Cost**	User Interaction	Location	**Service Time**	Dealership
0	0	0	**1**	0	0	**1**	0

Figure S1. BERT model architecture Representation of the BERT model architecture across the 8 topics of interest as a set of binary prediction outputs. For example, for the sample review shown *"Took 90 minutes . . . Ok. 5$ per charge"*, the model outputs 1 for *Cost, Service Time* topics, and 0 for the other topics.

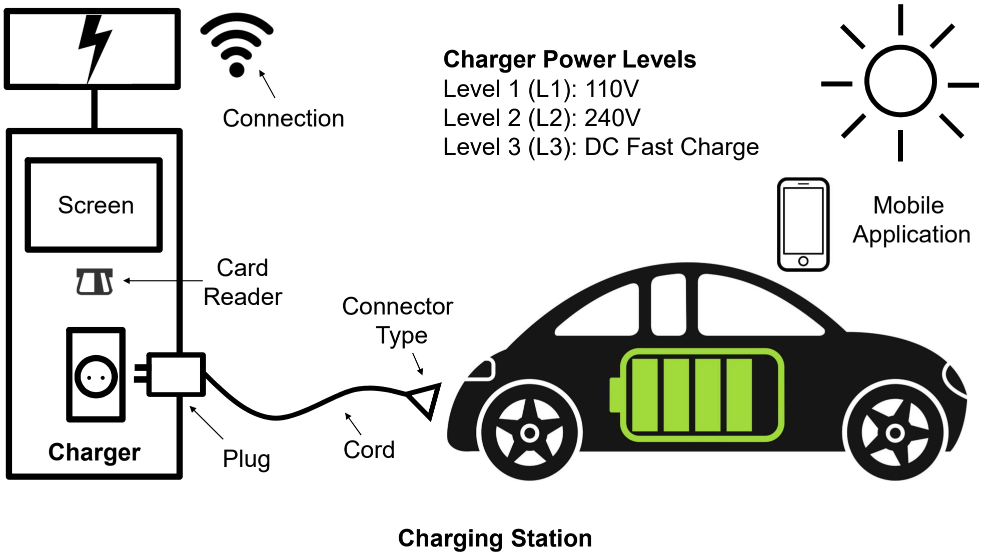
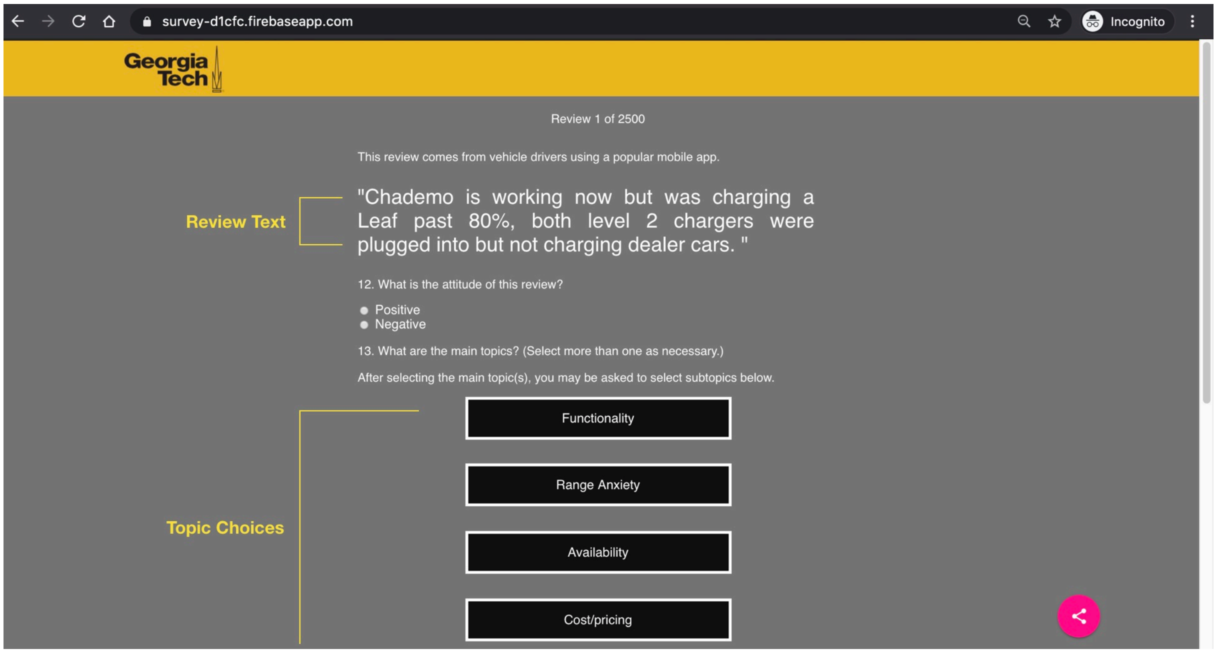


Figure S2. Diagram of EV charging station. Illustration of major EV charging components shown to the human annotators to help understand frequently occurring terms.



Review 1 of 2500

This review comes from vehicle drivers using a popular mobile app.

"Chademo is working now but was charging a Leaf past 80%, both level 2 chargers were plugged into but not charging dealer cars. "

12. What is the attitude of this review?

○ Positive
○ Negative

13. What are the main topics? (Select more than one as necessary.)

After selecting the main topic(s), you may be asked to select subtopics below.

Figure S3. Web App for training data collection A screenshot of the online interface for the human annotation.

Supplemental Tables

Table S1. Hyper-parameters for BERT and XLNet

Hyper-parameter	Value
Number of Epochs	20
Batch Size	8
Learning Rate	1e-4
Max Sequence Length	8
Weight Decay	0.01
Adam Epsilon	1e-8
Max Grad Norm	1
Warmup Steps	500
Train:Valid:Test	80:10:10

Table S2. Topic level accuracy

	Functionality		Range Anxiety		Availability		Cost		User Interaction		Location		Service Time		Dealership	
	Acc.	s.d.	Acc.	s.d.	Acc.	s.d.	Acc.	s.d.	Acc.	s.d.	Acc.	s.d.	Acc.	s.d.	Acc.	s.d.
BERT	86.0	0.445	96.7	0.259	91.5	0.414	96.8	0.260	89.4	0.460	85.2	0.448	97.1	0.334	91.2	0.341
XLNET	86.1	0.515	97.0	0.171	91.4	0.397	97.3	0.196	89.4	0.284	84.7	0.536	96.4	0.303	90.8	0.378
Majority Classifier	53.1		96.2		78.3		90.0		85.4		68.1		90.8		86.9	
LSTM	83.2	0.483	96.6	0.247	89.8	0.752	96.2	0.249	87.4	0.378	83.9	0.422	94.7	0.312	90.3	0.264
CNN	84.1	0.639	97.0	0.135	90.2	0.282	96.3	0.282	88.1	0.341	85.2	0.314	95.2	0.165	90.9	0.300

Table S3. Topic level F1 score

	Functionality		Range Anxiety		Availability		Cost		User Interaction		Location		Service Time		Dealership	
	F1	s.d.	F1	s.d.	F1	s.d.	F1	s.d.	F1	s.d.	F1	s.d.	F1	s.d.	F1	s.d.
BERT	0.860	0.005	0.707	0.028	0.869	0.008	0.912	0.007	0.754	0.011	0.824	0.006	0.909	0.010	0.807	0.010
XLNET	0.860	0.005	0.792	0.010	0.867	0.006	0.926	0.006	0.760	0.007	0.821	0.006	0.892	0.009	0.802	0.010
Majority Classifier	0.347		0.490		0.439		0.474		0.461		0.405		0.476		0.465	
LSTM	0.831	0.005	0.733	0.018	0.847	0.008	0.889	0.010	0.706	0.015	0.810	0.005	0.828	0.010	0.783	0.006
CNN	0.841	0.007	0.755	0.022	0.850	0.005	0.896	0.009	0.702	0.013	0.824	0.006	0.835	0.007	0.797	0.013

Table S4. **Computation times**

	Number of GPUs	Overall Computation Time†	Train Time per Epoch (seconds)	Test Time per Example (seconds)
CNN	1	00:00:56	2.8	2.7e-4
	4	00:00:50	2.5	2.6-e4
LSTM	1	01:25:38	257	3.0e-3
	4	00:57:42	173	2.2e-3
BERT	1	02:10:39	392	1.2e-2
	4	01:05:33	196	2.2e-2
XLNet	1	04:31:40	1,084	7e-2
	4	01:27:20	346	4e-2

†hours:minutes:seconds

Note: Computation times using PACE force-gpu cluster on 16GB memory.

Table S5. **Pairwise topic correlation**

	1	2	3	4	5	6	7	8
1. Functionality	1.000							
2. Range Anxiety	-0.027	1.000						
3. Availability	-0.256	-0.048	1.000					
4. Cost	-0.069	-0.003†	-0.053	1.000				
5. User Interactions	-0.177	-0.013†	-0.046	-0.018†	1.000			
6. Location	-0.218	0.048	-0.036	0.074	-0.066	1.000		
7. Service Time	-0.004	0.061	-0.073	0.055	-0.060	-0.022	1.000	
8. Dealership	-0.069	0.061	0.0525	-0.011†	0.068	0.011†	0.070	1.000

†Not significant (p >0.05)

Table S6. Accuracy conditional on changing the number of topics

Number of Categories	Functionality	Location	Availability	Cost	Dealership	User Interaction	Service Time	Range Anxiety
2	85.5	85.8						
3	85.5	84.8	91.5					
4	85.1	84.9	91.7	96.8				
5	86.0	85.4	91.2	97.1	91.7			
6	86.3	85.6	91.7	97.6	91.4	88.8		
7	85.6	84.7	91.3	97.1	91.2	88.4	96.5	
8	85.7	85.6	91.2	97.1	90.7	89.2	96.7	96.7
Average	85.7	85.3	91.4	97.1	91.3	88.8	96.6	96.7
Max Difference (%)	0.734	0.653	0.292	0.474	0.603	0.450	0.104	0.000

Table S7. F1 score conditional on changing the number of topics

Number of Categories	Functionality	Location	Availability	Cost	Dealership	User Interaction	Service Time	Range Anxiety
2	0.854	0.831						
3	0.854	0.820	0.872					
4	0.851	0.821	0.877	0.915				
5	0.860	0.826	0.866	0.923	0.825			
6	0.863	0.827	0.874	0.934	0.824	0.759		
7	0.856	0.815	0.868	0.919	0.816	0.742	0.897	
8	0.857	0.830	0.863	0.920	0.805	0.749	0.897	0.722
Average	0.856	0.824	0.870	0.922	0.818	0.750	0.897	0.722
Max Difference (%)	0.767	1.127	0.805	1.280	1.529	1.200	0.000	0.000

Supplemental Experimental Procedures

Robustness of typology

We assembled a pairwise correlation matrix to easily inspect whether the labels represent non-overlapping categories. In Table S5, we show that 25 out of 28 pairwise correlations are below 0.10 in absolute value, which indicates small correlation. In addition, none of the correlation coefficients are above 0.3 in absolute value, which indicates small to moderate correlation.

We also investigated the sensitivity of our models to having more or fewer topics. To do this, we computed the accuracy and F1 scores for a series of models trained with a varying number of topics. For example, in Table S6 and Table S7, we started with *Functionality* and successively added topics, starting from 2 topics and increasing up to 8. We report the accuracy measures in Table S6, and F1 scores in Table S7. For accuracy, we find that the maximum difference in performance varies by less than 1% for all topics, and for the F1 score, the maximum difference in performance varies by less than 2% for all topics. These additional results are well within the statistical uncertainty reported in our main results shown in Figure 1, Table S2, and S3. We therefore provide evidence that our technical approach is not very sensitive to having varying number of topics. We also evaluated impacts on the computing times, and found that the number of topics had limited impact on computing times (~1%).

BERT and XLNet implementations

BERT and XLNet are pre-trained contextual language models that leverage massive corpora such as the English Wikipedia and BooksCorpus to learn context from tokenized words [1]. These models leverage neural network architectures with information feeding in a bidirectional context. The language models are fine-tuned on our domain specific multi-label classification problem using training data.

To illustrate differences between BERT and XLNet in their ability to capture bidirectional context in our domain, we provide the following sample review, *"Fast charger working great!"* To understand the relational encoding, BERT and XLNet maximize the conditional probability of the word context in the forward and backward direction as follows:

$$\begin{aligned}\mathscr{L}_{BERT} = & \, logP(\text{Fast} \mid \text{working great! }) \\ & + logP(\text{Charger} \mid \text{working great!})\end{aligned} \tag{1}$$

$$\begin{aligned}\mathscr{L}_{XLNet} = & \, logP(\text{Fast} \mid \text{working great!}) \\ & + logP(\text{Fast} \mid \text{charger working great!})\end{aligned} \tag{2}$$

Here, $\mathscr{L}_{BERT}$ and $\mathscr{L}_{XLNet}$ refer to the log-likelihood functions for the two models. When comparing the equations 1 and 2, we see that the dependency between the tokens *Fast* and *Charger* in this example are learned as a relevant training signal in XLNet but not in BERT. For code implementation of BERT and XLNet, we followed the protocols in [2,3] as a starting point. Since there were no known references for optimal hyper-parameters for BERT or XLNet in this domain, we report our hyper-parameter values in Table S1, which we arrived at through minimal fine-tuning. We did not do an exhaustive hyper-parameter search. This further optimization could be done in future work. For seminal readings on BERT and XLNet, see [1,4,5].

CNN and LSTM implementations

The baseline models used for comparison with the transformer models are convolutional neural networks (CNN) [6,7] and long short-term memory (LSTM) classifiers [8]. Architecturally, while CNNs build feature representations of a sentence through convolution with filters of varying sizes [6], LSTMs encode hidden state representations via a recurrent neural network [8] which is updated by traversing the sentence in one direction. Although currently there is no consensus on which models are better for text classification tasks, CNNs and LSTMs provide complementary information. CNNs are hierarchical architectures, while LSTMs are sequential architectures, which tend to perform better in sequence modeling tasks. In this paper, we adapted code and protocols from [7,9] for CNN implementation and [10] for LSTM implementation. For a comparative review of CNNs and LSTMs in natural language processing, see [11].

Detecting if a certain station is functioning

To get an initial idea of how the method performs to detect if a certain station is functioning, we calculated the conditional probability of jointly detecting the *Functionality* topic and a negative sentiment in the review (e.g. the qualifying event).

To do this, we sampled reviews from charging stations with both high number of repeat check-ins and a low number of repeat check-ins in order to get a range of estimates across different station types. For this simulation, we assume that the joint probability of detecting a functionality topic and positive sentiment (e.g. "This station is working great!") is not a qualifying event. We provide an illustrative example below. To derive the negative sentiment probabilities, we used published numbers from [12] that uses EV charging reviews data from a similar date range. For example, for highly used stations in the 90th percentile by number of reviews, the negative sentiment probability is 0.495. Likewise, for less commonly used stations in the 25th percentile, the negative sentiment probability is 0.390. Next we calculated the prediction probabilities for the *Functionality* topic for these two groups of stations as 0.574 and 0.451 for the 90th and 25th percentile by number of reviews, respectively. The joint probability of a qualifying event, e.g. if a certain station is functioning and negative sentiment, gives us a range of 0.176 to 0.284.

This indicates that for every 100 reviews, we expect between 17 and 28 qualifying events on whether if a certain station is not functioning. In other words, this turns out to be one qualifying event every 3 to 5 reviews in this dataset. For example, a large CBSA such as San Jose-Sunnyvale-Santa Clara, CA, received 6,703 reviews between Aug 2011 and September 2015. This is approximately 4.5 reviews per day. Consequently, for large-scale implementation, the model will detect a qualifying event typically every day. On the other hand, a small CBSA such as Chattanooga, TN-GA, received 2,132 reviews between December 2011 and September 2015. This is approximately 1.5 reviews per day, which means that the model will detect a qualifying event typically every 2 to 3 days for a small CBSA. On a national basis, this means that our model would typically detect if a certain station is functioning, daily. Given the exponential growth of EV infrastructure data and usage, we expect this detection rate to get even better over time.

Software and resources

The deep learning algorithms used in this paper were written in Python, using PyTorch for BERT and XLNet; and TensorFlow for CNN and LSTM. Experiments for Table S4 were run on the PACE Force cluster using the NVIDIA Tesla P100 GPUs. The experiments for Table 2, S2, S3, S6, and S7 were run on Microsoft Azure Cloud, using the same NVIDIA Tesla P100 GPUs. We replicated these results across both high-performance computing clusters to within the statistical uncertainty reported.

Supplemental References

1. Devlin, J., Chang, M.-W., Lee, K., Toutanova, K. (2019). Bert: Pre-training of deep bidirectional transformers for language understanding. In North American Chapter of the Association for Computational Linguistics (NAACL-HLT), pp. 4171–4186.
2. Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., Cistac, P., Rault, T., Louf, R., Funtowicz, M., et al. (2020). Transformers: State-of-the-art natural language processing. In Empirical Methods in Natural Language Processing (EMNLP), pp. 38–45.
3. Marchetto, D. J., Ha, S., Dharur, S., Asensio, O. I. (2020) Extracting user behavior at electric vehicle charging stations with transformer deep learning models. In International Conference on Advanced Research Methods and Analytics (CARMA). https://dx.doi.org/10.4995/CARMA2020.2020.11613.
4. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., Polosukhin, I. (2017). Attention is all you need. In Advances in Neural Information Processing Systems (NeurIPS), pp. 5998–6008.
5. 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 (NeurIPS), pp. 5753–5763.
6. LeCun, Y., Bengio, Y. (1995). Convolutional networks for images, speech, and time series. In The Handbook of Brain Theory and Neural Networks, 3361.
7. Kim, Y. (2014). Convolutional neural networks for sentence classification. In Empirical Methods in Natural Language Processing (EMNLP), pp. 1746–1751.
8. Hochreiter, S., Schmidhuber, J. (1997). Long short-term memory. Neural Comput. 9, 1735–1780.
9. Zhang, Y., Wallace, B.C. (2017). A sensitivity analysis of (and practitioners' guide to) convolutional neural networks for sentence classification. In International Joint Conference on Natural Language Processing (IJCNLP), pp. 253–263.
10. Karpathy, A., Johnson, J., Fei-Fei, L. (2015) Visualizing and understanding recurrent networks. arXiv, 1506.02078.
11. Yin, Y., Kann, K., Yu, M., Schütze, H. (2017). Comparative study of cnn and rnn for natural language processing. arXiv, 1702.01923.
12. 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. Nat. Sustain. 3, 463–471. https://doi.org/10. 1038/ s41893-020-0533-6.