A Real-Time Twitter Data Mining Approach to Infer User Perception towards Active Mobility

Rezaur Rahman
Ph.D. Student
Department of Civil, Environmental, and Construction Engineering
University of Central Florida
12800 Pegasus Drive, Orlando, FL 32816
Email: rezaur.rahman@knights.ucf.edu

Kazi Redwan Shabab
Ph.D. Student
Department of Civil, Environmental, and Construction Engineering
University of Central Florida
12800 Pegasus Drive, Orlando, FL 32816
Email: redwanshabab@knights.ucf.edu

Kamol Chandra Roy
Ph.D. Student
Department of Civil, Environmental, and Construction Engineering
University of Central Florida
12800 Pegasus Drive, Orlando, FL 32816
Email: roy.kamol@knights.ucf.edu

Mohamed H. Zaki, Ph.D.
Assistant Professor
Department of Civil, Environmental, and Construction Engineering
University of Central Florida
12800 Pegasus Drive, Orlando, FL 32816
Email: mzaki@ucf.edu

Samiul Hasan, Ph.D. (Corresponding Author)
Assistant Professor
Department of Civil, Environmental, and Construction Engineering
University of Central Florida
12800 Pegasus Drive, Orlando, FL 32816
Email: samiul.hasan@ucf.edu

[This is the pre-print version of the following article: Rahman, Shabab, Roy, Zaki, and Hasan (2021). A Real-Time Twitter Data Mining Approach to Infer User Perception towards Active Mobility.Transporttation Research Record, which has been published in final form at https://doi.org/10.1177/03611981211004966*].*

ABSTRACT

This study evaluates the level of service of shared facilities through mining geotagged data from social media and analyzing the road-users' perception. In pursuit of this objective, we develop an algorithm adopting a text classification approach with contextual understanding to filter out relevant information related to user perception towards active mobility. Using a heuristic-based keyword matching approach produces about 75% out-of-context tweets. Hence this approach is deemed unsuitable for information extraction from Twitter. We implement six different text classification models and compare the performance of these models for tweet classification. We apply the model on real-world data to filter out relevant information and perform content analysis to check the distribution of keywords inside the filtered data. From the experiment, we find that the Term Frequency-Inverse Document Frequency (TF-IDF) vectorizer based logistic regression model (one of the six text classification models) performs best in classifying the tweets. To select the best model, we compared the performance of the models based on precision, recall, f1 score, and accuracy metrics. The findings from the analysis show that the proposed method can help produce more relevant information on walking and biking facilities as well as safety concerns. By analyzing the sentiment of the filtered data, we can infer the existing condition of biking and walking facilities in the D.C. area. This method can be a critical part of the decision support system to understand the qualitative level of service of existing transportation facilities.

1. INTRODUCTION

Widespread use of social media platforms such as Facebook, Twitter, Strava offers a unique opportunity to collect real-time information on existing transportation facilities in a cost-effective way. It is more relevant, especially when traditional data collection approaches such as travel demand surveys, user perception surveys are costly and time-consuming. Social networking sites facilitate users to share their daily activities, travel patterns, perceptions, and sentiments as small messages or posts. Such information is vital when managing infrastructure, traffic operation, and demand (1). It also encourages researchers and practitioners to explore the capacity of social media data in real-time traffic information sharing (2), travel behavior modeling (3,4), and qualitative service quality analysis of existing transportation facilities (5). In this study, we leverage social media data to create a framework to understand road user perception towards shared active transportation facilities.

Although active mobility is one of the significant components for developing sustainable transportation infrastructure, traditional transportation planning focuses on improving conditions of highways for car owners at the cost of safe sidewalks and bike facilities. Consequently, limited resources have been deployed in this sector to develop smart tools to understand pedestrian and bikers' safety and mobility concerns. In 2018 total federal funding on walking and biking was $916 million, which is only 2% of the total federal funding in transportation infrastructure (6,7), merely enough to encourage more people to use active transportation mode (6). Hence, this study's main objective is to propose a low-cost solution to engage more with the population and extract information regarding their perception towards active mobility, including conditions of existing facilities and safety concerns.

Moreover, city planners and transportation managers need to know the condition of existing facilities to develop a strategic plan for facility improvements. Transportation agencies mostly rely on qualitative perception surveys and quantitative analysis such as pedestrian volume count, number of bikers, and number of crashes to understand overall conditions for active mobility. Such data collection approaches require substantial resources, discouraging agencies from continuous real-time monitoring of shared space transportation facilities. Hence, we need a cost-effective alternative for real-time monitoring of the existing shared facilities.

In the recent past, social media platforms have gained popularity by allowing users to share their thoughts and concerns. Twitter is a notable example, with more than 330 million subscribed users. It is a microblogging service used to share views, activities, and thoughts through a 280-character message known as a ‘tweet.’ In the United States, Twitter is one of the most widely used social media platforms, with more than 67 million active users (8). Many transportation agencies (e.g., state DOTs) use Twitter to share real-time information to travelers, such as information related to traffic congestion, crashes, incidents, and planned road work. Users also share their views and concerns on existing transportation facilities via tweets. This information can be utilized to understand the overall condition of a transportation facility in a qualitative way. Hence, Twitter data has the potential to support decision support tools assisting transportation managers in understanding user's perception of existing facilities with respect to safety and service quality.

However, the flexibility of information sharing on social media platforms has created a significant challenge to extract relevant information related to specific content; most of the time, these social networking sites get flooded with random information, making it challenging to extract task-specific information. In recent years advances in natural language processing technologies create an opportunity to overcome these challenges to extract relevant information from social media data. Thus, this study's main objective is to develop a framework to extract information related to service quality and safety issues of biking and walking facilities from geotagged Twitter data. The study implements a systematic framework for Twitter data mining and text analysis to understand user perception towards active mobility. The research is motivated by three key prospects: 1) widespread use of social media platforms, 2) real-time data collection techniques, and 3) advances in natural language processing (NLP) technologies.

The proposed framework includes a three-step tweet filtering process; at first, we apply geolocation-based boundaries to filter out tweets for a specific region, later we apply a heuristic-based screening technique to filter out geotagged tweets based on specific keywords related to walking and biking. However, the heuristic-based screening technique fails to filter out the most relevant tweets related to a specific context, such as user perception towards walking and biking facilities; a higher proportion of the tweets provides random information. To overcome this challenge, we implement a text classification method based on the tweet context to filter out the most relevant information. As a final step of the tweet filtering approach, we apply a text classification model to filter out the most relevant tweets on active transportation facilities. We also validate our approach by analyzing the contents of each tweet; we apply the Latent Dirichlet Allocation (LDA) model on the filtered tweets to infer a high-level summary of users' thoughts on walking and biking conditions. Finally, we conduct sentiment analysis to understand the polarity of users' sentiments for existing active transportation facilities. The framework we proposed is based on real-time monitoring of the Twitter feed. Once the framework is deployed for real-world applications, all these operations can be completed in real-time. If a user posts any tweet related to a walking or biking facility, the proposed algorithm will collect this information and show the tweet's polarity and content in real-time.

Overall, based on our understanding of existing literature, we anticipate that this study will have three major contributions to existing literature and practices. First, it develops a new approach to collect information on user perception towards active mobility cost-effectively; second, it demonstrates qualitative tweet analysis technique to understand the level of service of an existing facility based on users' sentiments; third, it provides experimental evidence of the validity of the proposed method using geolocated Twitter data. Since the implemented approach identifies users' safety concerns at different locations, the proposed framework can be an alternative approach for near real-time monitoring of the qualitative level of service of existing active transportation facilities in terms of service quality and safety.

2. LITERATURE REVIEW

Social media offers an open-access platform for people to share their opinions about different issues (5) instantly. Information from these passive data sources establishes an alternative way to understand user perceptions towards existing transportation facilities such as availability and quality of sidewalks, bike lanes, safety concerns, etc. However, the raw data collected from Twitter are extremely noisy: flooded with random topics, similar keywords used in different contexts (9). Even though social media data provides a massive volume of information on user opinions, this information is meaningless unless we can extract the relevant information related to a specific topic. Traditional keyword-based filtering algorithms commonly handle text as straightforward successions of character strings; they only search if a given set of keywords is present in a sentence regardless of the context (10). These methods cannot extract context-wise information from Twitter; hence we need a robust context-wise text classification approach to overcome this challenge.

Text classification, one of the fundamental tasks in Natural Language Processing (NLP), is categorizing text according to its content. Text classification has widely been used for topic labeling, spam detection, and intent detection. Existing text classification methods can be divided into a traditional machine learning approach and a deep learning approach. Naïve Bayes (11), logistic regression (12,13), and support vector machine (14) are the most commonly used machine learning approaches for text classification. Naïve Bayes is commonly used as a standard for text classification since it is quick and simple to execute (15). It assumes all attributes of the class as an individual element, and this pattern simplifies the classification of the text (15). However, when the training data is noisy and small, Bayesian learning is not practical for text classification (16).

The logistic regression-based multiclass text classification has shown superior performance compared to other traditional approaches (13). This algorithm assigns weights to each input sequence to segregate potential classes from each other (17). However, Logistic Regression assumes that all the input features in the dataset are independent, which lowers the precision of text classification for a dependent set of variables (18). Support Vector Classification (SVC) works well for high dimensional features in texts (14); however, it takes a substantial amount of time to tune the parameters for SVC algorithms to improve the precision (19). Several studies (20) have also applied tree-based classifiers for text classification. Although these algorithms work well with categorical features, they are susceptible to a small perturbation in the data set and suffer from overfitting issues.

In the recent past, deep learning approaches have gained more attention due to its ability to deal with high dimensional data. Convolutional neural network (CNN) and recurrent neural network (RNN) are the two most commonly used deep learning methods for text classification. Although the CNN architecture was built for image processing, it has been successfully applied in text classification. However, CNN performs poorly for long sequences of text due to a limited capacity to learn consecutive connections (21). For long sequences of text RNN based classification models such as Long Short-Term Memory Neural Network (LSTM), Gated Recurrent Unit (GRU), Bidirectional LSTM (BiLSTM) have shown better performance (18,22–24).

One of the limitations of RNN based classification is that it becomes biased when later words are more influential than earlier ones for a sequence of texts. To overcome this issue, a CNN layer is introduced with the RNN architecture (25); Convolutional LSTM (convLSTM) model utilized the CNN model to extract a sequence of higher-level phrase representations and are fed into an

LSTM model. The ConvLSTM model captures both local features of sentences as well as global and temporal sentence semantics.

Overall a vast number of studies have used NLP to improve the efficiency of existing text classification and content analysis methods. These approaches can help us overcome the challenges of utilizing social media data for transportation planning and traffic management. In transportation research, text classification approaches are mostly applied for traffic incident detection from social media data (26–28); these studies apply a binary classification approach to separate the incident related text. Apart from incident detection, NLP has also been applied to infer weather-related events from social media data (29).

Although social media platforms generate massive information on user perceptions and sentiments related to different transportation facilities such as public transportation (30), shared mobility active transportation facilities (31), few studies have explored the capacity of social media data for real-time monitoring of qualitative service quality of these transportation facilities based on users perceptions and sentiments. Moreover, a few studies (31,32) explored the impact of user opinions and sentiments on social media to encourage sustainable mobility options such as biking, public transit. However, these studies are limited to social media data exportation; they do not address the challenges in collecting context-wise real-time information from social media posts.

In this study, we implement a framework based on Twitter data mining to analyze the qualitative level of service for active mobility to overcome this research gap. We implement advanced text classification approaches to extract the most relevant information from Twitter posts based on the context of the texts. We also perform sentiment analysis to represent users' polarity towards active transportation facilities. The proposed framework offers a new approach for evaluating the qualitative level of service of existing facilities for active mobility in terms of service quality and safety.

3. RESEARCH METHODOLOGY

This study proposes a framework to extract information on user perceptions towards active mobility options using real-time Twitter streaming data. The methodology is shown in Figure 1 and consists of three steps. First, we apply a geolocation boundary to collect Twitter data for a specific zone, followed by a keyword matching based searching approach to identify the relevant tweets related to active mobility. A context-based text classification approach is then applied to prune out tweets that include some information that contains keywords closely related to walking and biking but out of the context of mobility and transportation.

In the tweet filtering process, the keyword matching algorithm uses relevant keywords related to walking and biking (e.g., walk, bike, sidewalk etc.) to collect active mobility related tweets. The set of keywords assigned to the algorithm will control the number of collected tweet samples closely related to active transportation mode. In some cases, we might not get enough information because of missing some relevant keywords. Therefore, it is important to ensure feedback control in the Twitter filtering process depending on the outcome (Figure. 1.). The feedback control process will help check the collected information if the data collection approach is missing some relevant information related to active transportation facilities.

In the next step, we adopt a topic modeling approach to perform content analysis over-filtered tweets and generate clusters of topics related to active transportation facilities—the distributed keywords inside each cluster highlight user's activity, perceptions, and concerns at a higher level. By analyzing the topics and keywords inside each topic, we can identify the user concerns regarding active transportation facilities. Finally, we perform a sentiment analysis over the filtered tweets to understand users' polarity (a score representing whether a text is positive, neutral, or negative) towards existing walking and biking facilities. The sentiment analysis approach provides polarity metrics for users' perceptions.

The following section explains the different components of the proposed method, starting with model selection for text classification.

3.1 Tweet Classification

One of the critical components of the proposed framework is to implement a model that can filter out the most relevant information from the social media platform based on context of the tweets, which means the accuracy of the model will decide relevance of the collected information to a specific topic. In this study, we will implement multiple text classification models and will compare their performance in classification accuracy. However, to proceed with model implementation, at first, we need to extract the feature vectors from each tweet.

Extracting Feature Vectors

We use a vectorizer (33,34) to convert the tweet texts into a sparse matrix that consists of numbers or tokens. The size of the matrix depends on the vocabulary size if the vocabulary size is not given, the vectorizer estimates the vocabulary size by analyzing the data. So, the main function of the vectorizer is to convert the tweet texts as a vectorized input for the models. We also use a Term Frequency-Inverse Document Frequency (TF-IDF) (35) score to estimate the importance of different words inside a tweet. The TF-IDF of a word increases with an increase in frequencies but decreases if the word is present in many documents (e.g., stop words, punctuations); a high TF-IDF score of a word implies high importance within a collection of documents (tweets). The equation for calculating TF-IDF for a word (i) is as follows:

$$TF - IDF(i) = TF(i) \times IDF(i) \tag{1}$$

where

$$TF(i) = \frac{the\ number\ of\ times\ the\ word\ i\ appears\ in\ the\ document}{the\ total\ number\ of\ terms\ in\ the\ document} \tag{2}$$

$$IDF(i) = \log_{\mathrm{e}}\left(\frac{total\ number\ of\ documents}{number\ of\ documents\ with\ term\ i}\right) \tag{3}$$

We use both unigram and bigram of words to create feature vectors. The details about unigram and bigram are available in (36,37). Moreover, to remove the effect of total word counts in a document, we apply $l2$ normalization (sum of the squared value of TF-IDF = 1 for a document).

Model Selection for Text Classification

We adopt a multiclass classification approach to identify the tweets related to different categories of active mobility options (e.g., walking, biking). The objective is to find the best model to map the tweets into different categories based on the context. Let, $\mathcal{F}$ denotes the function that maps input tweets (X_m) into different categories (Y_m),

$$\mathcal{F}(X_m) \rightarrow Y_m \tag{4}$$

Here, m indicates the number of data samples; $X_m$is the input TF-IDF vector created in the previous step; $\mathrm{Y_m}$ is a vector that contains the labels $(y = i)$ for each category of tweets, i denotes one class out of our three classes. We consider three categories of tweets: walking related, biking related, and other random out of context tweets. Thus, the target vector (Y) has three labels where$, y \in \{"0": "walking", "1": "biking", "2": "other"\}$.

To select the best model, we check the predictive performance of each model in terms of precision, recall, f1 score, and accuracy. We generate a confusion matrix to estimate these performance measures. The confusion matrix also reveals the performance imbalance of a classifier: high accuracy for a class but low for another. Table 1 shows the components of a confusion matrix. The rows represent the actual labels, and the columns represent the predicted labels where positive means the existence of a particular label and negative means the absence of a particular label. For a given tweet, if the actual label is negative, a negative prediction by the model is assigned as true negative, and a positive prediction is assigned as false positive. Similarly, if the actual label is positive, a positive prediction is assigned as a true positive, and a negative prediction is assigned as a false negative.

$$\text{Precision} = \frac{TP}{TP+FP} \tag{5}$$

$$Recall = \frac{TP}{TP+FN} \tag{6}$$

$$F1\ Score = \frac{2 * Recall * precision}{Recall + precision} \tag{7}$$

$$Accuracy = \frac{\text{TP+TN}}{\text{TP+FP+FN+TN}} \tag{8}$$

The selection of the best classification model is challenging when the model performance changes for different performance metrics. In some cases, a classifier can show better accuracy, but a higher false-positive rate will cause a lower precision score. In the case of tweet classification, how accurately the model identifies the true positive cases is important. So, precision and recall are two essential criteria to evaluate model performance. However, we analyze all the metrics to develop an optimal model that shows consistent performance for each performance metric.

3.2 Topic Model for Content Analysis

To recognize the content of a posted tweet, we apply the Latent Dirichlet Allocation (LDA) or topic modeling approach *(3)*. The topic model specifies a probabilistic procedure of generating documents; the process starts with choosing a distribution over topics, and for each word in the document, a topic is randomly selected from the chosen distribution. Finally, a word is randomly drawn from that topic (39). The process can infer the set of topics responsible for generating a collection of documents (i.e., tweets) by applying standard statistical techniques inverting the process. The topic model has been widely used in machine learning; it has been recently used in transportation studies (40–44).

We implement the model using the genism library (45,46) in Python. We train the model over a sample of tweets to generate the topics and distribution of keywords inside each topic. However, generating meaningful topics from random tweet samples is more challenging; we need to decide on the optimal number of topics to prevent the repetition of similar topics and keyword distributions. To overcome this challenge, we use the coherence score metric (47), which measures the degree of semantic similarity (e.g., conceptually correlated words) between high scoring words within a topic. Thus, it helps differentiate between the semantically interpretable topics and the topics that are artifacts of statistical inference. A higher value of the coherence score indicates that the words in a topic are semantically relatable, so the topic has suitable interpretability. In our study, we adopt "topic coherence pipeline" (48,49) to estimate the aggregated coherence score for the topic model. To choose the optimal number of topics, we run the topic model with a different number of topics and estimate the aggregated coherence score. Finally, we select the optimal number of topics based on the maximum coherence score.

Once we find the optimal model, we can use it for topic generation and content analysis. The distributed keywords inside each topic will provide insights into user perceptions and concerns. So, rather than going over each topic, we can understand users' opinions by interpreting the topic keywords. Moreover, we use the trained model to find tweets closely related to a particular topic (e.g., safety concern, bike facility, etc.).

3.3 Sentiment Analysis

Sentiment analysis has widely been used to understand user perception of different products or facilities (50). In this study, we adopt a sentiment analysis approach to understand the user's experience of different active transportation facilities. To analyze the sentiment for each tweet, we use python's "vaderSentiment" library (51). VADAR (Valence Aware Dictionary for Sentiment Reasoning) is a pre-trained classification model that uses a rule-based approach to classify a text as positive, negative, and neutral. The model is being trained over social media texts and emoji, thus suitable for tweet analysis. We apply the pre-trained model to get the compound score (polarity of a text) for each tweet, which varies between -1 to 1. We categorize a tweet as positive if the compound score >= 0.05, neutral if the compound score is between -0.05 and 0.05, negative if the compound score <= -0.05.

4. DATA COLLECTION AND PRE-PROCESSING

In this study, we use Twitter data collected from the Washington, DC area using Twitter's streaming API (application programming interface), giving a geolocation boundary (Figure 2(a)). To collect the information on walking and biking, we filtered the tweets with relevant keywords (Table 2). In total, we have collected 3,533 tweets from October 7, 2019, to November 7, 2019. We check the duplicate entries for all the tweets based on tweet ids and remove all the duplicate tweets; the final dataset has in total of 3,273 tweets. We also check the number of unique users and find that the dataset has 2307 unique users, among which about 77.6 % tweeted only once between October 07, 2019 and November 07, 2019 (Figure 2 (b)). Only three users posted more than 30 tweets within this period, which means that in the data sample, majority of the users are occasional users. Figure 2(b) shows the distribution of users based on the number of tweets posted in the study period.

To create an annotated dataset, we manually labeled all of 3,273 tweets. To ensure that we retrieve the right labels of the tweets, we independently (three annotators) labeled each tweet and then matched the labels from different annotators to fix the final label. Each tweet can have at least one label out of three possible categories: walking related, biking related, and others. Figure 2(c) shows the distribution of different types of tweets. Although we apply a heuristic approach to remove irrelevant tweets from Twitter, we find that only 25% of the tweets are related to walking and biking; the rest contain random posts that involve words such as walk, bike, etc.

To understand the content of the collected data, we run a generalized topic model over the collected tweet samples. We estimate the coherence score for different numbers of topics and determine the optimal number of topics based on the maximum value of coherence. Figure 3 (a) shows that the maximum value of coherence (0.53) is obtained for ten topics. So, we run the final model with 10 topics (Figure 3 (b)). From the topic analysis, we find that few topics, such as topic#7 include keywords like "bike," "good," "ride," "well," indicating the user's perception of bike rides. Topic#6 also includes keywords related to bike and bike lanes. However, most of the topics seem irrelevant to the context and provide random information. Hence, we need further cleaning of the data to reduce the flooding of random information.

5. RESULTS AND DISCUSSIONS

5.1 Tweet Classification

To classify the tweets, we implement six different classification models: Naïve Bayes, Logistic Regression, Support Vector Classification (SVC), Long-Short Term Memory Neural Network (LSTM), Bidirectional LSTM, and convolutional LSTM. To train the model, we divide the data into train, test and validations set. We use 70% of the data to train the model and 15% of the data to test the model while tuning the parameters. Rest 15% of the data is used for validating the proposed approach for text classification and content analysis.

Before training the models, we apply both unigram and bigram based "CountVectorizer" (33) from scikit learn library (52) to create feature vectors from the tweets, however for this particular problem unigram based vectorizer performed best. The vectorizer generates a vectorized sparse matrix that includes 7070 features per tweet text, which means the vocabulary size of the tweet samples is 7070. We use this vectorized sparse matrix to estimate the TF-IDF score for each of the words using "TfidfTransformer" (53,54). In our final data set, we represent the tweet text as a vectorized matrix consisting of the TFIDF score of each word, which is then directly fed into the model as input.

In this experiment, we choose multinomial Naïve Bayes as our base model. We explored different values for the smoothing parameter alpha for the Naïve Bayes model; however, there is no significant increase in model accuracy. In the case of the logistic regression model, we do not have any open parameters to tune. The number of parameters for the logistic regression model is the same as the input size (7070). The SVM classifier is more laborious than the logistic regression model. For the SVM classifier, we explore three different kernels: linear, polynomial, and radial basis function. We also vary the penalty parameter ranging from 100 to 10000. We obtain the best result for both linear and radial basis functions with the value of penalty parameters as 1000.

In the case of deep learning models, we randomly searched different set of hyperparameters to obtain the optimal combination of hyperparameters that works best. However, none of them perform very well compared to the base model. Moreover, since we generate the feature matrix using vectorizer with TF-IDF, the dimension of features (7070) is higher. As such, it increases the model complexity by increasing the number of neurons at the input layers. Hence, most of the time, these complex models overfit and performed well on training data but not on the test data.

We apply another approach by using a simple “Tokenizer”(34) from Keras library (55) and a word embedding layer for the deep learning model, which significantly reduces the training time while increasing the model accuracy. For the embedding layer, we use a vocabulary size of 7106, which is obtained from the tokenizer. The length of the input features is 280, which is the same as the maximum allowable number of words in a single tweet. In Table 3, we report the hyperparameter sets for the deep learning model that produced the top result. The detail about each component of deep learning models for text classification can be found in (25).

Overall based on the performance measures, we find that TF-IDF based logistic regression (test accuracy=0.821) has performed better than all the other models, including the deep learning models. As shown in Table 4, both logistic regression and deep learning models performed well if we consider precision and recall measures. However, deep learning models would take more data, time, and computational power to complete the training process.

We also analyze the Receiver Operating Characteristics (ROC) curves to understand the performance of the models in terms of true positive and false positive rates. The ROC curve is based on the macro-average of the classification for each label. In the macro-average case, we compute the ROC value independently for each class and then take the average (hence treating all classes equally). ROC is a probability curve consists of a true positive rate and false positive rate, whereas the area under the curve (AUC) represents a degree or measure of separability. It explains the capability of the model to distinguish one class from another class. The higher the AUC is, the better the model performs in predicting 0 as 0 and 1 as 1. Based on the AUC value, we find that the Logistic regression model (0.890) performs better than all the other models. Figure 4 shows the ROC for the Logistic regression model over the test dataset.

5.2 Model validation and Content Analysis

We validate our model on 573 tweets, among which 121 tweets are walking related, and 20 tweets were biking related. To demonstrate the logistics regression model's performance, we estimate the confusion matrix over the validation data (Figure 5). From the confusion matrix, we find that the model can separate random tweets with 88% accuracy; the percentage of true positive rate (90%) is also very high for tweets related to biking. However, we see a higher percentage of false positive (about 38%) rate for classifying tweets related to walking. From our analysis, we find that most of the random tweets include the keywords walking or walk. However, many are out of context and unrelated to any perception of walking facilities. Such out of context tweets are difficult to separate from contextual tweets related to walking facilities; consequently, it generates a higher false positive rate. The model's overall accuracy on the validation data set is satisfactory, with 82.7% of correct classification.

We train the topic model on the filtered data (141 samples); we run the model with a different number of topics and check the corresponding coherence score. From the coherence measure, we find that the coherence score is the highest (0.57) for three topics (Figure 6a). Figure 6 (b) demonstrates the keyword distribution for each topic. From the keyword distribution, we find that users mostly address if they are enjoying walking and biking. Moreover, some keywords indicate negative aspects. For example, in topic#0, the keyword “kill” is associated with other active transportation-related keywords: walk and sidewalk. It is possible that this topic is closely related to the safety concerns of existing sidewalks.

Moreover, we apply the trained model to get the dominant topic category (i.e. topic#0, topic#1, and topic#2) for each tweet. Based on that, we group the tweet samples into topic#0, topic#1, and topic#2. Figure 7(a) demonstrates six tweets with their dominant topic category. We find that all the tweets closely related to topic#0 indicate safety issues such as pedestrian-vehicle collision while using the sidewalk or crossing the street. Moreover, we also observe a safety concern regarding e-scooters. There are also a few tweets on the benefits of existing bike facilities, such as protected bike facilities; requirements for facility improvement.

We conduct further analysis to understand user perception about walking and biking facilities. We perform sentiment analysis on the filtered tweets to obtain a tweet’s polarity: whether positive or negative. From the sentiment analysis result, we obtain the compound score. If the compound score is positive, it means the user's perception is positive, while if negative, it indicates a negative perception regarding the activity or facility. Figure 7 (b) shows the distribution of polarity scores for the classified tweets related to walking and biking. The overall distribution indicates that about 18.69 % of the walking related tweets are negative, and 61.11 % of them are neutral. This indicates that users show more positivity (38.2%) towards existing walking facilities (e.g., sidewalk, crossing, etc.) in the Washington DC area. Similarly, we observe that only 5% of the bike-related tweets are negative, which may indicate a better level of service quality for biking facilities. However, we do not have enough data sample to confirm this conclusion.

Another relevant benefit of this approach is that we can find the geolocation of the tweets; we can identify the area from where the tweets were posted. The geolocations come in two ways: exact point location and bounding box. The tweets that include the exact geo-coordinates are precise and indicate the incident or any facility's actual location. On the other hand, for tweets including location as a bounding box, we calculate the center of that bounding box to determine the location. In this case, we can understand the precision of the location by computing the diagonal distance of the bounding box. If the diagonal distance is small, the geolocation will be more precise. As shown in Figure 7(a), we map all the tweets based on their geolocation. From this information, we can identify the zones which are less safe or provide an insufficient safety measure for safe walking and biking.

6. CONCLUSION

Active mobility – mostly indicates walking and biking, one of the major components of sustainable transportation modes. The majority of the world's population relies on cycling, walking, and other forms of human-powered transport to commute to work, schools, and public transport stations (56). Moreover, active transportation promotes a healthy lifestyle, one of the most affordable and practical ways to reduce CO_2 emissions (57). Therefore, promoting walking and biking is critical for establishing people-oriented sustainable cities that are safe and equitable. One of the appropriate ways of maintaining a safer environment for active mobility is the continuous monitoring of existing infrastructure. However, traditional approaches to collect information on user satisfaction and safety concerns on active transportation modes are costly, require constant maintenance, and can be time-consuming. Addressing some of those concerns, our study leverages social media data and offers low-cost support to monitor existing active transportation facilities. The study brings forward a novel and systematic framework to overcome challenges in analyzing social media data in transportation using advanced language processing tools. The proposed method can be a cost-effective alternative to understand the qualitative level of service for different existing facilities for active movement. This paper presented different components of the framework, such as the tweet filtering mechanism, topics pattern identification, and user perception towards facilities (e.g., protected bike lanes, sidewalks). We validated the results on real-world data from Washington DC.

Some limitations would require our attention in future research. One main shortcoming of the study is the small sample (i.e., number of tweets) to carry out the analysis. A best practice is to collect the data for three months. In the future, we will continue collecting the data to analyze a larger dataset. Another urgent concern for social media data is misinformation or the spread of fake news; however, this is out of our research scope. Few existing studies (58–60) have proposed several methods to identify social media post misinformation. In our future research, we will consider these methods to filter out misinformation or fake posts.

Moreover, Twitter does not reveal the user's sociodemographic characteristics (e.g., age, gender). Hence, we cannot ensure a representative sample in terms of gender and age groups. However, according to a recently published statistic (61), the distribution of Twitter users by age group is 38% for age group 18-29, 43% for age group 30-64, and 7% for ages 65+. Thus, Twitter has a mixture of all population groups. We believe that inferring sociodemographic information of Twitter users and checking the Twitter dataset's representativeness compared to the actual population is an avenue for future research.

Another research question that arises from this study is comparing our results with real condition transportation facilities' activities. Due to limited data availability, we could not complete this analysis. However, we explore all the available active data collection methods to extract more information on active transportation facilities. Moreover, the implemented classification models need further tuning, as we find that LSTM, Bidirectional LSTM, Convolutional LSTM have reasonably high AUC values; however, they have lower accuracy values. Although we used different feature extraction methods as well as different combination of hyperparameters, TF-IDF based logistic regression method outperformed other methods. As a future research direction, we will study if users' perception towards bike facilities influences bike-sharing demands. The availability of the capital bike-sharing data would make it possible to expand on the framework presented in this paper.

AUTHOR CONTRIBUTIONS

The authors confirm contribution to the paper as follows: study conception and design: R. Rahman, KR Shabab, K.C. Roy, S. Hasan, MH. Zaki; data analysis and interpretation of results: R. Rahman, KR Shabab, KC Roy; draft manuscript preparation: R. Rahman, KR Shabab, S. Hasan, MH. Zaki. All authors reviewed the results and approved the final version of the manuscript.

REFERENCES

1. He K, Xu Z, Wang P, Deng L, Tu L. on Large-Scale Social Signals. 2016;17(9):2613–26.
2. Rahman R, Roy KC, Abdel-Aty M, Hasan S. Sharing real-time traffic information with travelers using twitter: An analysis of effectiveness and information content. Front Built Environ. 2019;5(June):1–15.
3. Rashidi TH, Abbasi A, Maghrebi M, Hasan S, Waller TS. Exploring the capacity of social media data for modelling travel behaviour: Opportunities and challenges. Transp Res Part C Emerg Technol [Internet]. 2017;75:197–211. Available from: http://dx.doi.org/10.1016/j.trc.2016.12.008
4. Liao Y, Yeh S, Jeuken GS. From individual to collective behaviours: exploring population heterogeneity of human mobility based on social media data. EPJ Data Sci [Internet]. 2019;8(1):1–22. Available from: http://dx.doi.org/10.1140/epjds/s13688-019-0212-x
5. Collins C, Hasan S, Ukkusuri S. A Novel Transit Rider Satisfaction Metric: Rider Sentiments Measured from Online Social Media Data. J Public Transp [Internet]. 2013;16(2):21–45. Available from: http://scholarcommons.usf.edu/jpt/vol16/iss2/2/
6. Buehler R, Pucher J, Bauman A. Physical activity from walking and cycling for daily travel in the United States, 2001–2017: Demographic, socioeconomic, and geographic variation. J Transp Heal [Internet]. 2020;16(September 2019):100811. Available from: https://doi.org/10.1016/j.jth.2019.100811
7. Chao EL. Budget Highlights Fiscal Year 2019 [Internet]. U.S. Department of Transportation. 2019. Available from: https://www.transportation.gov/sites/dot.gov/files/docs/mission/budget/304476/508-dot-bh2019.pdf
8. Twitter by the Numbers: Stats, Demographics & Fun Facts [Internet]. Omnicore. 2017. Available from: https://www.omnicoreagency.com/twitter-statistics/.
9. Sriram B, Fuhry D, Demir E, Ferhatosmanoglu H, Demirbas M. Short text classification in

twitter to improve information filtering. In: SIGIR 2010 Proceedings - 33rd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. 2010.

10. Rana MI, Khalid S, Akbar MU. News classification based on their headlines: A review. In: 17th IEEE International Multi Topic Conference: Collaborative and Sustainable Development of Technologies, IEEE INMIC 2014 - Proceedings. 2015.

11. Rish I. An empirical study of the naive Bayes classifier. IJCAI 2001 Work Empir methods Artif Intell. 2001;

12. Louisuille. M kantardzic. Data Mining_ Concepts, Models, Methods, and Algorithms - Mehmed Kantardzic - Google Books. 2011. p. 1–529.

13. Roy KC, Hasan S, Mozumder P. A multilabel classification approach to identify hurricane-induced infrastructure disruptions using social media data. Comput Civ Infrastruct Eng. 2020;1–16.

14. Chatterjee S, George Jose P, Datta D. Text Classification Using SVM Enhanced by Multithreading and CUDA. Int J Mod Educ Comput Sci. 2019;

15. McCallum A, Nigam K. A Comparison of Event Models for Naive Bayes Text Classification. AAAI/ICML-98 Work Learn Text Categ. 1998;

16. Qu Z, Song X, Zheng S, Wang X, Song X, Li Z. Improved Bayes Method Based on TF-IDF Feature and Grade Factor Feature for Chinese Information Classification. In: Proceedings - 2018 IEEE International Conference on Big Data and Smart Computing, BigComp 2018. 2018.

17. Raschka S. Python Machine Learning. Packt Publishing Ltd., Birmingham. - References - Scientific Research Publishing. 2015.

18. Kowsari K, Meimandi KJ, Heidarysafa M, Mendu S, Barnes LE, Brown DE. Text Classification Algorithms: A Survey. Inf. 2019 Apr;10(4).

19. Shanahan JG, Roma N. Improving SVM text classification performance through threshold adjustment. In: Lecture Notes in Artificial Intelligence (Subseries of Lecture Notes in Computer Science). Springer, Berlin, Heidelberg; 2003. p. 361–72.

20. Xu B, Guo X, Ye Y, Cheng J. An improved random forest classifier for text categorization. J Comput. 2012;7(12):2913–20.

21. Karita S, Wang X, Watanabe S, Yoshimura T, Zhang W, Chen N, et al. A Comparative Study on Transformer vs RNN in Speech Applications. In: 2019 IEEE Automatic Speech Recognition and Understanding Workshop, ASRU 2019 - Proceedings. 2019.

22. Wei F, Nguyen UT. Twitter bot detection using bidirectional long short-term memory neural networks and word embeddings. In: Proceedings - 1st IEEE International Conference on Trust, Privacy and Security in Intelligent Systems and Applications, TPS-ISA 2019. Institute of Electrical and Electronics Engineers Inc.; 2019. p. 101–9.

23. Mousa AED, Schuller B. Contextual bidirectional long short-term memory recurrent

neural network language models: A generative approach to sentiment analysis. In: 15th Conference of the European Chapter of the Association for Computational Linguistics, EACL 2017 - Proceedings of Conference. 2017.

24. Zhou P, Qi Z, Zheng S, Xu J, Bao H, Xu B. Text classification improved by integrating bidirectional LSTM with two-dimensional max pooling. In: COLING 2016 - 26th International Conference on Computational Linguistics, Proceedings of COLING 2016: Technical Papers. 2016.

25. Zhou C, Sun C, Liu Z, Lau FCM. A C-LSTM Neural Network for Text Classification. 2015;

26. Zhang Z, He Q, Gao J, Ni M. A deep learning approach for detecting traffic accidents from social media data. Transp Res Part C Emerg Technol. 2018;86:580–96.

27. Zhang S. Using Twitter to Enhance Traffic Incident Awareness. IEEE Conf Intell Transp Syst Proceedings, ITSC. 2015;2015-Octob(71403068):2941–6.

28. Yazici MA, Mudigonda S, Kamga C. Incident detection through twitter: Organization versus personal accounts. Transp Res Rec. 2017;2643(1):121–8.

29. Lin L, Ni M, He Q, Gao J, Sadek AW. Modeling the Impacts of Inclement Weather on Freeway Traffic Speed. Transp Res Rec J Transp Res Board [Internet]. 2015;2482:82–9. Available from: http://trrjournalonline.trb.org/doi/10.3141/2482-11

30. Pender B, Currie G, Delbosc A, Shiwakoti N. Social Media Use during Unplanned Transit Network Disruptions: A Review of Literature. Transp Rev A Transnatl Transdiscipl J [Internet]. 2014;(May 2014):1–21. Available from: http://www.tandfonline.com/doi/abs/10.1080/01441647.2014.915442

31. Das S, Dutta A, Medina G, Minjares-Kyle L, Elgart Z. Extracting patterns from Twitter to promote biking. IATSS Res [Internet]. 2019;43(1):51–9. Available from: https://doi.org/10.1016/j.iatssr.2018.09.002

32. Chen Y, Mahmassani HS, Frei A. Incorporating social media in travel and activity choice models: conceptual framework and exploratory analysis. Int J Urban Sci [Internet]. 2017;0(0):1–21. Available from: https://www.tandfonline.com/doi/full/10.1080/12265934.2017.1331749

33. sklearn.feature_extraction.text.CountVectorizer [Internet]. Available from: https://scikit-learn.org/stable/modules/generated/sklearn.feature_extraction.text.CountVectorizer.html#

34. Text data preprocessing [Internet]. Available from: https://keras.io/api/preprocessing/text/

35. Ramos J. Using TF-IDF to Determine Word Relevance in Document Queries. Proceedings of the first instructional conference on machine learning. 2003.

36. Jurafsky D, Martin JH. Chapter 3: N-Gram Language Models N-Gram Language Models. Speech Lang Process. 2019;

37. Pauls A, Klein D. Faster and smaller n-gram language models. ACL-HLT 2011 - Proc 49th Annu Meet Assoc Comput Linguist Hum Lang Technol. 2011;1:258–67.

38. Blei DM, Edu BB, Ng AY, Edu AS, Jordan MI, Edu JB. 10.1162/jmlr.2003.3.4-5.993. CrossRef List Deleted DOIs [Internet]. 2000;1:993–1022. Available from: http://www.crossref.org/deleted_DOI.html

39. Steyvers M, Griffiths T. Probabilistic Topic Models. In: T. Landauer, D McNamara, S. Dennis WK, editor. Handbook of latent semantic analysis. Mahwah, NJ, US: Lawrence Erlbaum Associates Publishers; 2007. p. 427–48.

40. Farrahi K, Gatica-Perez D. Discovering routines from large-scale human locations using probabilistic topic models. ACM Trans Intell Syst Technol. 2011;2(1):1–27.

41. Hasan S, Ukkusuri S V. Urban activity pattern classification using topic models from online geo-location data. Transp Res Part C Emerg Technol. 2014;44.

42. Montoliu R. Discovering mobility patterns on bicycle-based public transportation system by using probabilistic topic models. Adv Intell Soft Comput. 2012;153 AISC:145–53.

43. Sun L, Yin Y. Discovering themes and trends in transportation research using topic modeling. Transp Res Part C Emerg Technol [Internet]. 2017;77:49–66. Available from: http://dx.doi.org/10.1016/j.trc.2017.01.013

44. Hong J, Tamakloe R, Lee G, Park D. Insight from Scientific Study in Logistics using Text Mining. Transp Res Rec. 2019;2673(4):97–107.

45. Rehurek R, Sojka P. Software Framework for Topic Modelling with Large Corpora. In: Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks. Valletta, Malta: ELRA; 2010. p. 45–50.

46. Gensim [Internet]. 2009 [cited 2017 Dec 29]. Available from: https://pypi.python.org/pypi/gensim

47. Mimno D, Wallach HM, Talley E, Leenders M. Mimno et al_2011_Optimizing semantic coherence in topic models. 2009;(1).

48. Topic Coherence Pipeline [Internet]. [cited 2020 Mar 20]. Available from: https://radimrehurek.com/gensim/models/coherencemodel.html

49. Röder M, Both A, Hinneburg A. Exploring the space of topic coherence measures. WSDM 2015 - Proc 8th ACM Int Conf Web Search Data Min. 2015;399–408.

50. Jiang L, Yu M, Zhou M, Liu X, Zhao T. Target-dependent Twitter sentiment classification. In: ACL-HLT 2011 - Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies. 2011.

51. Hutto CJ, Gilbert E. VADER: A parsimonious rule-based model for sentiment analysis of social media text. In: Proceedings of the 8th International Conference on Weblogs and Social Media, ICWSM 2014. 2014. p. 216–25.

52. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, et al. Scikit-learn: Machine Learning in Python. J Mach Learn Res. 2011;12:2825–30.

53. Baeza-Yates R, Ribeiro-Neto B, others. Modern information retrieval. Vol. 463. ACM press New York; 1999.

54. Manning CD, Raghavan P, Schütze H. Introduction to Information Retrieval. Vol. 35, Computational Linguistics. 2008. 482 p.

55. Chollet F, others. Keras [Internet]. GitHub; 2015. Available from: https://github.com/fchollet/keras

56. Guell C, Panter J, Jones NR, Ogilvie D. Towards a differentiated understanding of active travel behaviour: Using social theory to explore everyday commuting. Soc Sci Med [Internet]. 2012;75(1):233–9. Available from: http://dx.doi.org/10.1016/j.socscimed.2012.01.038

57. Macmillan A, Connor J, Witten K, Kearns R, Rees D, Woodward A. The societal costs and benefits of commuter bicycling: Simulating the effects of specific policies using system dynamics modeling. Environ Health Perspect. 2014;122(4):335–44.

58. Davis CA, Varol O, Ferrara E, Flammini A, Menczer F. BotOrNot: A System to Evaluate Social Bots. 2016;4–5. Available from: http://arxiv.org/abs/1602.00975%0Ahttp://dx.doi.org/10.1145/2872518.2889302

59. Jain S, Sharma V, Kaushal R. Towards automated real-time detection of misinformation on Twitter. 2016 Int Conf Adv Comput Commun Informatics, ICACCI 2016. 2016;2015–20.

60. Qazvinian V, Rosengren E, Radev DR, Mei Q. Qazvinian et al. - 2011 - Rumor has it Identifying Misinformation in Microblogs(2). Conf Empir Methods Nat Lang Process. 2011;1589–99.

61. Statista. Percentage of U.S. adults who use Twitter as of February 2019, by age group. https://www.statista.com/. 2019.

List of Figures

Tables

TABLE 1. Confusion Matrix

Predicted Label Actual Label	Negative (0)	Positive (1)
Negative (0)	True Negative (TN)	False Positive (FP)
Positive (1)	False Negative (FN)	True Positive (TP)

TABLE 2. Keyword lists for data collection

Tweet types	Keywords
Geolocated tweets	*'jog', 'run', 'biking', 'cycling', 'sidewalk', 'walk lane', 'pedestrian', 'walking', 'walk', 'bike', 'walkway', 'walk way', 'bikelane', 'bike lane', 'footpath', 'foot path', 'pedway', 'ped way', 'running', 'pavement', 'footway', 'disability', 'bicycle', 'jogging', 'bike share', 'bike sharing', 'shared bike lane'*

TABLE 3. Selected hyperparameters for the models

Model	**Model Parameters**
Naïve Bayes	Smoothing parameter alpha =1.0
Logistics Regression	Number of parameters is the same as the input feature size
SVN Classifier	Penalty Parameter, C=1000, kernel = linear
Embedding + LSTM	Embedding layer: Vocabulary size =7106, Output dimension = 512 LSTM layer: Number of Neurons: 256
Embedding +Bidirectional LSTM	Embedding layer: Vocabulary size =7106, Output dimension = 512 Bidirectional LSTM layer: Number of Neurons: 256
Embedding + Convolutional LSTM	Embedding layer: Vocabulary size =7106, Output dimension = 256 Convolution layer: filters=16, kernel size=3 Max pooling layer: pool size =8 LSTM Layer: Number of Neurons: 32

TABLE 4. Tweet classification performance for different models on the test dataset

Model	**Train Accuracy**	**Test Accuracy**	**Precision**	**Recall**	**F1 Score**	**AUC**

TF-IDF + Naïve Bayes	0.779	0.767	0.420	0.340	0.300	0.740
TF-IDF + Logistic Regression	1.000	0.821	0.750	0.710	0.730	0.890
TF-IDF + SVM Classifier	1.000	0.796	0.690	0.670	0.680	N/A
Embedding + LSTM	0.999	0.804	0.710	0.680	0.700	0.860
Embedding +Bidirectional LSTM	0.996	0.814	0.740	0.690	0.720	0.880
Embedding + Convolutional LSTM	0.952	0.807	0.740	0.690	0.710	0.860