DETECTING BEHAVIORAL FAILURES IN EMERGING ELECTRIC VEHICLE INFRASTRUCTURE USING SUPERVISED TEXT CLASSIFICATION ALGORITHMS

Sooji Ha
School of Civil & Environmental Engineering and School of Computational Science & Engineering
Georgia Institute of Technology, Atlanta, GA, 30332
Email: s.ha@gatech.edu

Daniel J. Marchetto
School of Public Policy
Georgia Institute of Technology, Atlanta, GA, 30332
Email: daniel.marchetto@gatech.edu

Mary Elizabeth Burke
School of Public Policy
Georgia Institute of Technology, Atlanta, GA, 30332
Email: mburke38@gatech.edu

Omar Isaac Asensio*
School of Public Policy and Institute for Data Engineering & Science (IDEaS)
Georgia Institute of Technology, Atlanta, GA, 30332
Email: asensio@gatech.edu
ORCID Number: 0000-0003-2143-5022
* Corresponding Author

ABSTRACT

There is a growing interest in applying computational tools to the automatic discovery of social and economic behavior. For example, with decisions involving resource allocation related to public infrastructure, the ability to predict failures can allow for more efficient policy responses. In this paper, we use social data from a popular electric vehicle (EV) driver app to characterize the emerging EV charging station infrastructure. We introduce a typology of EV charging experiences collected from user reviews and deploy text classification algorithms, including convolutional neural networks (CNN), to automatically learn about potential failures. We use machine learning techniques as a pre-processing tool for econometric analyses on the quality of service delivery. After classifying the reviews into 9 main user topics and 34 subtopics, we find that the dominant issues in EV charging relate to station functionality and availability, which drive negative consumer experience. Contrary to the public discourse about EVs, range anxiety was not of large concern to existing EV drivers. Based on our findings, we move towards automated identification of failures in public charging infrastructure that can significantly reduce research evaluation costs through relatively simple computational solutions.

Keywords: electric vehicles, consumer behavior, convolutional neural networks, natural language processing, mobile data

INTRODUCTION

The transportation sector has become a dominant source of CO_2 emissions in the United States (*1*). In the last few years, there has been a growing attention on vehicle electrification as a strategy to reduce mobile source emissions with positive spillovers in air quality benefits (*2*). For example, in order to accelerate electric vehicle (EV) purchases, a majority of U.S. states are offering some type of financial incentive to complement federal tax credits, including rebates, tax exemptions and other incentives (*3–5*). An important complementarity to EV adoption is the availability of public charging infrastructure. It is now estimated that global investment in EV charging infrastructure from public and private sources will reach $80 billion USD by 2025 (*6*). In the US, this investment growth also marks an expected transition in policy support to a focus on charging infrastructure.

However, currently there is no easy way to evaluate the needs of drivers or assess performance. This is because the infrastructure upgrades needed to be able to monitor electric consumption and use in individual charging stations are at early stages of development or not available. Further, the large-scale data required to evaluate system performance cannot easily be aggregated across charging networks. Given the rise in real-time streaming data in transportation and mobility apps, much of the useful intelligence about charging infrastructure performance lies highly unstructured. Consumer reviews, for example, can be collected instantly from thousands of users, and manually processing or analyzing this unstructured data to collect useful information has not been possible. We argue that real-time, streaming data will be increasingly important for research evaluation of sustainable infrastructure. For example, since the release of a popular charging station locator app, there have been over 1.5 million user reviews of charging stations lying dormant as text (*7*). Given this volume of data, even at an expert processing rate of 120 reviews per hour, it will be prohibitively costly for humans to classify this unstructured text data for research evaluation.

We have shown in prior research that consumer sentiment can be automatically processed with high accuracy through computational aid (*8*). From this analysis, the evidence suggests that there is a significant amount of negative consumer sentiment related to the charging experience. While we demonstrated state-of-the-art performance from neural network-based models in this domain, learning about the sources of negative consumer experience, which is needed to conduct policy analysis, remains a challenging task due to the complexity of natural language processing (NLP). In this paper, we therefore introduce a computational solution to analyze the content of real-time text data as tailored to the domain of EVs and consumer behavior. Because it is known that consumer reviews may be subject to self-selection and other observational biases, we use machine learning as a pre-processing tool to conduct econometric analyses for statistical adjustment. We use this approach to automatically learn about large-scale barriers to EV infrastructure use nationally.

Contrary to the public discourse, we find that range anxiety is not a major concern among existing EV drivers. Instead, our results suggest that the major issues facing EV drivers relate to station functionality and availability—an insight that we make possible through large-scale data integration.

MACHINE LEARNING AND TEXT CLASSIFICATION

With the increasing popularity of social data from digital platforms, user-generated short texts have become an important data source for NLP. In the transportation domain regarding EV adoption, there are as yet few research studies that translate the unstructured data from user generated texts into actionable intelligence. One exception is a recent paper by Kuhl et al. (2019) in which the authors manually coded Twitter data and found that contrary to what has been the focus in the

literature, charging infrastructure was the most discussed topic (*9*). There is no definitive study that does large-scale analysis of EV behavior for policy analysis. Further, the implementation of recent advances, such as deep neural networks (DNNs), which have revolutionized the field of natural language processing (*10*) have not been implemented for transportation policy. There are two prevailing types of DNNs. The first are convolutional neural networks (CNNs). CNNs extract the most meaningful information from text data by decomposing the hierarchical structure of sentences or phrases (*11*). The other prevailing DNN architecture is that of recurrent neural networks (RNNs) (*12*). RNNs have the added benefit of flexibility in analyzing a sequence of text. For instance, a recent study by Ma et al., (2019) collected Chinese consumers' online comments about electric vehicles and processed the data with a variant of an RNN architecture, known as the long short-term memory (LSTM) in order to review EV purchase preferences such as retail prices, and EV makes and models (*13*). However, the authors do not report their classifier performance measures in their text mining analysis (e.g., accuracy and F1 score), so it is not possible to meaningfully access the relative merits of DNNs in this domain.

Choosing between CNN or RNN architectures for any type of data is currently an ongoing debate in the literature (*10*). In our prior work, we initially demonstrated that a CNN model produced state-of-the-art accuracy with good balance measures for sentiment classification tasks, which weakly dominated an RNN model (*14*). From the consumer analysis of charging station sentiment, we were also able to demonstrate that there were differences in station-level sentiment when looking at the geographical regions (e.g. urban, rural, etc.) even after controlling for observable station characteristics. However, sentiment analysis, although informative about quality perceptions, does not give us a window into the specific mechanisms or sources of the negative sentiment. As such, in this contribution, we are interested in applying innovations in NLP to introduce a typology of charging behavior and to provide insights on the use of computational tools for the automatic detection and discovery of barriers to infrastructure management.

DATA AND METHODS

We have a nationally representative sample of unstructured consumer reviews at 12,720 US charging station locations as provided by a popular EV charge station locator app. The text data consists of 127,257 reviews written in English from 29,532 registered and unregistered EV drivers during the period from 2011 to 2015. The sample represents charging stations from the entire U.S. market during the period of study. This includes data aggregated from 10 major EV charging networks in the US. In the sample, we also geocoded point of interest (POI) location information using Google places API for categories such as Dealerships, Government, Healthcare, Hotel/Lodging, Other, Park, Parking Garage/Lot, Residential, Restaurant, School/University, Shopping Center, Store/Retail, and Workplace. For more information, ref. (*8*).

EV Charging Infrastructure Consumer Reviews

The charging station reviews can be considered social interactions within the community of EV drivers. After analyzing the contents of over 8,000 reviews, two research assistants were able to identify the main categories discussed by users and were able to determine which issues are most prominent regarding the charging experiences of this community. In preliminary experiments, we investigated several unsupervised topic modeling techniques that did not provide theoretically meaningful clusters. Therefore, we took the approach of hard coding labels based on human intelligence. We introduce 9 main categories and 34 subcategories that make up a typology of charging

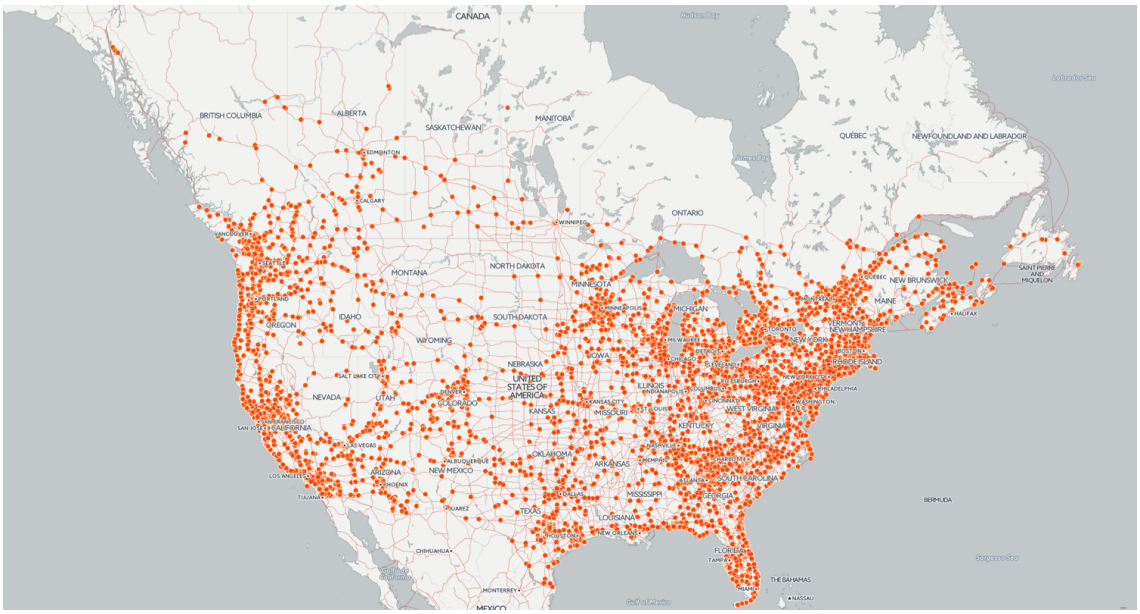

FIGURE 1: Map of Charging Stations in North America

behavior that allows for easier identification and eradication of inefficiencies of the charging process. The typology is provided in Table 1. *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 station. *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* category refers to comments reporting charging rates (e.g. 10 miles of range per hour charged) experienced in a charging session. The *Dealerships* category refers to comments concerning specific dealerships and user's associated charging experiences. The *Other* category refers to comments that do not fall into the previous eight categories. From our sample of human labeled reviews, the *Other* category occurs 6.0% of the time. For the full frequency counts by label, see Table 2.

Approach to Curating the Training Data

Classification techniques employed by many scholars often assume that observational data is a random sample from a given distribution that is believed to be representative of the population. However, well-known biases in learning and evaluating classifiers can include researcher bias, sample selection bias, and other statistical sampling issues (*15–17*).

For this reason, we actively curated a population of human annotators that were prescreened to be representative of the US general population (age 18+). With the support of a Qualtrics panel, a sample of 1,000 participants were recruited with nationally representative demographic characteristics such as age, income and education level, sex and ethnicity. This allowed us to explore techniques to mitigate, although not completely eliminate, potential individual bias by

TABLE 1: EV Mobile App Typology of User Reviews

Category	Subcategory
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

eliciting the wisdom of crowds and reducing the potential impact of conflicting biases.

We deployed an online survey questionnaire from Nov 27 to Dec 11, 2018 to build a training dataset for supervised machine classification. Each participant was given a total of 20 charging station reviews and was tasked with labeling the sentiment of the review (positive or negative), as well as selecting categories and sub-categories that applied to the review. Of the 20 charging station reviews labeled by each participant, 5 of them were sample of 830 reviews designed for reliability checks, while the remaining 15 reviews were randomly selected from the superpopulation. The sample 830 reviews for the reliability checks were randomly drawn from a set that had been previously labeled by research assistants. The holdout samples were designed to be distributed to at least three of the 1,000 recruited participants on average for the purpose of calculating inter-rater agreement. All reviews that include potentially sensitive information such as cell phone numbers or email addresses were redacted. Because subcategories of many categories such as *Functionality* had domain-specific terminology that could be unfamiliar to a general population, we also provided a diagram of a charging station with information along with tips to help guide the human classification. A view of the online survey interface is shown in Figure 2. After the labeling tasks were completed, we also asked participants for voluntary demographic information. No personally identifiable information was collected or shared (IRB protocol No. H18250).

Inter-Rater Reliability

Because we had multiple annotators, we used Fleiss' Kappa (κ) as a measure of agreement between raters. Fleiss' Kappa was selected because it offers the benefit of a single metric to assess agreement between n-raters (*18*). On average, we had an average of 3 raters per review, ranging from 1 to 7 in the experiment. The Fleiss' Kappa, κ is calculated as below:

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

where $\bar{P}$ is the average number of agreement on all category assignments between rater pairs for the questions, and $\bar{P}_e$ is the average proportion of assignment to the categories. For example,

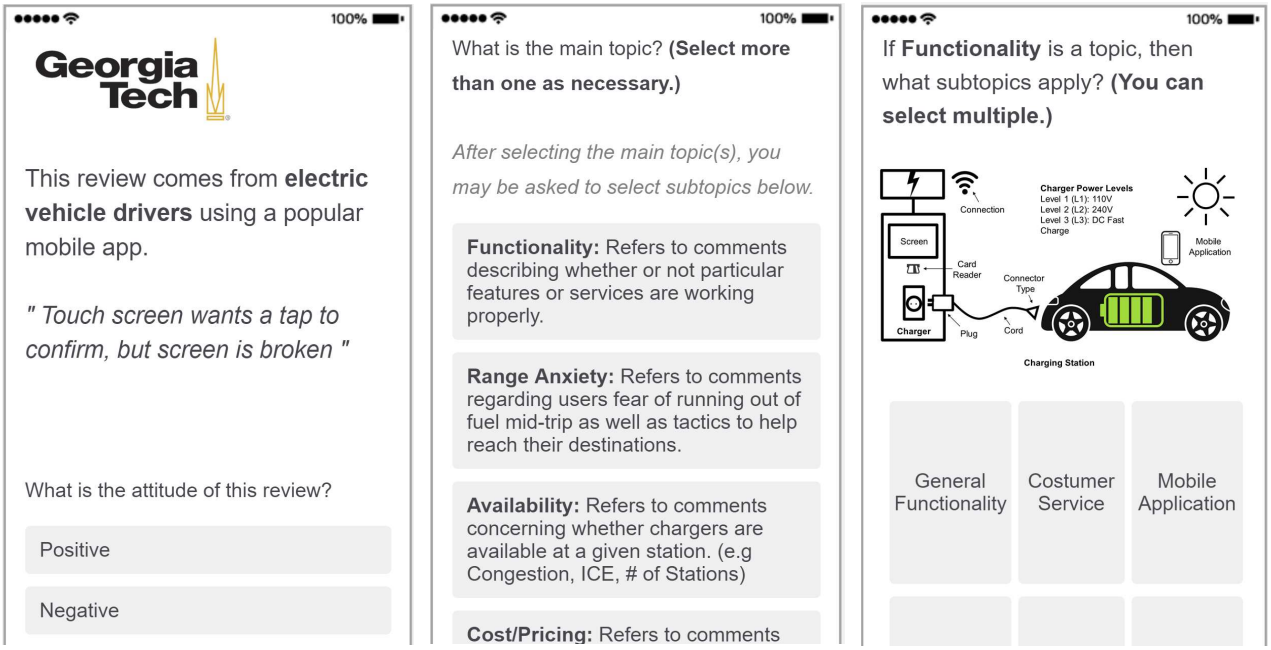

FIGURE 2: Online Survey Interface

assume there are 3 raters labeling 3 questions with 3 categories each. Suppose that the 3 raters agree on 2 questions with distinct categories and completely disagree on 1 question, choosing non-overlapping categories. In this case, $\bar{P} = \frac{2}{3}$, and $\bar{P}_e = \frac{11}{27}$, resulting in a κ score of 0.44. 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 above what would be expected at random (*19*). From the numerical example above, we would interpret the κ as moderate agreement. For more information, see ref. (*18*). We provide an analysis of inter-rater reliability in the Results and Discussion section.

CNN Implementation

In this study, we expand on CNN implementation protocols introduced in ref. (*14*), which is part of a larger literature of consumer analysis using DNNs (*20–22*). Given that many labels can occur relatively infrequently leading to unbalanced class data for training and testing, a simple majority classifier that assigns the majority class to all test data is provided as a baseline for performance. The performance of the CNN classifier is also compared to other commonly used baseline classification models, e.g. the bag-of-N-grams based logistic regression (LR) and support vector machines (SVM).

Convolutional Neural Network for Text Classification

Here we provide a high level overview of a single layer CNN model for binary topic classification. Figure 3 shows a generalized CNN architecture, which is comprised of three main parts: the vector representation of words, the convolutional layer, and the fully connected layer that yields a binary category prediction as the output. First, all words in a consumer review are changed to tokens, each having a certain dimensional vector composed of numeric values. These values are called word embeddings, which represent similarity between vocabularies quantitatively; similar words

have similar values across their word vectors and vice versa. In this study, we used pre-trained, publicly available word vectors, namely the *word2vec* (*23*), which is a predictive model trained on 100 billion words and phrases in Google News. We also tried other open-sourced, pre-trained word vectors, *GloVe* (*24*), a count-based model which gave comparable results. The use of pre-trained word vectors in text classification task is a widely used method to improve performance in the absence of a large corpus of domain-specific training data (*20, 25*). In the convolutional layer, filters scan through the word vectors creating what are referred to as "feature maps" using activation functions that are representative of the word vector of the chosen filter size. This process is done empirically by multiple filters, and the results are combined and extracted in a process called "pooling." Extracted information by pooling process is gathered as one feature vector, and it creates prediction through the fully connected layer. A CNN model can have various hyperparameters for the process of filtering, convolution, and pooling processes. For more information about details of the algorithm, see refs, (*11, 20*).

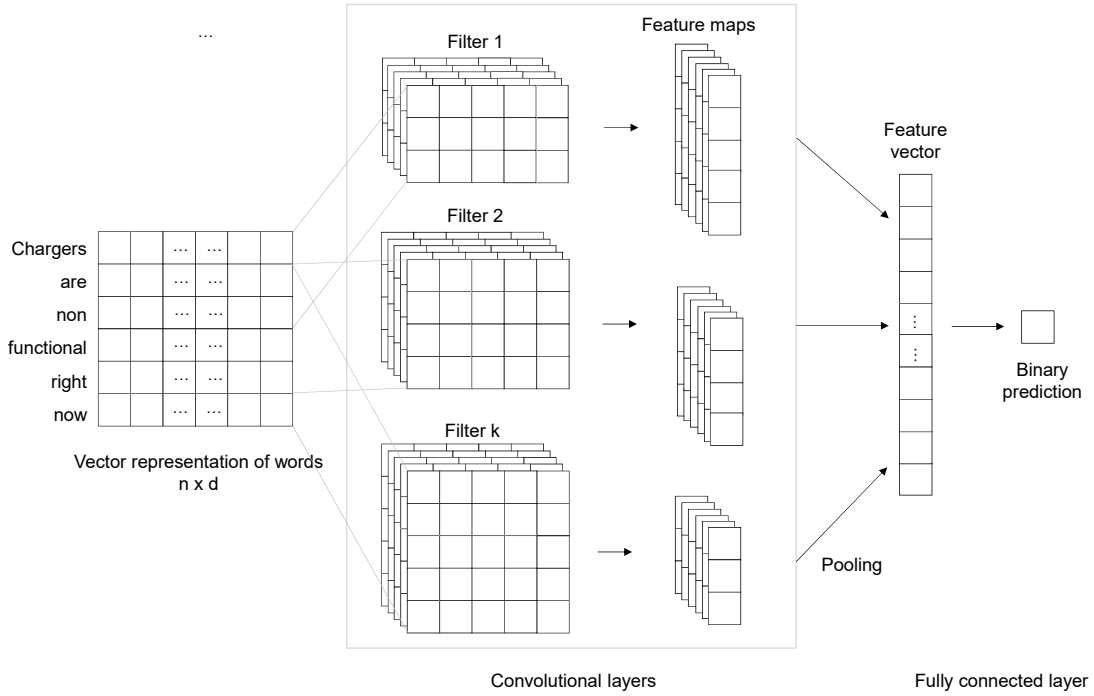

FIGURE 3: Architecture of the Convolutional Neural Network Model

Training and Testing Data

We created nine binary classifiers for each of the main categories: *Functionality, Range Anxiety, Availability, Cost, User Interactions, Location, Service Time, Dealership* and *Other.* For the training data, we used the 20,000 reviews labeled by the 1,000 participants. The models are validated by a set 5,229 true labels provided by research assistants as human experts. The counts of labels provided by participants are summarized in the Table 2. As the categories labeled in the training set are largely imbalanced, it is important to evaluate balance measures in classifier performance and to verify the level of learning compared with the simple majority class model. Based on the most common labels, the most important topics to consumers are *Functionality* and *Availability*.

TABLE 2: Counts of Labeled Reviews per Category in Training Data

	Labeled	Not Labeled	Percent Labeled[†] (%)
Functionality	7,370	12,630	36.9
Range Anxiety	1,809	18,191	9.0
Availability	4,968	15,032	24.8
Cost	2,006	17,994	10.0
User Interactions	3,174	16,826	15.9
Location	3,046	16,954	15.2
Service Time	1,805	18,195	9.0
Dealership	1,075	18,925	5.4
Other	1,209	18,791	6.0

[†] Reviews can have multiple labels therefore Percent Labeled does not sum to 100%.

Hyperparameter Optimization

We perform basic hyperparameter tuning, beginning with suggested values introduced in previous studies (*14, 20, 26*). These include filter region sizes of [3, 4, 5]; 100 filters; use of rectified linear unit (ReLU) activation function; learning rate of 0.001; dropout rate of 0.3 for regularization, and no l_2 norm constraint. We also followed guidelines provided by (*26*) that each dataset has its own optimal filter region size and numbers, therefore perfromed basic grid search to find hyperparameters that could yield higher performance in our specific dataset. For the *Functionality* category, we found that filter region sizes of [12, 12, 12], 400 number of filters, learning rate 0.0001, 0.6 dropout rate, 3 epochs with 128 batch size resulted in the highest performance within our search. However, with other categories, the performance dropped as changes occurred to our initial hyperparameters.

Outcomes of Interest

We are interested in evaluating the factors that predict the performance of outcomes across stations. Given our objective of detecting behavioral failures, we focus our analysis on *Functionality* and *Availability*, which are the 2 most frequently observed labels in the training data. We created an index that measures the probability that a label is likely to be chosen. For a given station review i, at location group g, in *year*, the Label Score is defined as follows:

$$\text{Label Score}_{m,i,g,year} = \frac{\text{Count of label reviews}_{m,i,g,year}}{\text{Total count of reviews}_{i,g,year}} \tag{2}$$

where m is the label of interest. A score near 0 would indicate that the specific label has a low incidence at that location group in that year, while a high score near 1 would indicate a high incidence at the location group in that year were assigned that label by the classifier.

Econometric Estimation using Fractional Response Models

The most commonly used implementation for models with fractional dependent variables only requires some specification of the correct functional form of the fractional dependent variable (*27*). In our case, the variables of interest are fractional vectors bounded between 0 and 1 with probability masses accumulating at the boundaries of the interval (See Figure 4). It would be

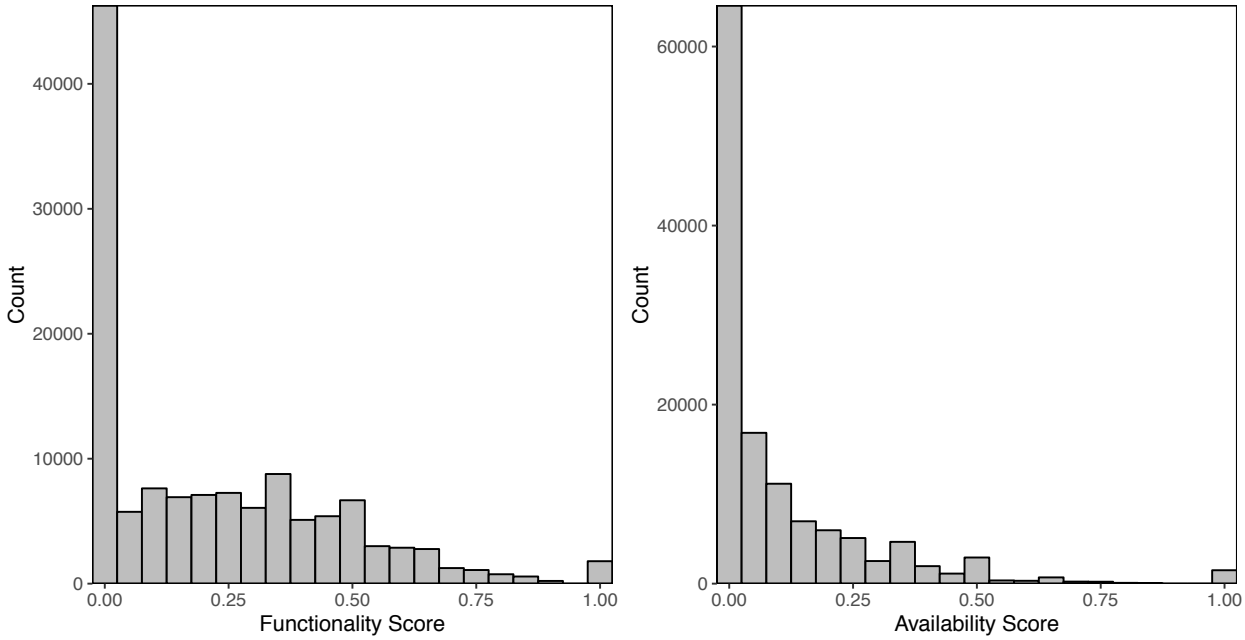

FIGURE 4: Histograms of Label Scores

inappropriate to use an OLS or log-odds estimator, for reasons we have described elsewhere (*14*). In the general form for the fractional response model, the interest is on the conditional expectation of the fractional response variable $y_{i,t}$ on the vector of explanatory variables $\mathbf{x}_{i,t}$ such that,

$$E(y_{i,t}|x_{i,t}) = G(\mathbf{x}_{i,t}\theta),\ i = 1,\ldots,N, \tag{3}$$

where $G(\cdot)$ is a non-linear transform function where the cdf satsifies $0 \leq G(\cdot) \leq 1$, the fractional dependent variable is defined only on $0 \leq y_{i,t} \leq 1$, and θ is a parameter vector of interest. Estimates of the effects are directly computed using the Bernoulli log-likelihood function given by (*28, 29*),

$$LL_{i,t}(\theta) \equiv y_{i,t} log[G(\mathbf{x}_{i,t}\theta)] + (1 - y_{i,t}) log[1 - G(\mathbf{x}_{i,t}\theta)]. \tag{4}$$

Given the presence of boundary observations of 0 in the fractional dependent variables of our dataset, the pooled Bernoulli quasi-maximum likelihood estimator (QLME) of θ can be computed as,

$$\hat{\theta} = \underset{\theta}{\operatorname{argmax}} \sum_{i=1}^{N} LL_{i,t}(\theta). \tag{5}$$

For our analysis, the main functional specification regresses the outcomes of interest on a vector of dependent variables, including geographical area dummies, station characteristics, POI dummies, Negativity Score, and related interactions. For the geographical area dummy variables we utilize the Census definitions for urban centers, urban clusters, and rural areas. From the Census definitions, urban centers are urbanized areas with more than 50,000 people, urban clusters are

urbanized ares with populations between 2,500 and 50,000 people, and rural areas are any such areas outside of urbanized ones (*30*). The station characteristics include measures of the number of connectors at a station, the number of networks at a station, and a proprietary station quality rating that ranges between 1 and 5. The general functional specification is as follows:

$$\begin{aligned}\text{Label Score}_{m,i,g,year} = \alpha_{i,year} + \text{Geographical Area Dummies}_{g} + \text{Station Characteristics}_{i,g,year} \\ + \text{Negativity Score}_{i,g,year} + \text{Interaction Effects}_{i,g,year} + \text{POI Dummies}_{g} \end{aligned} \quad (6)$$

RESULTS AND DISCUSSION

Descriptive Analysis of the Training Data

We performed a basic analysis to learn about the characteristics of the collected data. We provide a frequency distribution of the occurrences of all hard-coded labels in Figure 5 for main categories and in Figure 6 the most frequently occurring subcategories for our top category. The most selected main categories were *Functionality* and *Availability*, with counts of 7,370 and 4,968 respectively. We also learned that *Range Anxiety* is not a dominant theme among EV drivers' discussions (only 9% of the training data). We also break out the *Functionality* subcategories which were most discussed. For example, many consumers report issues with the proper functioning of the charger itself, station features, customer service, wireless connection, power issues and list of other problems, see Figure 6.

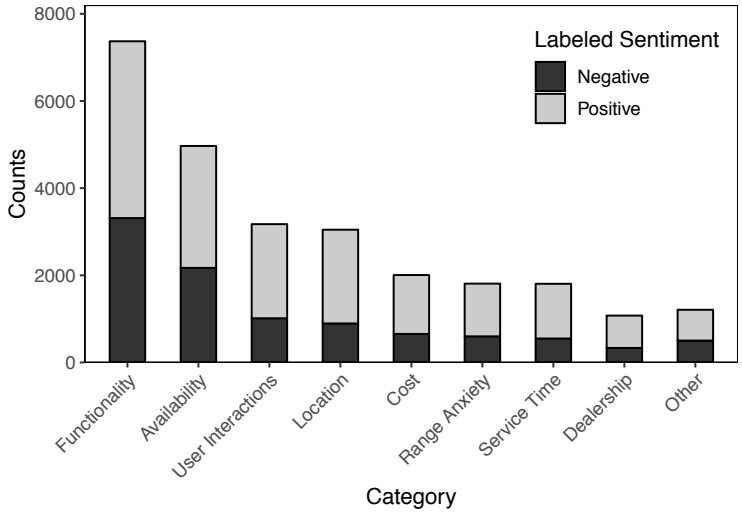

FIGURE 5: Frequency Count of Training Data per Main Category

Using the holdout sample labels collected from the training data, the Fleiss' κ was measured to be $\kappa = 0.31$, which is considered "fair agreement" in prior literature (*19, 31*). Although we note that there are relatively few existing studies that experimentally curate the training data, or report their reliability measures, we found some recent studies that begin to report reliability measure from human annotators using social data from posts in online forms (*32*). Further details on the effects of inter-rater reliability on classifier performance are discussed in the next section.

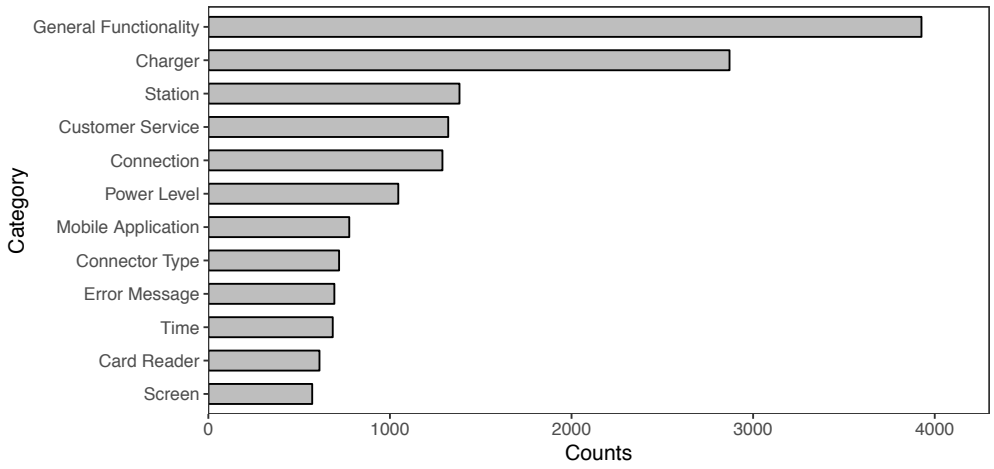

FIGURE 6: Frequency Counts per Subcategory of Functionality.

Classification Results

We are interested to learn whether a neural net-based model can accomplish the task of detecting a performance category from text with good accuracy compared to humans, while still achieving good balanced measures, using the F1 score as an indicator. For the analysis of classifiers we report model metrics for our top 3 categories which are *Functionality, Cost*, and *Availability*. Table 3 shows the comparison of CNN versus the other baseline models.

TABLE 3: Classification Results for Top Categories

	Accuracy (%)				F1 Score			
	CNN	SVM	LR	Majority Class	**CNN**	SVM	LR	Training Data Class Balance
Functionality	**78.1**	73.3	74.5	54.7	**0.72**	0.66	0.66	Balanced
Cost	**94.1**	93.3	92.9	90.6	**0.59**	0.57	0.44	Highly Imbalanced
Availability	**85.2**	83.2	0.85	88.7	**0.50**	0.37	0.46	Moderately Imbalanced

Despite the fact that we have imbalanced training data, we see that the CNN model outperformed the three baseline models for most categories, and is significantly better at detecting true positives and true negatives as shown by the F1 score. For example, one review states *"Right not working. Terrible"*. Our human annotators identified this review as a *Functionality* category. For the machine classification, LR and SVM failed to label *Functionality*, while CNN learned from semantic context to properly identify the label. In the opposite context of misclassification, CNN also performed better. For example, a consumer writes *"Stephen, you can use the Greenlots app if you don't have a fob."*. LR and SVM both predicted this to be *Functionality* label, while both human annotators and the CNN classifier correctly identified this text as not relating to *Functionality*. These examples illustrate that the CNN model has strong potential to automatically learn about the major issues in this domain. The results we report in Table 3 represent state-of-the-art

classification performance known for this domain.

One exception to the strong performance of CNN is the *Availability* category, where the accuracy of CNN was slightly lower than the simple majority class algorithm. This may be due to the difference in labeling rules between participants and trained annotators, where in the training data, *Availability* was labeled for 24.8% of the training reviews, while the trained annotators were more strict and labeled *Availability* for only 11.3% of the testing reviews. This suggests that in future work, we could provide more clear instructions to general population of annotators about the definitions of the labels. We also found that the *Cost* category had better F1 measure as compared ton *Availability* with higher imbalance in the training data.

However this strong performance is not without limitations. It did not do well on the labels such as *Location, Service Time, User Interactions, Dealership, Range Anxiety*, and *Others*. Based on our survey questionnaire, our results suggest that many of these domain specific terms can be complex for the general US population. To investigate this further, we calculated the Fleiss' κ of 0.31, which indicates that the inter-rater reliability score can be significantly enhanced. For this reason, in future work, we suggest curating crowd sourced human labels for machine learning using crowds of experts who might be more proficient in this domain. We leave that as future work. Further, we have implemented one neural net-based model, but it is not the only possible architecture. In addition to further tuning of the model we presented in this study, we also suggest exploring other deep learning architectures, particularly the recurrent neural networks which could have the benefit of learning from sequences of text data.

Having shown the good performance of our classier, we then use this model as a pre-processing tool to conduct econometric analyses in order to evaluate large-scale consumer issues in charging infrastructure. For more details on model performance, results are available upon request.

Fractional Response Model Results

We use the training data to classify the full population of 127,257 reviews in order to focus on our top two categories for the fractional response models (FRMs). Our main objectives are to uncover the main categories driving negative experience in the charging infrastructure in the United States. We evaluate geographical areas, both urban and non-urban, and how explanatory features are moderated by negative experience automatically classified using machine intelligence. The main results of the FRMs can be seen in Table 4.

Negative Charging Experiences with Functionality More Likely in Urban Center

In model (I) we estimate a basic specification that regresses the Functionality Score on observable station characteristics. We cluster our standard errors at the location group level. We find no evidence that geographical area is a significant predictor of functionality labels; however, in models (II) and (III) we investigate the sub-population of reviews with high Negativity Scores and find that reviews urban centers are 46.6% more likely to be about *Functionality*. For example, one user posted a review about *Functionality* in an urban center that state, "one of two chargers has fault error and doesn't work both charge a fee:(", while the same user posted a review outside of an urban center that stated, "120v outlet on lamp by entrance/sign very nice hotel excellent food next door", which is not about *Functionality*. This result answers a previously open question about the sources of negative consumer sentiment. By contrast, urban clusters and rural areas were not significant predictors of *Functionality* labels net of all controls.

The models used were robust to the inclusion of the quality rating as well as to various clustering alternatives for standard errors. These additional results are available upon request.

Negative Charging Experiences with Availability More Likely in Urban Cluster

We also evaluated factors related to station availability. In model (IV), we estimate a basic specification that regresses the Availability Score on all observable station characteristics. We do find evidence that urban centers have more reviews relating to *Availability* topics as compared to rural areas. Our results show that the sources of negative consumer experiences related to *Availability* are primarily in the urban clusters with reviews being 39.6% more likely. By contrast we do not find statistically significant results in with negative consumer sentiment in urban center or rural areas. This is interesting because one would expect *Availability* issues to be an urban phenomenon related to congestion. However, our large-scale analysis points to smaller urban clusters could need additional resources to broaden the availability of charging stations.

TABLE 4: Fractional Response Model Results

	Functionality Score			**Availability Score**		
	(I)	*(II)*	*(III)*	*(IV)*	*(V)*	*(VI)*
Geographical Area						
Urban Center	-0.080	-0.273**	-0.280**	0.455***	0.457***	0.461***
	(0.094)	(0.132)	(0.132)	(0.091)	(0.091)	(0.091)
Urban Cluster	0.185	0.179	0.188	-0.243**	-0.399***	-0.426***
	(0.153)	(0.147)	(0.148)	(0.122)	(0.147)	(0.149)
Station Characteristics						
Number of Connectors	0.450***	0.451***	0.456***	-0.477***	-0.476***	-0.482***
	(0.025)	(0.025)	(0.025)	(0.037)	(0.037)	(0.037)
Number of Networks	-0.148*	-0.147*	-0.162*	0.179	0.180	0.177
	(0.085)	(0.085)	(0.088)	(0.154)	(0.154)	(0.156)
Quality Rating			-0.059***			0.075***
			(0.017)			(0.017)
Negativity Score	1.744***	1.339***	1.246***	0.813***	0.798***	0.901***
	(0.054)	(0.220)	(0.216)	(0.055)	(0.056)	(0.055)
Urban Center x Negativity Score		0.466**	0.474**			
		(0.227)	(0.224)			
Urban Cluster x Negativity Score					0.396**	0.433**
					(0.201)	(0.206)
Point of Interest Control Dummies	Yes	Yes	Yes	Yes	Yes	Yes
Clustered SE	Yes	Yes	Yes	Yes	Yes	Yes
Number of Observations	127,257	127,257	127,257	127,257	127,257	127,257
R2	0.154	0.155	0.158	0.053	0.053	0.056
Note:				*p<0.1;	**p<0.05;	***p<0.01

Among the observable station characteristics, it turns out the the number of connectors is a significant predictor of both functionality and availability topics. In the case of *Availability*, more connectors predicts less reviews with availability labels. With *Functionality*, more connectors predicts more reviews with functionality labels. This is intuitive as more connectors should help availability issues while also providing more chances for functional issues to occur at a charge station. Overall, sources of negative consumer sentiment appear to be an urban phenomenon with urban centers being related to negative sentiment in reviews about functionality and urban clusters

being related to negative sentiment in reviews about availability.

POLICY IMPLICATIONS

In this study, we have been able to demonstrate that advances in computational algorithms can be deployed at relatively low cost with promising performance measures. We also demonstrate that through large-scale data aggregation, it may be possible to build a framework that could capture consumer intelligence about the functioning of the infrastructure, in near-real-time. Such capabilities could revolutionize how we manage, evaluate, and invest in charging infrastructure for electrified transportation. Based on our results, we have three main policy recommendations. First, a necessary criteria for building frameworks for real-time analysis is data and information sharing. We suggest the expansion of policies that can allow for greater real-time data sharing regionally and between jurisdictions. Second, given the discovery of negative consumer experiences as an urban phenomena, we suggest strategies for local and regional government to push greater standards and investment to help ensure that the quality and reliability of the charging experience is core to policies for EV growth. Third, the discussion of range anxiety as consumer barrier appears to be overstated, whereas station functionality and charging availability at the point of use may be the more critical limitation. Real-time streaming data is already changing the nature of mobility decisions for consumers. With a trained model present, thousands of reviews can be processed and analyzed in matter of minutes, if not faster. This should yield significant cost reductions for infrastructure performance evaluation.

ACKNOWLEDGEMENTS

We thank the generous support of the National Science Foundation Award No. 1931980, the Anthony and Jeanne Pritzker Family Foundation, Microsoft Azure for Research (Award CRM:0518988), the Sustainable LA Grand Challenge, and the Ivan Allen College Dean's SGR-C Award. For valuable research assistance, we thank Cade Lawson and Soobin Oh. We thank Suzie Lee and Hannah Weirich at Qualtrics for participant support. For valuable comments and feedback, we thank Richard Fujimoto, Emily Grubert, Haesun Park and Iris Tien. This research was supported in part through research cyber-infrastructure resources and services provided by the Partnership for an Advanced Computing Environment (PACE) at the Georgia Institute of Technology, Atlanta, Georgia, USA.

REFERENCES

1. EPA, *Inventory of U.S. Greenhouse Gas Emissions and Sinks: 1990-2016*, 2018, document No. 430-R-18-003.
2. Council, N. R. et al., *Hidden costs of energy: unpriced consequences of energy production and use*. National Academies Press, 2010.
3. Department of Energy, *Electric Vehicles: Tax Credits and Other Incentives Database*, 2019, access date: 07/31/2019, `https://www.energy.gov/eere/electricvehicles/electric-vehicles-tax-credits-and-other-incentives`.
4. Carley, S., R. M. Krause, B. W. Lane, and J. D. Graham, Intent to purchase a plug-in electric vehicle: A survey of early impressions in large US cites. *Transportation Research Part D: Transport and Environment*, Vol. 18, 2013, pp. 39–45.
5. Sheldon, T. L., J. DeShazo, and R. T. Carson, Electric and plug-in hybrid vehicle demand: lessons for an emerging market. *Economic Inquiry*, Vol. 55, No. 2, 2017, pp. 695–713.
6. Navigant, *Market Data: EV Market Forecasts: Global Forecasts for Light Duty Plug-In Hybrid and Battery EV Sales and Populations: 2017-2016*, 2017.
7. Recargo, *PlugShare Key Features and Benefits*, 2019, access date: 07/31/2019, `https://recargo.com/plugshare.html`.
8. Alvarez, K., A. Dror, E. Wenzel, and O. I. Asensio, Evaluating Electric Vehicle User Mobility Data using Neural Network based Language Models. In *Proceedings of 98th Annual Meeting of the Transportation Research Board, ADC80 Standing Committee on Alternative Transportation Fuels and Technologies, Washington, D.C.*, 2019.
9. Kuhl, N., M. Goutier, A. Ensslen, and P. Jochem, Literature vs. Twitter: Empirical insights on customer needs in e-mobility. *Journal of Cleaner Production*, Vol. 213, 2019, pp. 508–520.
10. Yin, W., K. Kann, M. Yu, and H. Schütze, Comparative study of CNN and RNN for natural language processing. *arXiv preprint arXiv:1702.01923*, 2017.
11. LeCun, Y., L. Bottou, Y. Bengio, P. Haffner, et al., Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, Vol. 86, No. 11, 1998, pp. 2278–2324.
12. Elman, J. L., Finding structure in time. *Cognitive science*, Vol. 14, No. 2, 1990, pp. 179–211.
13. Ma, S. C., Y. Fan, J. F. Guo, J. H. Xu, and J. Zhu, Analysing online behaviour to determine Chinese consumers' preferences for electric vehicles. *Journal of Cleaner Production*, Vol. 229, 2019, pp. 244–255.
14. Asensio, O. I., K. Alvarez, A. Dror, E. Wenzel, C. Hollauer, and S. Ha, Evaluating popular sentiment of electric vehicle owners in the United States with real-time data from mobile platforms. *Working paper*, 2019.
15. Shepperd, M., D. Bowes, and T. Hall, Researcher bias: The use of machine learning in software defect prediction. *IEEE Transactions on Software Engineering*, Vol. 40, No. 6, 2014, pp. 603–616.
16. Zadrozny, B., Learning and evaluating classifiers under sample selection bias. In *Proceedings of the twenty-first international conference on Machine learning*, ACM, 2004, p. 114.
17. Dwork, C., V. Feldman, M. Hardt, T. Pitassi, O. Reingold, and A. L. Roth, Preserving statistical validity in adaptive data analysis. In *Proceedings of the forty-seventh annual ACM symposium on Theory of computing*, ACM, 2015, pp. 117–126.

18. Fleiss, J. L., Measuring nominal scale agreement among many raters. *Psychological bulletin*, Vol. 76, No. 5, 1971, p. 378.
19. Landis, J. R. and G. G. Koch, The Measurement of Observer Agreement for Categorical Data. *Biometrics*, Vol. 33, No. 1, 1977, pp. 159–174.
20. Kim, Y., Convolutional Neural Networks for Sentence Classification. *arXiv:1408.5882 [cs]*, 2014, arXiv: 1408.5882.
21. Kalchbrenner, N., E. Grefenstette, and P. Blunsom, A convolutional neural network for modelling sentences. *arXiv preprint arXiv:1404.2188*, 2014.
22. Johnson, R. and T. Zhang, Effective Use of Word Order for Text Categorization with Convolutional Neural Networks. *arXiv:1412.1058 [cs, stat]*, 2014, arXiv: 1412.1058.
23. Mikolov, T., I. Sutskever, K. Chen, G. S. Corrado, and J. Dean, Distributed Representations of Words and Phrases and their Compositionality. In *Advances in Neural Information Processing Systems 26* (C. J. C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K. Q. Weinberger, eds.), Curran Associates, Inc., 2013, pp. 3111–3119.
24. Pennington, J., R. Socher, and C. Manning, Glove: Global Vectors for Word Representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Association for Computational Linguistics, Doha, Qatar, 2014, pp. 1532–1543.
25. Iyyer, M., J. Boyd-Graber, L. Claudino, R. Socher, and H. Daumé III, A Neural Network for Factoid Question Answering over Paragraphs. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Association for Computational Linguistics, Doha, Qatar, 2014, pp. 633–644.
26. Zhang, Y. and B. Wallace, 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)*, Asian Federation of Natural Language Processing, Taipei, Taiwan, 2017, pp. 253–263.
27. Ramalho, E. A., J. J. Ramalho, and J. M. Murteira, Alternative estimating and testing empirical strategies for fractional regression models. *Journal of Economic Surveys*, Vol. 25, No. 1, 2011, pp. 19–68.
28. Papke, L. E. and J. M. Wooldridge, Econometric methods for fractional response variables with an application to 401 (k) plan participation rates. *Journal of applied econometrics*, Vol. 11, No. 6, 1996, pp. 619–632.
29. Papke, L. E. and J. M. Wooldridge, Panel data methods for fractional response variables with an application to test pass rates. *Journal of Econometrics*, Vol. 145, No. 1-2, 2008, pp. 121–133.
30. US Census, *2010 Urban Area FAQS*, 2010.
31. Gwet, K. L., *Handbook of inter-rater reliability: the definitive guide to measuring the extent of agreement among raters ; [a handbook for researchers, practitioners, teachers & students]*. Advanced Analytics, Gaithersburg, MD, 3rd ed., 2012.
32. Bobicev, V. and M. Sokolova, Inter-Annotator Agreement in Sentiment Analysis: Machine Learning Perspective. In *RANLP*, 2017, pp. 97–102.