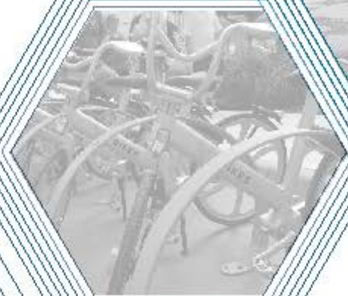

SOCIAL MEDIA ANALYSIS FOR TRANSIT ASSESSMENT

Won Hwa Kim
Kate Kyung Hyun
Ge Gordon Zhang
Anthony Giarrusso

CENTER FOR TRANSPORTATION EQUITY, DECISIONS & DOLLARS

The Center for Transportation, Equity, Decisions and Dollars (CTEDD) is a USDOT University Transportation Center, leading transportation policy research that aids in decision making and improves economic development through more efficient, and cost-effective use of existing transportation systems, and offers better access to jobs and opportunities. We are leading a larger consortium of universities focused on providing outreach and research to policy makers, through innovative methods and educating future leaders of the transportation field.

UNIVERSITY OF TEXAS ARLINGTON
CAL POLY SAN LUIS OBISPO
Georgia Tech
USF UNIVERSITY OF SOUTH FLORIDA
WISCONSIN UNIVERSITY OF WISCONSIN–MADISON

FINAL

CENTER FOR TRANSPORTATION, EQUITY, DECISIONS & DOLLARS

CENTER FOR TRANSPORTATION EQUITY, DECISIONS & DOLLARS

SOCIAL MEDIA ANALYSIS FOR TRANSIT ASSESSMENT

FINAL PROJECT REPORT

By:

Principal Investigator: Won Hwa Kim[†]
Co-principal Investigators: Kate Kyung Hyun[†],
Ge Gordon Zhang[‡], Anthony Giarrusso[‡]
Student Researchers: Anna Philip[†], Farah Naz[†], Vivek Patel[†],
Xin Ma[†], Lijun Tong[‡], Yiou Zhu[‡]

[†]University of Texas at Arlington
[‡]Georgia Institute of Technology

Sponsorship:
CTEDD

For:
Center for Transportation, Equity, Decisions and Dollars **(CTEDD)**
USDOT University Transportation Center
The University of Texas at Arlington
601 W.Nedderman Dr. Suite 103
Arlington TX 76019-0108 United States
Phone: 817-272-5138 | Email: C-Tedd@uta.edu

Disclaimer

The contents of this report reflect the views of the authors, who are responsible for the facts and the accuracy of the information presented herein. This document is disseminated under the sponsorship of the U.S. Department of Transportation's University Transportation Centers Program, in the interest of information exchange. The Center for Transportation, Equity, Decisions and Dollars (CTEDD), the U.S. Government and matching sponsor assume no liability for the contents or use thereof.

Technical Report Documentation Page

1. Report No. CTEDD 018-08	**2. Government Accession No.**	**3. Recipient's Catalog No.**
4. Title and Subtitle SOCIAL MEDIA ANALYSIS FOR TRANSIT ASSESSMENT		**5. Report Date** 12/1/2019 **6. Performing Organization Code**
7. Author(s) Won Hwa Kim, Kate Kyung Hyun, Ge Gordon Zhang, Anthony Giarrusso		**8. Performing Organization Report No.**
9. Performing Organization Name and Address Center for Transportation, Equity, Decisions and Dollars (CTEDD) USDOT University Transportation Center The University of Texas at Arlington 601 W.Nedderman Dr. Suite 103 Arlington TX 76019-0108 United States		**10. Work Unit No. (TRAIS)** **11. Contract or Grant No.**
12. Sponsoring Organization Name and Address United States of America Department of Transportation Research and Innovative Technology Administration		**13. Type of Report and Period Covered** **14. Sponsoring Agency Code**
15. Supplementary Notes		

16. Abstract

Stakeholders and transportation planners often use users' feedback to assess transit services including ride hailing platforms and reflect them for future plans. Interestingly, social network services (SNS) provide such information in a large set of text by individuals exchanging event base attitude and sentiment. This information is very useful, however, these data are often unorganized and it is intractable to process this extremely large set of text data by human effort whose size is continuously increasing. In this regime, we collected ride hailing service relevant text data from Twitter and created a database, and developed a novel Deep Learning (DL) framework that process and classify sentences that will automatically categorize the texts uploaded by service users according to transportation service specific criteria's. Our model uses multiple kernels for convolution to capture local context among neighboring words in texts and is simplified by summarizing parameters in traditional models using a kernel function. Using our DL model, we trained a classifier that identifies 1) to which transit service a text corresponds (e.g., reliability, mobility and cleanness), and 2) which sentiment the text contains (i.e., positive vs. negative). Its prediction performance is comparable to state-of-the-art DL methods but our model converges much faster during training which means it trains much more efficiently. We expect that our framework will provide feedbacks for policy makers who explore communication and information technology to create strategies to improve system efficiency and transit ridership.

17. Key Words Social Media Analysis, Text Analysis, CNN, Twitter, Transportation Evaluation		**18. Distribution Statement**	
19. Security Classification (of this report) Unclassified.	**20. Security Classification (of this page)** Unclassified.	**21. No. of Pages** 27	**22. Price** N/A

Table of Contents

Abstract
Chapter 1: Introduction
Chapter 2: Related Works
Chapter 3: Novel Deep Learning Framework for Text Analysis
Chapter 4: Performance Evaluation of the Developed Model
Chapter 5: Deploying MKCNN on Geographic Information System
Chapter 6: Conclusion
Reference

List of Tables

Table 1. Transit Performance Measures (TCRP Report 88)
Table 2. Performance of different models for sentiment analysis
Table 3. Performance of different models for transit performance category classification
Table 4. Examples of correctly classified tweets using MKCNN

List of Figures

Figure 1. Illustration of Gaussian kernel with different parameters σ and μ (localization and dilation) affecting different dimensions in a word embedding.
Figure 2. Overall Architecture of our MKCNN framework. It consists of input layer (word embedding), convolution layer, pooling layer, fully connected layer and output layer. Multi-kernel convolution is performed at the convolution layer.
Figure 3. Label distributions in our Social Media Transit Text Dataset. Left: distribution of 11 performance categories, Right: distribution of sentiments.
Figure 4. Confusion matrix of classification result on our dataset using MKCNN. The diagonal elements represent correct prediction, which are dominant compared to off-diagonal elements. Row: true labels, Column: predicted labels.
Figure 5. A web application to view MTA transit information in New York City.
Figure 6. An information box after clicking the marker.
Figure 7. A customized map using filters.
Figure 8. A 3D map by zooming in.

Abstract

Stakeholders and transportation planners often use users' feedback to assess transit services including ride hailing platforms and reflect them for future plans. Interestingly, social network services (SNS) provide such information in a large set of text by individuals exchanging event base attitude and sentiment. This information is very useful, however, these data are often unorganized and it is intractable to process this extremely large set of text data by human effort whose size is continuously increasing. In this regime, we collected ride hailing service relevant text data from Twitter and created a database, and developed a novel Deep Learning (DL) framework that process and classify sentences that will automatically categorize the texts uploaded by service users according to transportation service specific criteria's. Our model uses multiple kernels for convolution to capture local context among neighboring words in texts and is simplified by summarizing parameters in traditional models using a kernel function. Using our DL model, we trained a classifier that identifies 1) to which transit service a text corresponds (e.g., reliability, mobility and cost), and 2) which sentiment the text contains (i.e., positive vs. negative). Its prediction performance is comparable to state-of-the-art DL methods but our model converges faster during training which means it trains much more efficiently. The model is deployed on a Geographic Information System (GIS) map which can be interactively used by any public users. We expect that our framework will provide feedbacks for policy makers who explore communication and information technology to create strategies to improve system efficiency and transit ridership.

Chapter 1: Introduction

Global ride hailing market is expected to grow over $120 Billion by 2025 and North America will be the highest contributor to the global market. Various surveys, including Allied Market Research, mention that the user penetration of ride-hailing mobility service is expected to exceed over 20% in the next five years. This emerging technology is largely led by two operators in US, Uber and Lyft where Uber was estimated around $50 and Lyft around $11 billion in their values, respectively as of 2018 (Henao and Marshall, 2018).

Historically, stakeholders and transportation planners collect various information to maintain reliable and quality transit service focusing on spatial coverage (e.g., distance to stops/stations; operating hours), accessibility to transit infrastructure (e.g., physical access to stops/stations), and operation performance (e.g., punctuality and reliability) as given in Transit Cooperative Research Program (TCRP) Report 88 (Ryus, 2003). On the other hand, decision-makers need to value more on individual feedback of the ride-hailing operation for evaluating the system because success of ride-hailing largely depends on individual-to-individual experiences. As opposed to the public transit system that is operated based on fixed route and schedule, ride-hailing match on-demand rides between customers and drivers via website and mobile apps. Therefore, individuals evaluate the ride hailing system with either demand-responsive measures such as spatial availability or reliability based on their travel itineraries or subjective and perceptional measures such as driver behaviors and safety.

Since individuals experience different ride hailing trips spatially and temporarily, a large amount of data collected from anonymous users should be quite beneficial to generate more diversified and un-biased information to understand how individuals perceive the service and what aspects they value from their ride-hailing trips. Social network service (SNS) is one of many sources of such information that provides a large but unorganized database of information where individuals exchange event base attitude and sentiment (i.e., experience from individual transportation activity). This information often leads a chain effect that encourages others to react the message (e.g., a single post on a Twitter is visible to those who are connected to the commenter and recursively propagates beyond them). While these posts reflect users' exhaustive experience on transportation service quality and performance, it is extremely difficult to derive meaningful information by human force since the data are large, arbitrary and complex.

This study investigated a strategy for extracting and mining on ride-hailing information from a social media platform, i.e., Twitter, and to develop a machine learning model to identify hidden information from data that are relevant on self-assessment from active riders, such as sentiment and feedback on the service. This study aimed at categorizing and analyzing ride-hailing experiences by various performance measures based on individual perception and service quality. The main contributions of this study are: 1) to evaluate current ride-hailing systems, ~2 million tweet data were collected, filtered and annotated to create a new database of transit related text from Twitter, 2) for the data analysis, we developed a novel Deep Learning framework that trains and performs classification on text data, 3) and provide an extensive analysis to understand current transit services from users' perspective. Since understanding users' experiences and opinions on the individual ride-hailing trips is a key for successful planning and implementation, this study is expected not only to measure the efficiency and equality of ride-hailing transportation service but also to support decision-making strategies to upgrade current operation. The research results expected to help policy makers and planners understand the role of social network data for better understanding transportation service efficiency and performance.

Chapter 2: Related Works

Social Network System (SNS) Sentimental Analysis for Transit Service Assessment

With the advent of social media platforms, individuals are able to express their feelings instantly. Researchers have been discussing the advantages of using social media to understand individual's perception and their sentiments on different context. Sentiment analysis, known as opinion mining or emotion AI, uses the natural language processing to extract subjective information from a body of text. Transit agencies are one of the entities that could benefit from SNS analysis. Their competition with other transportation modes such as automobile and ridehailing for long-distant trips, and active transportation of walking and biking for shorter distance trip has been growing in recent years. Understanding perceptions, willingness-to-use, barriers, and challenges of current transit users directly relates to decision-makings to introducing new service and expanding current operation.

Researchers have found that commuters tend to choose their mode of transportation considering cost-effectiveness, comfort, punctuality, safety, efficiency, and cleanliness altogether to maximize their overall utility and satisfaction (Andreassen, 1995, Hanna and Drea, 1998, Thevathasan and Balachandran, 2007). However, transit agencies are not always successful to satisfy the needs of passengers due to their nature of operation and management strategies such as fixed route and limited operation hours (Collins et al., 2013). However, management may not aware of their failure from coming from human interactions or technical issues that could be easily fixed in current transportation system. This is because traditional performance matrices (e.g., punctuality, realibility) focus on operational efficiency and do not capture the customers' dissatisfaction to the service from non-operational issues. Lack of knowledge on custermers' needs and gap often misallocates limted resources and ultimately lose customers (Koushki et al., 2003). The Transit Capacity and Quality of Service Manual (TCQSM 1999) showed a huge discrepancy on expectations to the transit service between transit agencies and passengers. Agencies also do not tend to make sufficient invest to the areas that already showed high riderships because they tend to believe that the high riderships is resulted from good service. Researchers suggested to use social media to hear diverse opinions to capture commuters concerns in real time even from those areas with high activities. Collins et al. (2013) showed that SNS analysis is advantageous for agencies to collect user specific needs and meaningful insights on the service, particularly for the areas that require improvement. Sentiment analysis simply serves as a tool for transit agencies to predict future ridership and to make informed decisions for a better travel experience.

Transportation Performance Assessment using SNS data

Various transportation projects have already begun to adopt machine learning and data mining techniques for transportation planning and operation for decision-making processes (Kavanaugh et al., 2012, Bregman, 2012, Collins et al., 2013, Gal-Tzur et al., 2014). (Pratt et al., 2019) recently analyzed over 2000 tweets commenting shared ride-hailing (UberPool and Lyft Shared/Line) and found that other passengers' behavior is one of the most important factor negatively affecting shared-ride experiences. Zulkarnain *et al.* also used twitter to evaluate user sentiments on six categories of service such as cost, payment, food delivery, and application stability for the ride hailing service in Indonesia (Zulkarnain et al., 2018). Pratt et al. (2019) analyzed 2000 tweets of shared ride-hailing service, Uber Pool and Lyft Shared, to investigate how travelers and drivers understood the service and communicated through social network platform. The study found that the negative tweets significantly outnumbered the positive tweets, as people tend to post their complaints rather than expressing gratitude for their experiences. The study also found that most of the tweets had humorous tone to post mostly about their fellow passenger's behavior. Zulkarnain et al. (2018) used the Latent Dirichlet Allocation (LDA) to generate nine common topics from the tweets for ride hailing providers in Indonesia. The topics include responsiveness to customers' complaints, experience and issues in food delivery service, the reliability of the service, loyalty points and rewards, fraudulence occurrence and possibility, drivers' behavior and customers' trip experience, electronic money system, instant courier service, and the fare of the services. The study showed how ride-hailing service providers could use customers' opinion, experience and complaints about the service based on the topics detected by the algorithm.

SNS data has been widely used to evaluate transit service performance. Bregman (2012) introduced several social media metrics that transit agencies used to capture public insight on services and contents. For example, likes and shares of a post in Facebook, the number of followers and tweets in twitter, and the number of website visits from google analytics could help agencies determine popular service that users are favor of or unpopular service that requires improvement. Gal-Tzur et al. (2014) provided an overview of different social media platforms, characteristics of their contents and how these characteristics are reflected in transport-related posts. Kavanaugh et al. (2012) conducted focus group interviews with officials from Arlington County in Virginia to understand how various social media contents such as Facebook, Twitter, YouTube and other local webpages were used to manage routine and emergency events in their region. In particular, semantic analyses appear to widely adopt to identify popular service and understand frequent users' opinions. The study found that local government could benefit from social media analysis to draw sensible conclusion since overwhelming amount of data easily reveal hidden information about users and services.

Researchers however pointed out a few limitations of SNS analysis. Pratt et al. (2019) indicated that humor or human emotion within the texts was found to be highly difficult to capture in SNS data analysis. Biased tones and contents towards negative perception and experience may also mislead the overall quality (Kavanaugh et al., 2012). Owyang et al. (2010) found similar limitation and concluded that solely using SNS data for decision-making may not provide meaningful overall insight on transit service to. Kavanaugh et al. (2012) advised to systematically arrange and pre-process the SNS information to archive data as the data collection API of social media platforms such as Twitter tend to mix past and present data altogether.

Text Data Mining and Machine Learning

Adopting machine learning algorithms for text classification has a rich history in Natural Language Processing (NLP) community. If one can represent text data with specific labels as a vector of the same length (i.e., word embedding), then he/she can adopt a supervised machine learning algorithm to train a label prediction model to classify future incoming texts, e.g., its sentiment (positive vs. negative). Support Vector Machine (SVM) was able to successfully carry out various text classification problems such as topic classification (Colas and Brazdil, 2006, Joachims, 1998, Tong and Koller, 2001), sentiment analysis (Pang and Lee, 2004, Vinodhini and Chandrasekaran, 2012), spam filtering (Caruana et al., 2013, Caruana et al., 2011, Drucker et al., 1999) and many others. For similar applications, Naïve Bayes and similar Bayesian statistical models has been utilized for its simplicity and high performance with smaller dataset (Schneider, 2003, Metsis et al., 2006, Pang et al., 2002, McCallum and Nigam, 1998). Many ML algorithms use ``frequency'' of terms for word embedding such as term frequency – inverse document frequency (*tf-idf*) (Ramos, 2003, Aizawa, 2003), which is a text length invariant feature by counting how frequent specific words appear in a document. This approach is powerful in practice since similar types of documents may contain the same words that occur more often than others. However, it does not consider spatial information of the words which is quite useful to characterize a sentence, since the relationships among neighboring words contain significant semantic information. Recent Deep Learning methods using Convolution Neural Network (CNN) and Recurrent Neural Network (RNN) are the perfect examples that explore the relation between words. CNN methods use convolution in signal processing to derive local context information between words in a text as a feature for downstream classification (Kim, 2014, Kan and Thi, 2005, Lee and Dernoncourt, 2016, Joulin et al., 2016), while RNN approaches takes a sentence as a sequence of words into their model to characterize the order and associations of words (Zhou et al., 2016, Wang et al., 2016, Yin et al., 2017, Zhou et al., 2015). These models can "learn" the word embedding during their training phase and are often more suitable to analyze relatively shorter text such as sentences from social media since 'frequency' of specific words may not manifest well.

Chapter 3: Novel Deep Learning Framework for Text Analysis

We consider a problem where a set of text (i.e., a collection of sentences) is available from which we extract meaningful information (e.g., specific pattern). In this study, the information relates to ride-hailing service which can be used in policy making and strategy development for urban transportation planning. In general, such a process requires substantial human effort to go through a large amount of data to identify ride-hailing relevant text and identify meaningful information (e.g., sentiment and category of services). It eventually becomes infeasible with limited resources especially if the dataset is extremely large (e.g., millions of tweets). In this regime, this study developed a novel Convolution Neural Network (CNN) model that performs classification on big text data.

Preliminary: Convolution Neural Network in Natural Language Processing

To keep this report self-contained, a conventional CNN framework for text classification is first introduced in (Kim, 2014) as a preliminary. Given a set of sentences of s_i and their labels y_i, the objective of a CNN model for text classification is to learn a model $\mathrm{f}(s_i) = \hat{y}_i$ that accurately predicts its label $\hat{y}_i$. A CNN model extracts "features" from the raw data using layers of convolution operation, which was originally studied in signal processing (Haykin and Van Veen, 2007). A convolution operation is formulated as a linear combination between two functions, i.e. an original signal and a filter. For texts, we first need to embed the text data into a vector representation, and only then convolution can be performed on two vectors which are representations of sentences and filters in the vector space. This text-to-vector embedding can be trained inside a machine learning model, or one can use a pretrained word-to-vector embedding frameworks such as word2vec (Goldberg and Levy, 2014) and GloVe (Pennington et al., 2014) that transform each word to a fixed sized vector.

Given the word embedding of training text data, the word embeddings are inputted to a CNN framework that consists of convolution layers, pooling layers and fully connected (FC) layer to yield a prediction outcome (e.g., prediction of a class). CNN first extracts various features from the word embedding in the convolution layer which is equivalent to higher dimensional mapping, and downsamples the data by pooling the maximum values in small windows (i.e., dimension reduction). Then these pooled features are fed into a fully connected layer which is a classification model for a final outcome. The prediction errors (e.g., cross entropy or L2-norm) from the entire or partial training set is backpropagated to update necessary parameters in the neural network model using gradient descent approaches.

Given the word embedding of training text data, the word embeddings are inputted to a CNN framework that consists of convolution layers, pooling layers and fully connected (FC) layer to yield a prediction outcome (e.g., prediction of a class). CNN first extracts various features from the word embedding in the convolution layer which is equivalent to higher dimensional mapping, and downsamples the data by pooling the maximum values in small windows (i.e., dimension reduction). Then these pooled features are fed into a fully connected layer which is a classification model similar to logistic regression. The prediction errors (e.g., cross entropy) from the entire or partial training set is backpropagated to update necessary parameters in the neural network model using gradient descent.

Formally, when a word is represented as a k-dimensional vector $x_i \in \mathbb{R}^k$, a sentence s that consists of n words can be represented as a concatenation of these vectors as $\mathrm{s} = x_1 \oplus x_2 \oplus \cdots \oplus x_n$ where $\oplus$ is a concatenation operator. These word embeddings can be randomly initialized and learned while training a CNN model which yields a look-up table to map a word to a fixed sized vector representation. The embedded matrix is constructed by concatenating the word embeddings vertically that result in a matrix of size $\mathrm{N} \times \mathrm{k}$ where N is the maximum number of words in a sentence and k is the length of the word embedding. Specifying a small receptive region of h words in a sencence, weights $\mathrm{w} \in \mathbb{R}^{h \times k}$ (i.e., a filter) are associated with the word embedding to define a convolution in that specific field yielding an outcome $\mathrm{c}(l)$ at index l as

$$c(l) = \sum_{j=1}^{k} \sum_{i=1}^{i+h} w(i,j) s_{l:l+h-1}(i,j) + b$$

where b is a bias term and $s_{l:l+h-1}$ are words from the l-th to $(l + \mathrm{h} - 1)$-th location in a sentence. In the convolutional layer, multiple filters of various sizes are convolved with the input data to obtain useful feature maps. The weights w of the filters are randomly initialized and learned during a training phase. The convolution, a linear combination between a word embedding of an input text and a filter as in equation above, is performed by sliding down the filters through the word embedding matrix. These filtered input c are fed into an activation function which is an element-wise non-linear function such as Rectified Linear Units (ReLU) or sigmoid to yield a feature vector M of size $(\mathrm{N} - \mathrm{h}) + 1$.

Performing convolution with P different filters yields multiple feature vectors M_p, and then max pooling is performed at the pooling layer, i.e., picking the largest value in each M_p, to select the most significant features from the feature vectors and reduce their dimensions. These pooled features are then inputted to the FC layer to obtain a prediction $\hat{y}_i$ whose error between y_i will be backpropagated to train all the parameters in the FC layer, the filters and look-up table (if necessary) via optimization methods such as gradient descent. The necessary error is formulated as a loss function typically using cross-entropy measures to be described shortly.

Multi-kernel Convolution Neural Network

Convolution in general can be described as a filtering/smoothing operation that extracts useful information from data from local context (i.e., neighboring) information. For example, if we are analyzing text data (i.e., sentences), the local context information of each word can be derived by looking at how it is related with its neighboring words. While traditional CNN models define a $\mathrm{w} \in \mathbb{R}^{h \times k}$ as a filter whose number of parameters are large and learns them independently ignoring distances among different words. Therefore, the size of the window h often remains small (e.g., 3 or 5) focusing only in a small region in a sentence. We therefore define a kernel function, e.g., Gaussian kernel $g_{\sigma,\mu}(\cdot)$ with mean μ and standard deviation σ, that are characterized by a small number of parameters and covers much longer spectrum of the given sentence as

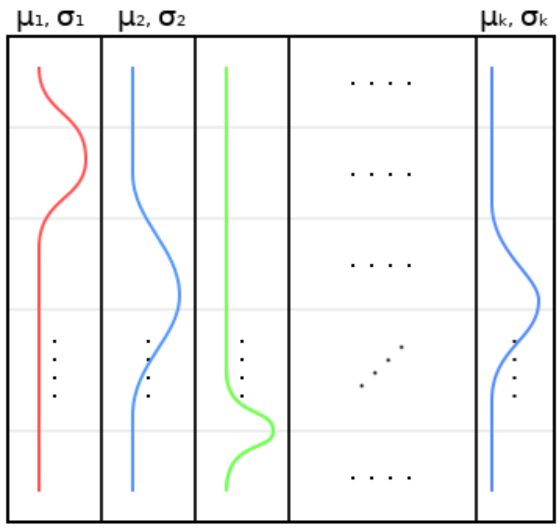

Figure 1. Illustration of Gaussian kernel with different parameters σ and μ (localization and dilation) affecting different dimensions in a word embedding.

$$c(l) = \sum_{j=1}^{k} \sum_{i=1}^{i+h} g_{\sigma,\mu}(i,j) s_{l:l+h}(i,j) + b$$

whose outcome is a vector $c(l)$ that will go through an activation function (i.e., ReLU) to yield a feature map M. This concept of kernel convolution is visualized in Fig. 1 with a Gaussian kernel of two parameters. Here, the μ defines the localization of the kernel (i.e., the center of the filter) and σ controls its dilation (i.e., scale of the filter) that determine the width of local regions to cover with the filter. We define this kernel function for every dimension of the word embedding to extend our framework to a Multi-kernel Convolution Neural Network (MKCNN), where we learn different kernel functions adaptively for individual dimension of a word embedding.

Model Architecture

Similar to other fundamental CNN architectures, our framework consists of an input layer, convolution layer, pooling layer and FC layer as shown in Figure 2. The input to the model is a text s_i (i.e., sentence) which has a label y_i assigned using a one-hot-vector encoding representing different classes (a vector in the size of the number of available labels with all 0s except for one 1 representing a specific class). First, an s_i is converted into an word embedding matrix which has a size of N × k. The dimension of the matrix is fixed and padded with 0s as necessary for differing lengths of sentences.

A layer of multi-kernel convolutions of s_i with P different filters sizes, which gives us P feature maps of M_p. Then, max pooling is performed over each feature map to extract the most important features and concatenated as a vector z. This z becomes the input to the FC layer whose parameters are denoted as β with the final output layer with a softmax activation function. The

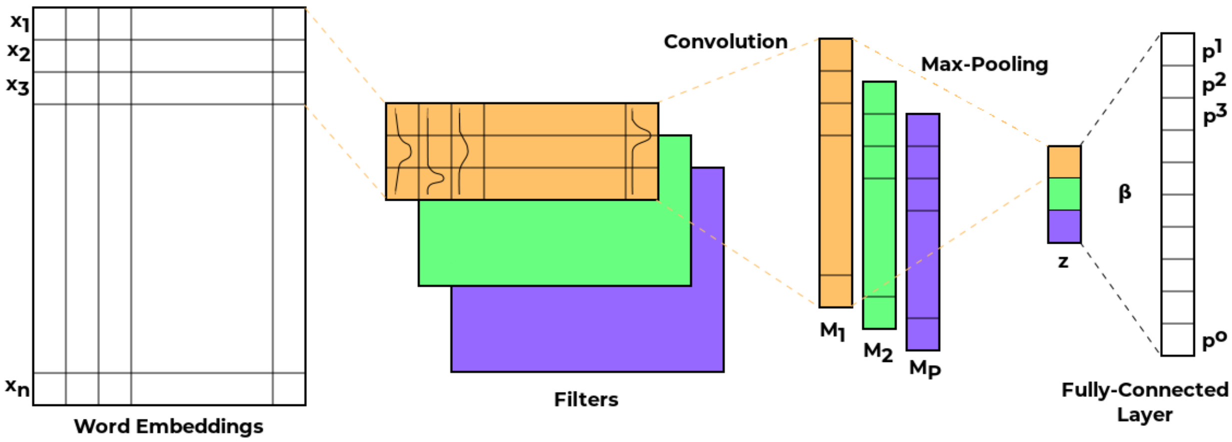

Figure 2. Overall Architecture of our MKCNN framework. It consists of input layer (word embedding), convolution layer, pooling layer, fully connected layer and output layer. Multi-kernel convolution is performed at the convolution layer.

output layer has O number of nodes that correspond to different labels to classify. To an input sentence s_i, the softmax function assigns probability distribution p_i^o for each class label that identifies the most likely label $\hat{y}_i$ (i.e., predicted label) as

$$p_i^o = softmax(\beta^T z_o) = \frac{e^{\beta^T z_o}}{\sum_o e^{\beta^T z_o}}$$

whose sum equals to 1. We select an output node that has the largest probability and assign a label that corresponds to that node, i.e., $\hat{y}_i = \ argmax_o p_i^o$.

To train the model, we formulate a loss function that measures the error between the prediction p_o and the true label y_i which we want to minimize. One of the loss functions commonly used for multi-class classification is the cross entropy which is defined as

$$\mathrm{L}(p) = \sum_i \sum_o y_i^o \log\,(p_i^o) + \ \lambda \|\beta\|_2$$

where y is the true class label, p is the model's predicted probability, β are the parameters in the FC layer and λ is a user parameter to control balance between cross-entropy error and ℓ_2 −regularization to avoid overfitting of the model. The error is back propagated through the neural network and used to update the learnable of the model which are β, μ, σ and the look-up table (i.e., word embedding) in such a way to reduce the $\mathrm{L}(p)$ that eventually reduces the error between prediction and groud truth. The gradient descent is performed using partial derivatives from each of these learnable parameters, and we used the Adam optimizer in *Tensorflow* package in Python for backpropagation.

Social Media Dataset

Data Collection

The text data were collected through Twitter APIs provided by Twitter. Twitter offers two different APIs to collect tweets, i.e., historical and real-time tweet collecting APIs. Considering that the developed CNN framework requires a large dataset to train a CNN model, this study chose to use the streaming API which is much more efficient than collecting historical tweets with limited access. This study developed a tweet collection framework that calls the streaming API and saves the returned tweets into the database. The framework was written in python, and twitter4J library was used to connect the twitter streaming API to our software. With the streaming API, there was no rate limit and maximum of 60 tweets per second were collected. In addition, the API provides the keywords and geolocation filter to prescreen the tweets and only qualified tweets will be collected. The titles of popular ride-share companies like Uber and Lyft were used as keyword for filtering as the scope of this study is social media data analysis. Total of 1,925,952 Uber/Lyft relevant tweets were collected between January 23 and February 1 in 2019.

Sentiment Analysis

A basic task of sentiment analysis is to classify the polarity (positive, negative or neutral) of a given text, even beyond polarity (angry, sad or happy), and classify the subject matter etc. Sentiment analysis can capture the connotation of an opinion and change the way transit agencies measure rider satisfaction (Collins et al. 2013). Sentiment analysis can provide agencies with an insight on how a brand or service is perceived in relation to value and quality.

Performance Category and Annotation

Numerous studies have been conducted to develop and implement criteria/categories to evaluate performance of quality and performance of transportation system. However, scant research and effort has been made to understand rider's perception and develop performance measure specifically for ride-hailing services. Performance measures are typically adopted transit agencies for reporting purpose, self-improvement, and communicating results (TCRP Report 88) as shown in Table 1. In addition to the performance measure that is required to be reported at the National Transit Database, transportation planners collect supplementary measures to improve their services, accomplish agency goals, attract new riders, and provide necessary information to the funding authority and stakeholders.

Measuring transportation system performance has gained increasing attention since 1990 through Transit Capacity and Quality of Service Manual (TCQSM) which provides a tool for transit agencies to measure quality of service from passengers' point of view (Kittelson et al., 2003). TCQSM defines the quality of service as the reflection of passenger's perception on system performance. The performance measures that are used to explore riders' point of view are diverse from the typical financial and output focused measures required for National Transit Database reporting. It reflects the extent to which a transit service meets the need of its customers and has an important implication in the context of ride hailing services. Komanduri at al. (2018) conducted a study on Ride-Austin, a non-profit mobility-on-demand service in Austin Texas to measure the impact of the service in urban context. They used traditional performance metrics such as total trips, vehicle-miles traveled, total travel times, and fares paid, along with more sophisticated measures such as deadheading, terminal times, and the number of active riders and drivers. Xu et.

Al. (2019) developed four performance measures from the operators' point of view including percentage of abandonment, average queueing time for riders, average pick-up time and queueing time variance for riders. Developing performance indicators for ride hailing services and understanding the transferability of the traditional performance metrics to this service are crucial for providing effective and comfortable mobility option for individuals

TABLE 1. Transit Performance Measures (TCRP Report 88)

Performance Measure Categories	Description	Example Measures
Availability	Assess how easily potential passengers can use transit service. Include spatial availability and temporal availability.	Service coverage, service density, stop spacing, stop accessibility, frequency of service, hours of service, response time, fleet composition, percent-person minutes served, service denials, pass ups, transit accessibility index
Service Delivery	Assess passengers' day-to-day experience using transit. Include reliability and customer service	On-time performance, headway regularity, missed trips, lost service, scheduled miles per minute of delay, run time ratio, compliant rate, percent of missed phone calls, customer service response time, driver courtesy, passenger environment, customer satisfaction
Community	Assess transit's role in meeting broad community objectives. Include mobility and community outcome.	Mobility, trip generation, welfare-to-work accessibility, service equity, community economic impact, personal economic impact, employment impact, land development impact, environmental impact
Travel Time	Evaluates how long it takes to make a trip by transit in compared to other mode. Include travel time and speed measure	travel time, travel time variability, transit-auto travel time, number of transfers, transfer time, delay, travel speed, system speed
Safety and Security	Measures the likelihood that passengers can get involved into accidents or become a victim of a crime. Include driver's behavior, safety and vehicle condition.	Accident rate, passenger safety, percent positive alcohol/drug test, number of traffic tickets issued to operators, percent of buses exceeding the speed limit, number of crimes, ratio of transit police officers to transit vehicles, number of incidents of vandalism
Economic	Evaluate transit performance from a business perspective.	ridership, passenger miles traveled, cost efficiency, service miles per revenue miles, cost effectiveness, productivity

Capacity	Assess the ability of transit facilities to ove both vehicle and people. Include service denials, seat capacity and ridership.	passenger capacity, station/terminal capacity, vehicle capacity, volume to capacity ratio
Comfort	Overlap of different categories; especially service delivery, quality of maintenance, noise and other impacts of transit service	mean vehicle age, driver courtesy, customer satisfaction, fleet cleaning

This study largely adopted the conventional transit performance measures and revised them for ride-hailing services to capture users' experience and perception about the services. First, traditional performance categories were reviewed and evaluated based on their applicability to the ride hailing service environment. Nine main categories were created including availability, travel time, cost, human interaction, reliability, technology, safety, vehicle quality, and community outcomes. Descriptions of each category are as follow.

1) '*Cost*' refers to the cost that occurs to the user of the service. Tweets containing information about the user cost, cost efficiency and cost comparisons to other mobility options were labeled under this category.
2) '*Human Interaction'* measures driver behavior and interaction with passengers. Generally, a driver's behavior is not accounted as a performance measure for traditional transit service with little interaction between drivers and riders. However, in ride hailing service environment, one-to-one interaction plays a crucial role in service experiences for customers.
3) '*Reliability*' indicates how much ride hailing services provide promised experiences. For example, users evaluate if their vehicle shows up on-time or arrives at the destination within the expected arrival time window.
4) '*Safety*' refers to user perceptions on safety for their ride-hailing trips. Traditional transit performance measures captures passenger safety using accidents rate or response time to that incident. In ride hailing services, the spectrum of safety is broader including unsafe driving behavior, speeding, and traffic rule violation.
5) '*Technology*' measures both customer service and app functionality. Tweets containing discussions about difficulty or ease of reaching customer service to resolve their personal issues in booking a trip or making a payment falls under the category of customer service. The app functionality measure captures any technological issues or suggestions on the mobile app platforms.
6) '*Availability*' refers to spatial and temporal coverage of the service. Unlike fixed-route transit services, ride-hailing services offer wider mobility without temporal limit. Tweets indicating the existence or unavailability of Uber/Lyft service at locations and time of their desired travel has been labeled under this category.
7) '*Travel Time*' refers to the time that a person experiences while traveling from one point to another. This is an important measure for the ride hailing services since this service is attractive in the basis of offering relatively shorter time by eliminating dwell time at stops

associated to the traditional transit services. Tweets including indication to in-vehicle travel time and wasting time for the vehicle to arrive has been labeled under this category.

8) '*Vehicle Condition'* is another factor that contributes to the overall perception of the quality of service in ride hailing environment. Tweets containing complaints, impressiveness, and cleanliness about the driver and vehicle were labeled into these categories.
9) '*Community outcome*' considers any benefits that community expect from the ride-hailing operation such as better job access, economic improvement, and increased mobility for vulnerable users.

This study annotated the collected tweets based on the performance measures defined above as ground-truth dataset. The annotators first assigned one performance measure label for every tweet and added sentiment label to identify if the users' experience with the performance category were positive or negative.

Dataset Description

The final dataset contained 3,329 tweets. Each tweet has a sentiment assigned either *positive* and *negative*, and a performance measure from among 9 categories described in the previous section. Figure 3(left) shows label distributions for the 9 performance categories, and Figure 3(right) shows that there are 472 positive and 2651 negative tweets in the dataset. If a single tweet is annotated with two or more performance categories, we copied the tweet and annotated it with different performance criteria. The dataset includes 579 duplicated tweets, most of these tweets are related to human interaction or reliability. There is a disproportionate number of tweets which are labeled as human interaction, which make sense as Uber and Lyft follow a peer to peer ridesharing model and riders interact a lot with the drivers. This dataset will be archived and released to help other studies that utilize social media text data for transportation analysis. The distribution of sentiment and performance category distributions in the dataset are shown in Figure 3. Due to the restrictions from Twitter, the actual text data cannot be made public. However, we are currently actively communicating with Twitter team to get permission to publish the dataset.

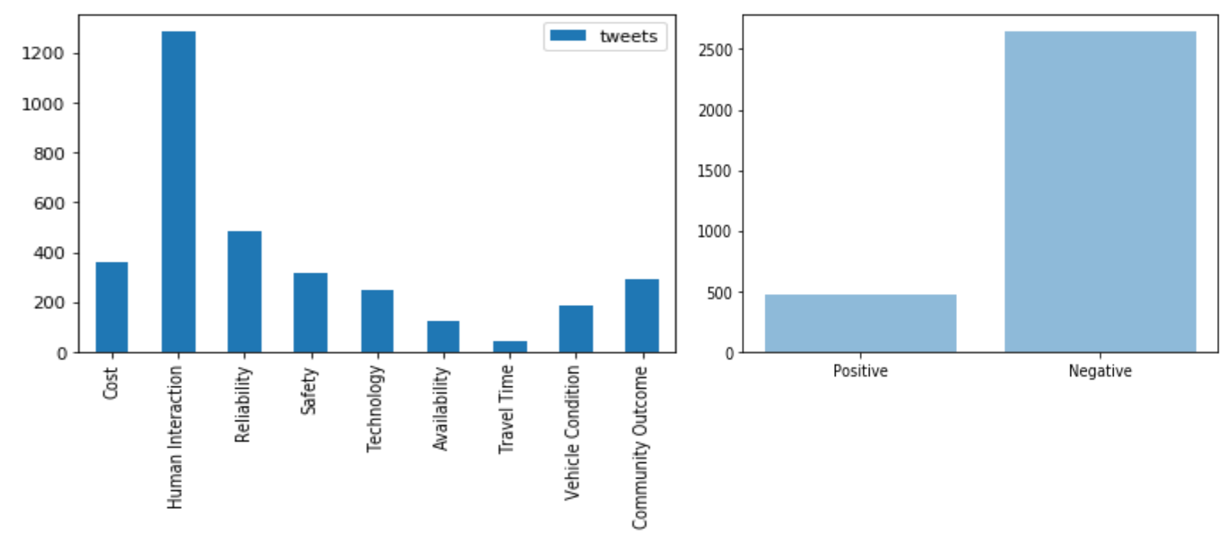

Figure 3. Label distributions in our Social Media Transit Text Dataset. Left: distribution of 11 performance categories, Right: distribution of sentiments.

Chapter 4: Performance Evaluation of the Developed Model

This study performed various experiments on the ride hailing text dataset that we collected, comparing our MKCNN with three other baselines: 1) CNN introduced in (Kim, 2014), 2) traditional Recurrent Neural Network (RNN) (Williams and Zipser, 1989) and 3) Long Short Term Memory (LSTM) inside Recurrent Neural Network (RNN) architecture which is popular for text analysis (Sak et al., 2014, Hochreiter and Schmidhuber, 1997). This study divided the dataset into training (70%) and testing (30%) and the final accuracy on the test dataset was evaluated.

Sentiment Analysis

Sentiment analysis finds whether a specific sentence/text yields positive or negative sentiment of users and is one of the most popular and useful applications in text analyses. We performed sentiment analysis to identify ride-hailing user perceptions to their overall experiences using the tweets collected and annotated in this study. We divided the dataset into training (70%) and testing (30%) dataset to train and compare the performance of our framework with three other baseline methods mentioned above.

Table 2 summarizes the sentiment analysis result comparisons with other baseline models. All models converged with high training accuracy which means they have learned the discrimination patterns for sentiments of tweets in the training set. As seen with the testing accuracy, the developed model outperformed RNN and LSTM which are the most popular deep learning frameworks for text analysis and achieved comparable result with a state-of-the-art CNN model. Using our framework, the accuracy for sentiment classification on the testing dataset was 94.8%, which is high considering that there are many ambiguous and somewhat meaningless text on typical social media platforms. This sentiment analysis is a binary classification task which is the simplest form of a machine learning task, and it validates that our model is showing sound performance on a social media text classification task.

TABLE 2. Performance of different models for sentiment analysis

Models	**Training Accuracy**	**Testing Accuracy**
MKCNN (ours)	0.983	0.948
CNN	0.992	0.949
RNN	0.986	0.861
RNN – LSTM	0.85	0.86

Transit Performance Analysis

Data Preprocessing

Out of 9 categories introduced in the earlier section, the final model used five categories -- cost, human interaction, reliability, safety, and technology -- that have sufficient samples (over 250 tweets) to properly train the ML models. Community outcome category was also removed since it reflected various aspects on the service such as income improvement or mobility, which could potentially be divided into sub-categories. To handle class imbalance problem, this study randomly selected 492 tweets from human interaction category (out of 1286), so that the category with the largest sample size does not exceed double of the number of samples of the category with the least samples. From a total of 1,901 tweets, 1,331 (70%) were selected as the training set and the rest 570 tweets were designated as the testing set for the downstream machine learning experiments.

Performance Category Classification

The developed model and three baseline models were applied on the preprocessed data for five category classification problem and the training and testing accuracies as show in Table 3. All models converged with high training accuracy of above 90% showing that all these models were properly trained to distinguish the five performance measures given the training text data. The same or similar hyper parameters were used to train the models. For models using convolutions, number of filters =24, $\mathrm{k} = 32$ and the number of different window sizes were set to 3. CNN used $\lambda = 3$ and $\mathrm{h} = [3, 4, 5]$ which were used in (Kim, 2014) and $\lambda = 1$ and $\mathrm{h} = [7, 9, 11]$ for MKCNN. Number of hidden layers for RNN were 2 and number of hidden nodes for each layer was 16. Increasing these numbers increased training time but did not improve the result.

The trained models showed different testing accuracies. Notice that this is a 5 class classification problem, therefore 20% accuracy is expected by a random guess. Even considering the non-uniform distribution of classes, the highest accuracy that one can get is 25.9% by predicting all the test samples as the human interaction category. The developed model MKCNN achieved accuracy of 47.2% which is comparable to CNN model (47.5%) and performs much better RNN or LSTM as seen in Table 3. There is an overfitting showing significant differences between training accuracy and testing accuracy mainly due to significant amount of noise and desultory text which are natural in social media. Due to a randomness inherent in deep learning models, this study observed some cases where both CNN and our model achieve slightly better accuracy (~50%) with different parameter settings, however the results are reported based in testing the models under the same or similar conditions.

TABLE 3. Performance of different models for transit performance category classification

Models	Training Accuracy	Testing Accuracy
MKCNN (ours)	0.941	0.472
CNN	0.96	0.475
Simple RNN	0.951	0.267
LSTM	0.95	0.333

A confusion matrix showing the category prediction result using our model is given in Fig. 4. The rows in the confusion matrix correspond to the true label and the columns are our predictions. We can see that the diagonal elements, which show the correct prediction, are dominant in the matrix compared to other off-diagonal terms for all categories. This shows that our trained model is relatively correctly classifying unseen tweets in the testing dataset in the regardless of their performance categories.

The actual examples of the correctly classified tweets with both category and sentiment using MKCNN are given in Table 4, three tweets extracted from each category. As mentioned earlier, the texts from social media contain significant noise and ambiguity from grammar error,

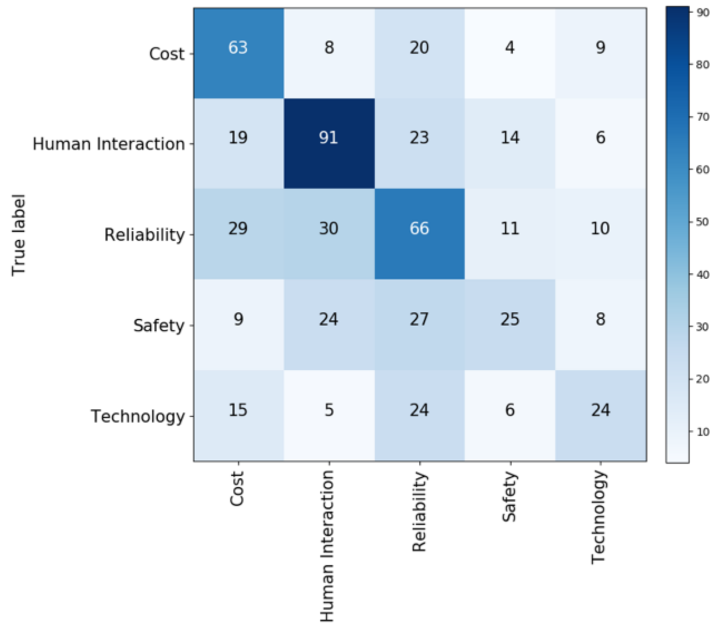

Figure 4. Confusion matrix of classification result on our dataset using MKCNN. The diagonal elements represent correct prediction, which are dominant compared to off-diagonal elements. Row: true labels, Column: predicted labels.

typo, meme and special notations; nevertheless, our model was able to successfully categorize these tweets to an extent that make sense to human understanding.

TABLE 4. Examples of correctly classified tweets using MKCNN

Tweet	Category	Sentiment
@infernal_monkey You better hire an uber or something my dude. That's 100% worth the fare!	Cost	Positive
@lastborn0805 @Taxifyng Well to me uber is cheaper	Cost	Positive
And the following ride cost us more than the double. Because was the only available... Please contact me ASAP! @AskLyft @lyft	Cost	Negative
these lyft drivers need to shut up - no i don't wanna hear about your dog tammy	Human Interaction	Negative
why is my uber driver telling me about how he disowned his dad for destroying his marriage 911	Human Interaction	Negative
You ever meet the most genuine nicest people ever and being around them make you so happy?? Bc that was me tonight w my lyft driver and his friends wow	Human Interaction	Positive
Four layers: heat teach leggings, fleece lined tights, regular leggings and jeans. The cold went right through. Thank God for @uber. No way was I waiting for a bus and two trains in this cold.	Reliability	Positive
Currently 15 minutes late to a meeting thanks to @uber	Reliability	Negative
update on this: my lyft driver messed up picking me up from my apt this morning so the 2 mi trip turned into a...30 mi one sksksksksk	Reliability	Negative
Yo this uber driver doing like 100 mph just going down Southern Avenue	Safety	Negative
How many hours do these ola/uber drivers drive in a day ?some say they drive for almost 18 to 20 hours ?is it acceptable > isnt it risky for the driver and the passenger ?	Safety	Negative
uber man ran a red light im https://t.co/gO3A34lwOX	Safety	Negative
Does anyone know how to finesse uber cos blud this app is finessing me.	Technology	Negative

@Uber I send uber for other people sometimes, and find your app constantly makes it a difficult experience, I have to go thru it several times, it reverts to my location when I enter to location	Technology	Negative
I have been a loyal customer and have always preferred using uber over any other cab service but this is really the limits. They have no helpline number & their emails keep directing you back to the app and report your complaint from their pre-made list of issue.	Technology	Negative

Chapter 5: Deploying MKCNN on Geographic Information System

Our pre-trained MKCNN model is deployed on a Geographic Information System (GIS) located at http://geospatial.gatech.edu/MTATransit/. To deploy MKCNN on geographic information systems (GIS), the following tasks were accomplished:

1) Fetching related Twitter feeds by Twitter streaming API and Python and then store the tweets in MySQL database
2) Categorizing the tweets from the content of Twitter messages using MKCNN
3) Visualizing the composite records on the map based on the geotags of the Twitter feeds using Mapbox
4) Establishing an interactive web platform by using JavaScript, CSS, and HTML

The developed web application provides interactive options for users to view MTA transit information at any location in New York City (see Figure 5). The transit performance evaluations, such as cost, human interaction, reliability, safety and technology, and perceived satisfaction of transit, such as positive and negative, are represented as markers with different colors on the map. The markers give more details on each tweet when users click them. The details include created time, content, and category of the tweet (in Figure 6). To provide more customized visualization on the map, we designed the filter of the time range, category, and confidence score of the tweets. The users can use those filters to find filtered transit information, depending on the users' needs (in Figure 7). Mapbox was chosen to be the provider of the base map to show the 2D/3D environmental details of the tweets (in Figure 8). The users can click on the map or search by address to zoom into a point of interest (POI).

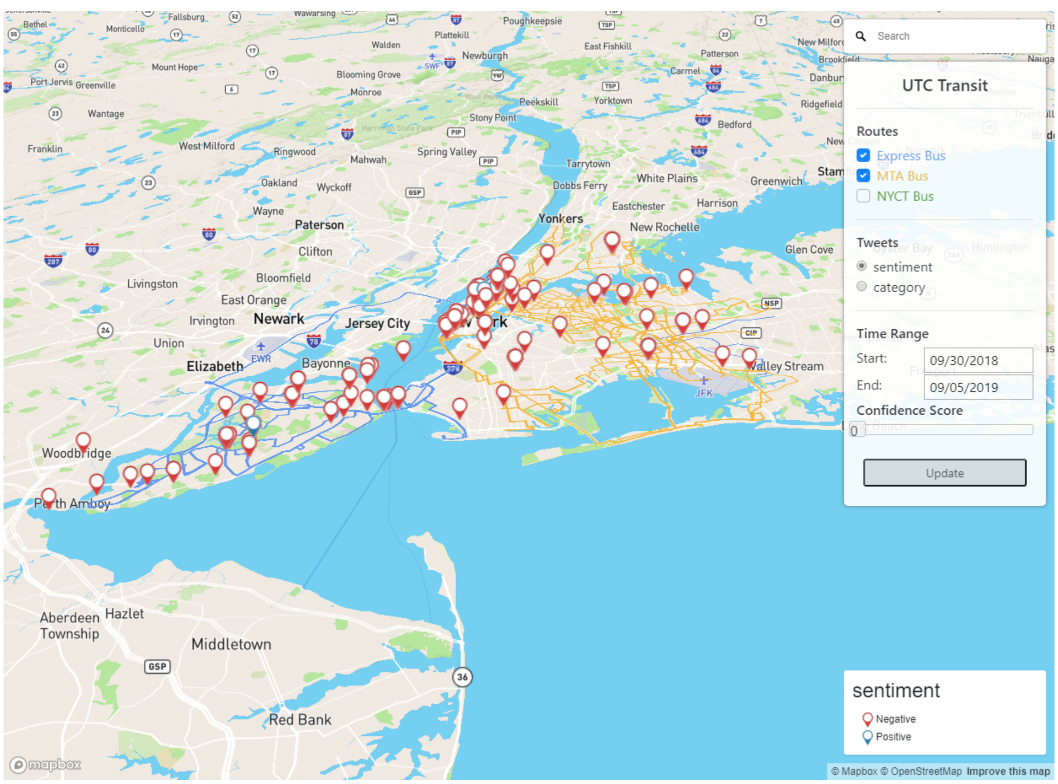

Figure 5. A web application to view MTA transit information in New York City

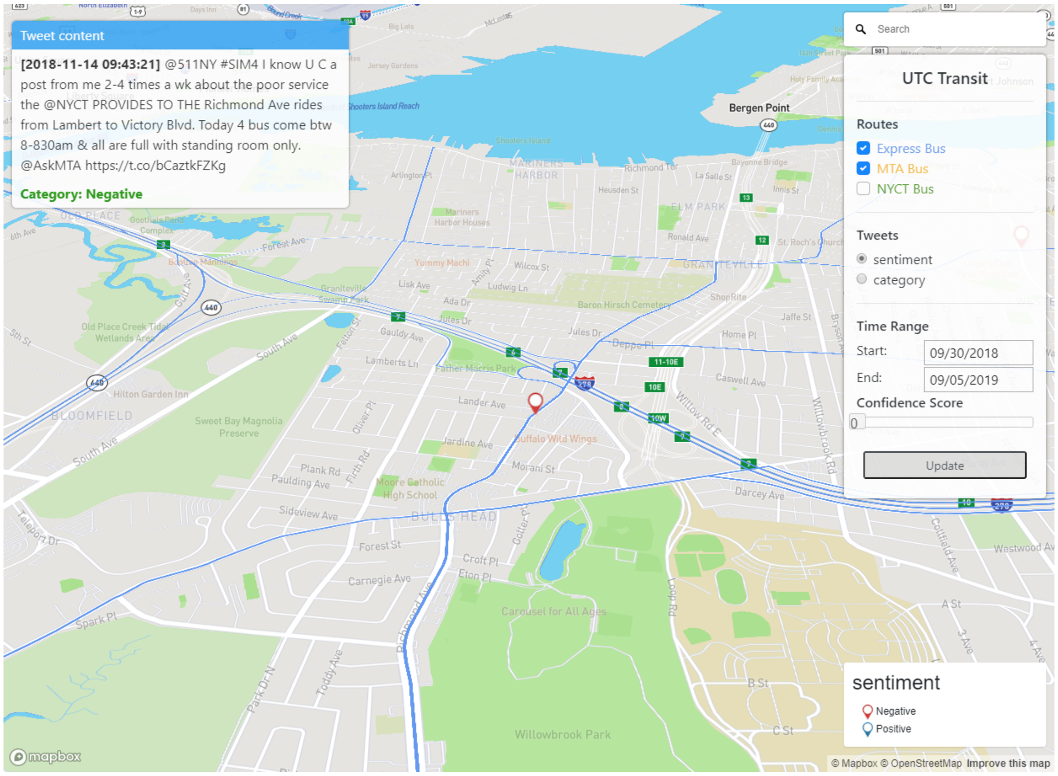

Figure 6. An information box after clicking the marker

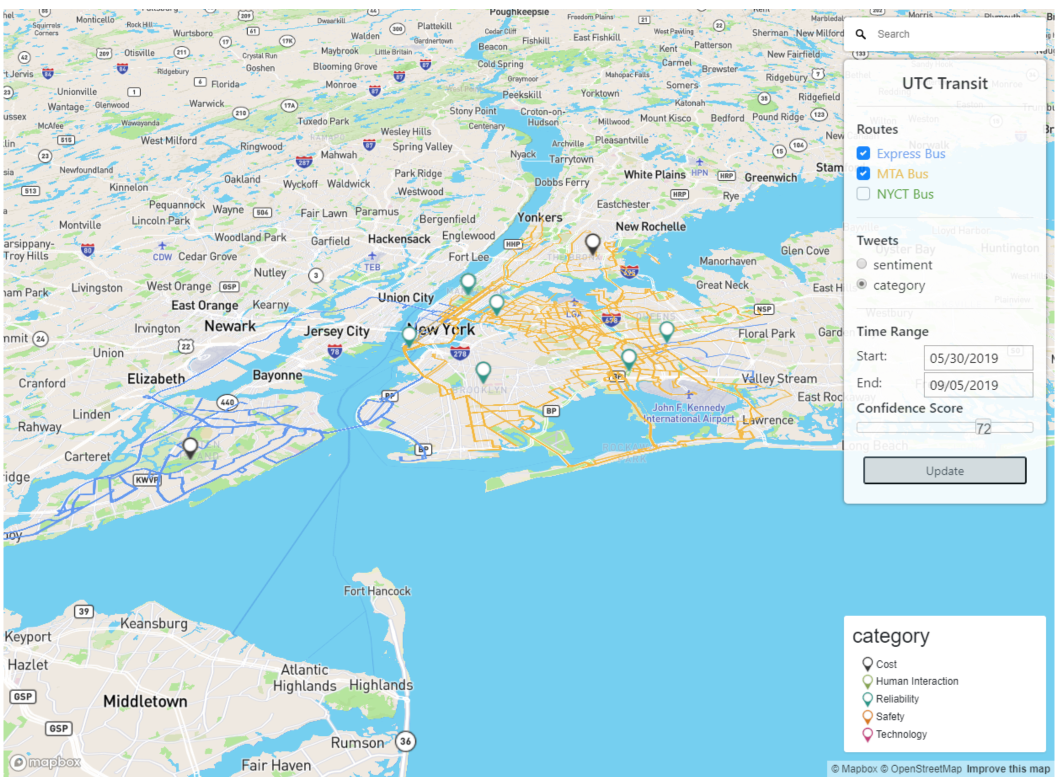

Figure7. A customized map using filters

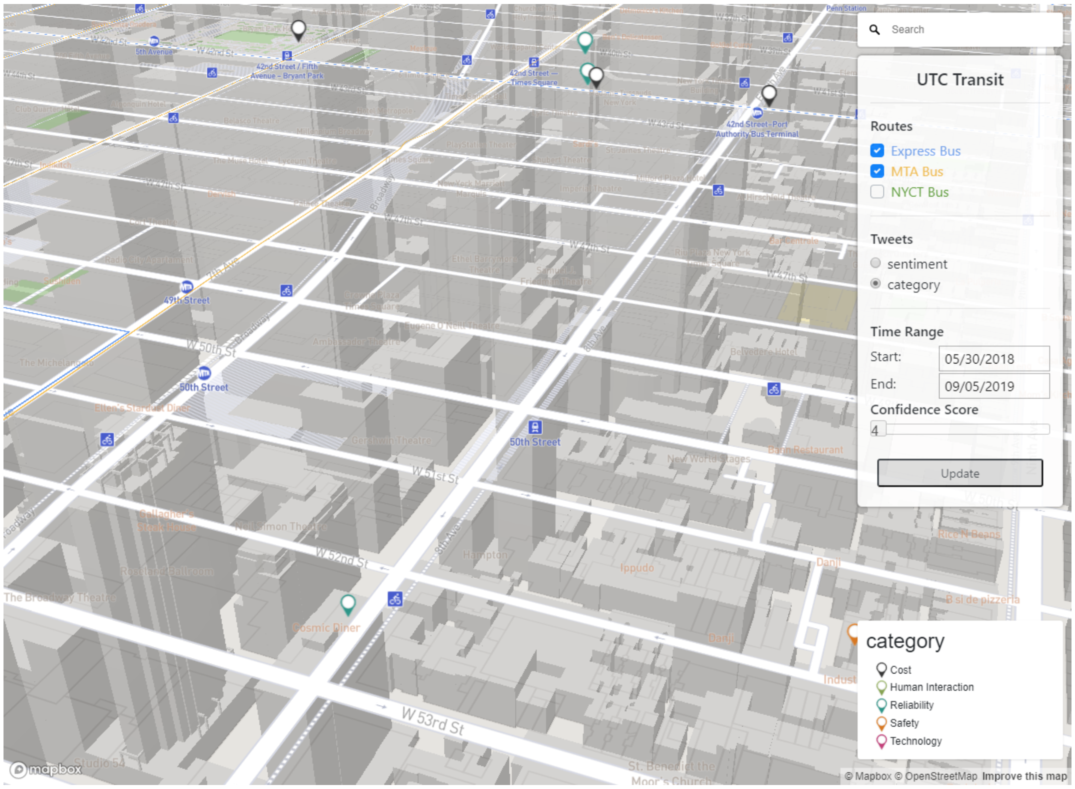

Figure 8. A 3D map by zooming in

Chapter 6: Conclusion

Social media fosters communication between individuals. Transportation planners have begun to actively adopt the social media to directly engage with transportation users to understand their barriers and challenges to use transportation service and obtain suggestions to enhance the systems. Since social media produces enormous amount of feedbacks every day, stakeholders' needs to use big data from social media has grown in the recent years, especially when they conduct a system evaluation for transportation planning. This study used text data from a social media platform, i.e., Twitter, to understand users' perception and experiences on ride-hailing system by developing the novel deep learning algorithm, MKCNN. This study used five performance measures including cost, human interaction, reliability, safety, and technology, based on the most frequent texts that ride-hailing users posted to their tweets. This study particularly added technology measure to track if users have any issues in using their smart phone app in making payment or booking a trip. The developed model produced over 65.3% accuracy for cost, human interaction and reliability measures, with overall 94% and 47% accuracies for training and testing dataset, respectively. The study found that users mainly discussed their concerns or challenges for single or multiple performance categories. Notably, many users described negative human behavior and reliability in the same tweet, which may cause higher misclassification rates between those categories in the model. The pre-trained model is deployed on a GIS map that any public can interactively experience the classification results on transportation related tweets from NY metropolitan area.

References

AIZAWA, A. 2003. An information-theoretic perspective of tf–idf measures. *Information Processing & Management,* 39**,** 45-65.

ANDREASSEN, T. W. 1995. (Dis) satisfaction with public services: the case of public transportation. *Journal of Services Marketing*.

BREGMAN, S. 2012. *Uses of social media in public transportation*, Transportation Research Board.

CARUANA, G., LI, M. & LIU, Y. 2013. An ontology enhanced parallel SVM for scalable spam filter training. *Neurocomputing,* 108**,** 45-57.

CARUANA, G., LI, M. & QI, M. A MapReduce based parallel SVM for large scale spam filtering. 2011 eighth international conference on fuzzy systems and knowledge discovery (fskd), 2011. IEEE, 2659-2662.

COLAS, F. & BRAZDIL, P. Comparison of SVM and some older classification algorithms in text classification tasks. IFIP International Conference on Artificial Intelligence in Theory and Practice, 2006. Springer, 169-178.

COLLINS, C., HASAN, S. & UKKUSURI, S. V. 2013. A novel transit rider satisfaction metric: Rider sentiments measured from online social media data. *Journal of Public Transportation,* 16**,** 2.

DRUCKER, H., WU, D. & VAPNIK, V. N. 1999. Support vector machines for spam categorization. *IEEE Transactions on Neural networks,* 10**,** 1048-1054.

GAL-TZUR, A., GRANT-MULLER, S. M., KUFLIK, T., MINKOV, E., NOCERA, S. & SHOOR, I. 2014. The potential of social media in delivering transport policy goals. *Transport Policy,* 32**,** 115-123.

GOLDBERG, Y. & LEVY, O. 2014. word2vec Explained: deriving Mikolov et al.'s negative-sampling word-embedding method. *arXiv preprint arXiv:1402.3722*.

HANNA, J. B. & DREA, J. T. 1998. Understanding and predicting passenger rail travel: an empirical study. *Transportation Journal,* 38**,** 38-46.

HAYKIN, S. & VAN VEEN, B. 2007. *Signals and systems*, John Wiley & Sons.

HENAO, A. & MARSHALL, W. E. 2018. The impact of ride-hailing on vehicle miles traveled. *Transportation***,** 1-22.

HOCHREITER, S. & SCHMIDHUBER, J. 1997. Long short-term memory. *Neural computation,* 9**,** 1735-1780.

JOACHIMS, T. Text categorization with support vector machines: Learning with many relevant features. European conference on machine learning, 1998. Springer, 137-142.

JOULIN, A., GRAVE, E., BOJANOWSKI, P. & MIKOLOV, T. 2016. Bag of tricks for efficient text classification. *arXiv preprint arXiv:1607.01759*.

KAN, M.-Y. & THI, H. O. N. Fast webpage classification using URL features. Proceedings of the 14th ACM international conference on Information and knowledge management, 2005. ACM, 325-326.

KAVANAUGH, A. L., FOX, E. A., SHEETZ, S. D., YANG, S., LI, L. T., SHOEMAKER, D. J., NATSEV, A. & XIE, L. 2012. Social media use by government: From the routine to the critical. *Government Information Quarterly,* 29**,** 480-491.

KIM, Y. Convolutional neural networks for sentence classification. EMNLP, 2014.

KITTELSON, ASSOCIATES, PROGRAM, T. C. R. & CORPORATION, T. D. 2003. *Transit capacity and quality of service manual*, Transportation Research Board.

KOUSHKI, P., AL-SALEH, O. & AL-LUMAIA, M. 2003. On management's awareness of transit passenger needs. *Transport Policy,* 10**,** 17-26.

LEE, J. Y. & DERNONCOURT, F. 2016. Sequential short-text classification with recurrent and convolutional neural networks. *arXiv preprint arXiv:1603.03827*.

MCCALLUM, A. & NIGAM, K. A comparison of event models for naive bayes text classification. AAAI-98 workshop on learning for text categorization, 1998. Citeseer, 41-48.

METSIS, V., ANDROUTSOPOULOS, I. & PALIOURAS, G. Spam filtering with naive bayes-which naive bayes? CEAS, 2006. Mountain View, CA, 28-69.

OWYANG, J., LOVETT, J., PETERSON, E., LI, C. & TRAN, C. 2010. Social marketing analytics: A new framework for measuring results in social media. *Altimeter Group & Web Analytics Demystified, San Mateo*.

PANG, B. & LEE, L. A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts. Proceedings of the 42nd annual meeting on Association for Computational Linguistics, 2004. Association for Computational Linguistics, 271.

PANG, B., LEE, L. & VAITHYANATHAN, S. Thumbs up?: sentiment classification using machine learning techniques. Proceedings of the ACL-02 conference on Empirical methods in natural language processing-Volume 10, 2002. Association for Computational Linguistics, 79-86.

PENNINGTON, J., SOCHER, R. & MANNING, C. Glove: Global vectors for word representation. Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP), 2014. 1532-1543.

PRATT, A. N., MORRIS, E. A., ZHOU, Y., KHAN, S. & CHOWDHURY, M. 2019. What do riders tweet about the people that they meet? Analyzing online commentary about UberPool and Lyft Shared/Lyft Line. *Transportation research part F: traffic psychology and behaviour,* 62**,** 459-472.

RAMOS, J. Using tf-idf to determine word relevance in document queries. Proceedings of the first instructional conference on machine learning, 2003. Piscataway, NJ, 133-142.

RYUS, P. 2003. A summary of tcrp report 88: A guidebook for developing a transit performance-measurement system. *TCRP Research Results Digest*.

SAK, H., SENIOR, A. & BEAUFAYS, F. Long short-term memory recurrent neural network architectures for large scale acoustic modeling. Fifteenth annual conference of the international speech communication association, 2014.

SCHNEIDER, K.-M. A comparison of event models for Naive Bayes anti-spam e-mail filtering. Proceedings of the tenth conference on European chapter of the Association for Computational Linguistics-Volume 1, 2003. Association for Computational Linguistics, 307-314.

THEVATHASAN, A. & BALACHANDRAN, B. Customers' perceptions of metropolitan train services in Melbourne. 30th Australasian Transport Research Forum. Melbourne, Australia: Australasian Transport Research Forum. Available from:/http://www. patrec. org/web_docs/atrf/papers/2007/1584_Balachandran, Thevathasan_Customers% 20Perceptions% 20of% 20Metropolitan% 20Train% 20Service s% 20in% 20Melbourne. pdfS, 2007.

TONG, S. & KOLLER, D. 2001. Support vector machine active learning with applications to text classification. *Journal of machine learning research,* 2**,** 45-66.

VINODHINI, G. & CHANDRASEKARAN, R. 2012. Sentiment analysis and opinion mining: a survey. *International Journal,* 2**,** 282-292.

WANG, Y., HUANG, M. & ZHAO, L. Attention-based LSTM for aspect-level sentiment classification. Proceedings of the 2016 conference on empirical methods in natural language processing, 2016. 606-615.

WILLIAMS, R. J. & ZIPSER, D. 1989. A learning algorithm for continually running fully recurrent neural networks. *Neural computation,* 1**,** 270-280.

YIN, W., KANN, K., YU, M. & SCHÜTZE, H. 2017. Comparative study of CNN and RNN for natural language processing. *arXiv preprint arXiv:1702.01923*.

ZHOU, C., SUN, C., LIU, Z. & LAU, F. 2015. A C-LSTM neural network for text classification. *arXiv preprint arXiv:1511.08630*.

ZHOU, P., QI, Z., ZHENG, S., XU, J., BAO, H. & XU, B. 2016. Text classification improved by integrating bidirectional LSTM with two-dimensional max pooling. *arXiv preprint arXiv:1611.06639*.

ZULKARNAIN, Z., SURJANDARI, I. & WAYASTI, R. A. Sentiment Analysis for Mining Customer Opinion on Twitter: A Case Study of Ride-Hailing Service Provider. 2018 5th International Conference on Information Science and Control Engineering (ICISCE), 2018. IEEE, 512-516.

CTEDD

CENTER FOR TRANSPORTATION EQUITY, DECISIONS & DOLLARS

The **Center for Transportation, Equity, Decisions and Dollars (CTEDD)** is a USDOT University Transportation Center, leading **transportation policy research** that aids in **decision making** and **improves economic development** through **more efficient, and cost-effective use of existing transportation systems,** and offers **better access to jobs and opportunities.** We are leading a larger consortium of universities focused on **providing outreach** and **research to policy makers**, through **innovative methods** and **educating future leaders of the transportation field.**

UNIVERSITY OF TEXAS ARLINGTON

CAL POLY SAN LUIS OBISPO

Georgia Tech

USF UNIVERSITY OF SOUTH FLORIDA

WISCONSIN UNIVERSITY OF WISCONSIN–MADISON

C-TEDD@UTA.EDU CTEDD.UTA.EDU @UTACTEDD @C__TEDD @ CENTER FOR TRANSPORTATION, EQUITY, DECISIONS AND DOLLARS 817.272.5138