Using Social Media to Predict Traffic Flow under Special Event Conditions

Ming Ni
Department of Industrial and Systems Engineering
University at Buffalo, The State University of New York
429 Bell Hall, Buffalo, NY 14260
Phone: (716)645-2357
Email: mingni@buffalo.edu

Qing He
Department of Civil, Structural and Environmental Engineering and
Department of Industrial and Systems Engineering
University at Buffalo, The State University of New York
313 Bell Hall, Buffalo, NY 14260
Phone: (716)645-3470
Email: qinghe@buffalo.edu

Jing Gao
Department of Computer Science and Engineering
University at Buffalo, The State University of New York
350 Davis Hall, Buffalo, NY 14260
Phone: (716)645-1586
Email: jing@buffalo.edu

Submitted to TRB 93rd Annual Meeting for Presentation and Publication
January 2014, Washington D.C.
November 15, 2013

Word counts (<7500): 7486
Abstract and Manuscript: 4486
Number of Tables and Figures: 12 (=250*12=3000)

* Corresponding Author
Abstract

Social media is a great resource of user-generated contents. Public attention, opinion and hot topics can be captured in the social media, which provides the ability to predict human-related events. Since social media can be retrieved in real time with relatively small building and maintenance costs, traffic operation authorities probably identify the social media data as another type of sensor for traffic demand. One of those challenges is how to extract reliable traffic-related features from big and noisy social media data. The other challenge is how to locate a feasible traffic study that fits well with social media data. In this paper, we aim to use social media information to assist traffic flow prediction under special event conditions. Specially, a short-term traffic flow prediction model, incorporated with tweet features, is developed to forecast the incoming traffic flow prior to sport game events. Both tweet rate features and semantic features are included in the prediction model. We examine and compare the performance of four regression methods, respectively autoregressive model, neural networks model, support vector regression, and k-nearest neighbor, with and without social media features. To the end, we show the benefit gained by including social media information in the prediction model and its computational efficiency for potential practical applications.
1. Introduction

Social media has become an indicator of modern people and lifestyle in the Internet virtual community. Vast of user-generated contents strengthen linkage and interaction between each individuals within the community circle, also provide large amount of information related to various area. Examples include Facebook, Twitter, YouTube, Google Plus and Wikipedia. The trend of easy accessing social media will continuously grow with the development and commercialization of wearable computer devices, like Google Glass and other smart watches.

On the other hand, Traffic congestion is one of the interesting and long-lasting problems in the world. In the last few decades, the sophisticated traffic network has been established around the world, but coming with both day-to-day recurrent traffic congestion and event-caused non-recurrent traffic congestion, which significantly affect the quality of life and impact the U.S. economy. Among different traffic mitigation counter-measures, traffic flow prediction alerts traffic congestion beforehand and facilitates pro-active traffic management and control. With an accurate prediction of traffic flow, road users are able to choose their routes accordingly, and traffic management agencies are capable to operate and control the traffic to alleviate possible congestion. Thus traffic flow prediction is one of the key methods to alleviate traffic congestion, especially under non-recurrent traffic conditions.

Motivated by the potentially great value of the social media, in this paper, we intend to incorporate social media data into traffic flow prediction models. Specially, the following question have been raised: Will social media help to predict prior-event traffic? Different with traffic data from traditional detectors, social media data is unorganized, noisy, big, and most of times unrelated to traffic. One of those challenges is how to extract reliable traffic related features from big and noisy social
media data. The other challenge is how to locate a feasible traffic study that fits well with social media data.

Due to data availability, the scope of traffic analysis in this paper is constrained in the freeway. The scope of the social media data is restrained to the tweets, which created, shared and modified by the Twitter users. As a basic introduction of tweets, each user of Twitter can be subscribed by other users known as followers. The message posts or status updates, called as tweets, is able to contain 140 characters or less, which typically is one kind of user-generated contents. The "@" sign followed by a username is used for mentioning or replying to other users. Hashtags of tweets are the words or phrases prefixed with a "#" sign, which is used to group posts together or indicated to a topic.

This paper will utilize those above components of Twitter massages, like content of tweet, number of hashtags, number of users, number of tweets with URLs, and number of tweets mentioned by other users. We focus on fusing social media data with event traffic data, in order to predict the dynamics of prior-event traffic. In this paper, we build a short-term traffic flow prediction model, incorporated with tweet features to forecast the incoming traffic flow prior to sport game events. We examine and compare the performance of four regression methods, respectively autoregressive model, neural networks model, support vector regression, and k-nearest neighbor, with and without social media features. It is observed that the prediction errors are notably decreased with the aid of social media data.

2. Literature review

There has not been considerable published research involved to use social media to do traffic prediction. Therefore, this section investigates literature from two perspectives seperately. One is related to the existing forecasting studies utilizing social media data. The other comes from the prior short-term traffic prediction models under atypical conditions.
Social Media

Although the social media includes a variety of web services, the previous studies generally fall into two parts: social relations and user-generated contents (1). Social relations reflect the virtually social network and connections between people on the web. The user-generated contents are created and published by the people, which intend to strengthen linkage and interaction between each individual within the community circle. In this paper, we focus on the previous studies with the user-generated contents.

Positive or negative emotions were able be extracted from Twitter data about certain company brands and products (2), which can be an auxiliary survey data applying to the company marketing strategy. Public opinion of election in US was studied by some researches (3)(4), effectively using the social data as a public survey to predict the election results with low cost comparing to the traditional telephone survey. As a different idea, a model was built by the rate of Twitter messages, related to movie, to predict movie box-office revenue and confirmed that the Twitter feeds can be effective indicators of real-world performance(5). An evaluation system about stock market was proposed to collect public mood and sentiment form Twitter service as an economic indicator (6). Other research also used social media to predict stock market indicator by analyzing Twitter posts. It is concluded that the correlation between stock market and social media indeed existed by examining emotional Twitter posts and market index such as Dow Jones, NASDAQ and S&P 500 (7).

Yu and Kak (1) identified three characteristics of event when social media data can be treated as a good predictor. The three characteristics are, respectively, human related event, masses of people involved, and the event should be easy to be talked in public.

Traffic Prediction
Since people intended to publish on social media corresponding to the non-recurrent traffic conditions, such as traffic special events, we focus on the literatures related to the traffic prediction under atypical conditions.

Neural Network was used to build the traffic volume prediction model which was based on the time-series data (8). In a similar manner, a supervised statistical learning technique called Online Support Vector machine for Regression, or OL-SVR, was applied for the prediction of short-term freeway traffic flow under both typical and atypical conditions (9). Guo et al. (10) pre-processed traffic data using Singular Spectrum Analysis, and utilized k- nearest neighbor method to predict traffic. Since pre-processing step reduced the noisy sensor inputs, this model can be used under normal and special conditions. An Online boosting nonparametric regression (OBNR) model also was used to perform traffic prediction, which consists of two major parts, respectively, the base part and the boosting part. The base part was constructed under normal conditions, while the gradient boosting part undertook special conditions into account (11).

3. Prediction Analysis

The framework of prediction analysis is presented as Figure 1. Starting with data preprocessing, both Twitter and traffic features are introduced and extracted by the following four components, traffic data de-trending, game traffic impacted detector identification, game tweet extraction and aggregation, and tweet semantics extraction.

![Flow diagram of prediction analysis](image)

Figure 1 The flow chart of the prediction analysis
Data Description

In this paper, traffic data were obtained from the Caltrans Performance Measurement System (PeMS). PeMS is a system designed to maintain California freeway traffic data and compute annual congestion for facilities with surveillance systems in place, typically loop detectors spaced approximately 0.5 mile apart on each freeway lane (12). The analysis uses 1-hour aggregated volume data, collected in four months from February 1, 2013 to May 30, 2013. Some of the detectors may miss or report invalid data in the practice. In order to compensate the missing or incorrect data samples, the diagnostics algorithm and imputation regression models developed in (13) were applied to detect the bad detectors and fill the missing value. This method generated total 12,242,699 entries of hourly traffic flow records in the database.

Twitter data was collected by the same spatial and temporal window accordingly, through the Twitter Streaming API with geo-location filter. Right after tweets collected, the filtration of spam and commercial tweets is implemented by the list of Twitter users. This results in a total number of 5,444,527 valid tweets.

The sport events are considered as good venues to perform both traffic and social media analysis, since it can be observed by the tweet posts, namely quantity and semantics. In addition, the traffic volume is dramatically influenced by the local sport events due to its popularity. In the study, we specially consider game traffic impact on I880 near Oracle Arena and O.co Coliseum in Oakland, California. Oracle Arena and O.co Coliseum was home to the Oakland Athletics (Athletics) and Golden State Warriors (Warriors) in 2013 game seasons. Oakland Athletics are a Major League Baseball team and Golden State Warriors are an American professional basketball team. There were 51 home games of the Athletics and Warrior from February to May in 2013. Oracle Arena and O.co Coliseum is located right besides Interstate Highway 880 (I-880). Six traffic detectors, located before the exits of I-880 to the entrance of
the Oracle Arena, were chosen to analyze incoming prior game traffic, shown as Figure 2.

Figure 2 Detectors layouts around Oracle Arena and O.co Coliseum

**Data Pre-processing**

In this part, we study the tweet and traffic data sets in order to obtain the potential features in the prediction modeling.

**Game and Non-game Day Comparison**

**Twitter**

The sport games, like Athletics’ and Warriors’, were considerably interested topics in the Twitter, especially among the Twitter users in the Bay area in order to comment and support their home teams. In order to identify the tweets which are relevant to
each game and each team, the appropriate keywords, shown as Table 1, were used to
select the game-related tweets with the help of disambiguation checks.

Table 1 Keywords to select game-related tweets

<table>
<thead>
<tr>
<th>Gold State Warriors</th>
<th>Oakland Athletics</th>
</tr>
</thead>
<tbody>
<tr>
<td>· letsgowarriors</td>
<td>· greencollar</td>
</tr>
<tr>
<td>· warrior</td>
<td>· letsgooakland</td>
</tr>
<tr>
<td>· dubnation</td>
<td>· athletics</td>
</tr>
<tr>
<td>· letsgodubs</td>
<td>· astalk</td>
</tr>
<tr>
<td>· warriorsground</td>
<td></td>
</tr>
<tr>
<td>· warriorgame</td>
<td></td>
</tr>
</tbody>
</table>

The keywords are chosen by the related phrases about the teams and the hashtags which are used to group the messages about the teams or related games. Those keywords are tested by a fraction of tweet data to verify whether the relevant tweet messages would be included or not. After iterative trials, those keywords in the Table 1 have been finalized.

The game-relevant tweets were aggregated into hourly time series in order to be compatible with Traffic data. For each hour in the 4 months, we extracted 5 features of tweet rates, shown as Table 2.

Table 2 Features for Tweet Rates

<table>
<thead>
<tr>
<th>Features for Tweet rates</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>no_of_tws</td>
<td>Number of tweets related to game in one hour</td>
</tr>
<tr>
<td>no_of_users</td>
<td>Number of independent users sent those tweets</td>
</tr>
<tr>
<td>no_of_hashtags</td>
<td>Number of hashtags in all those tweets</td>
</tr>
<tr>
<td>no_of_tws_mention</td>
<td>Number of tweets mentioned other twitter user</td>
</tr>
<tr>
<td>no_of_tws_urls</td>
<td>Number of tweets which has URLs</td>
</tr>
</tbody>
</table>
Figure 3 Comparisons of tweet rates on game days and non-game days

Note the non-game days stand for the days which don’t carry on either home game or away game, while game days denotes only for the days carrying on home game at Oracle Arena or O.co Coliseum.

Figure 3 shows the number of tweets per hour for Golden State Warriors at their game days and non-game days. One can see that there are much more tweets on the game day than that on the non-game day. Therefore, we can use the tweet rate series on the same day to indicate the total public attention for this game.
First of all, in order to eliminate the effects of day-to-day traffic fluctuations, we detrended the traffic flow by subtracting the average hour-of-day and day-of-week traffic volumes. For every record of traffic flow, the below equation was used to reduce the periodic variance so that we can only concentrate on the increment of the traffic flow.

\[ \Delta y_{wt} = y_{wt} - \bar{y}_{wt} \]

Where \( y \) stands for the traffic flow; \( w \) indicates it is weekday or weekend; \( t \) denotes the hour of day; \( \bar{y}_{wt} \) is the average time-of-day and day-of-week traffic flow.

The time-of-day patterns of traffic flow are demonstrated in Figure 4 for both weekday and weekend. Figure 4 shows the different periodic fluctuations of traffic between weekday and weekend. In traffic prediction, it is common practice to exclude such fluctuations. Therefore, the estimated variations of traffic flow, respectively weekday and weekday, are used here to de-trend the hourly traffic flow data.

![Figure 4 Hourly traffic flow for detectors](image-url)
Figure 5 The de-trended prior-game traffic flow

The de-trended prior game traffic is plotted in Figure 5, which takes the last 2 hours before the game event starting into account. The majority of de-trended game traffic series is greater than zero, which indicates more traffic demand on the game day.

In order to test the impact of game on traffic data, a single-factor analysis of variance (ANOVA) was used to differentiate between two populations of data sets through factor “game day”. One dataset contains the de-trended traffic flows within 4 hours before the game starting time. The other includes the binary game indicator during the same time on every non-game day.

Table 3 ANOVA test results for factor “game day”

<table>
<thead>
<tr>
<th>Detector</th>
<th>P-value</th>
<th>Significant or not</th>
</tr>
</thead>
<tbody>
<tr>
<td>400498</td>
<td>0.0064</td>
<td>YES</td>
</tr>
<tr>
<td>400955</td>
<td>1.24e-13</td>
<td>YES</td>
</tr>
<tr>
<td>400190</td>
<td>6.3e-07</td>
<td>YES</td>
</tr>
<tr>
<td>400956</td>
<td>0.029</td>
<td>NO</td>
</tr>
<tr>
<td>400360</td>
<td>4.52e-13</td>
<td>YES</td>
</tr>
<tr>
<td>400333</td>
<td>7.45e-13</td>
<td>YES</td>
</tr>
</tbody>
</table>
Table 3 presents the results from ANOVA analysis. As one can see, sport games generate significantly impact on the traffic volume of most surrounding detectors on game days. Note that the traffic flow at detector 400956 does not demonstrate significant differences caused by game events, since the detector is located at downstream of the exit towards Oracle Arena. In addition, the distance between detector 400333 and Oracle Arena is much larger comparing with other detectors. For the purpose of simplification, these 2 traffic detectors are treated as only the neighboring detectors rather than the analytical targets to be predicted in the following analysis.

**Tweet Semantics Extraction**

One of important differences between tweets and other types of data is that tweets consist of rich sentiments. For our analysis, we believe the public sentiment about one particular game has influence on the attendance, just like the relationship between the public opinion about a company and the stock price. Psychologically, more biased emotion about one topic, more attention would be gained in the public. Therefore, sentiment extraction is necessarily introduced to the analysis, which served as the potential semantic features of tweet data to be investigated in the following study.

Sentiment analysis is an important part in machine learning. In here, sentiment results are treated as potential tweet features. In this paper, the sentiment analysis is performed based on an R (14) package, called “tm.plugin.sentiment”, which provides the functions for natural language sentiment processing. The words of text in each Tweet are labeled as positive, negative and neutral references. Before conducting the sentiment analysis, the text part of tweets was initially preprocessed by the following steps.

1. Make each letter lowercase,
2. Remove punctuation and stopwords.
3. Replace the abbreviations from the formal English words.
Our study implemented a unified letter transformation approach (15) to normalize the Tweet post, to replace the non-standard tokens to the Standard English words. This method tends to alleviate the biased effect of abnormal text on the sentiment analysis.

In order to quantify the sentiments of tweets, a sentiment scores system was implemented and based on the Lydia/TextMap system (16)-(17). There are 5 measures listed as follows.

- **polarity** \( \frac{p - n}{p + n} \): difference of positive and negative sentiment references / total number of sentiment references
- **subjectivity** \( \frac{p + n}{N} \): total number of sentiment references / total number of references
- **pos_refs_per_ref** \( \frac{p}{N} \): total number of positive sentiment references / total number of references
- **neg_refs_per_ref** \( \frac{n}{N} \): total number of negative sentiment references / total number of references
- **senti_diffs_per_ref** \( \frac{p - n}{N} \): difference of positive and negative sentiment references / total number of references

Above 5 measures were calculated for every tweet post.

**Prediction Analysis and Modeling**

After the traffic and tweet data have been processed, there are many usable features of those two, respectively the traffic flowing of neighboring detectors, and 5 tweet rate features plus 5 tweet sentiment score features. This section, two kinds of features are combined together to build the short-term traffic prediction model under the special event conditions.

**Prediction Features Selection**

Because there are many potential features of traffic and tweet at hand, it is essential to select the right significant features to build a parsimonious prediction model.
It varies that the correlations between tweets features and traffic flow for each detectors, as Figure 6 shows. Therefore, the features selection is an essential step and very significantly affects the performance of the prediction modals.

Figure 6 Correlation analysis between tweet features and game traffic incremental

(a) Correlation plot of twitter polarity and 400190 traffic flow
(b) Correlation plot of No. of Twitter user and 400955 traffic flow

Where the “W” stands for games of Golden State Warriors. “A” stands for games of Oakland Athletics. “AW” denotes the game time happened to have two teams together at the Oracle Arena. “ALL” is for correlation for the all of games together.

Those traffic and tweet features were combined into an optimal matrix for each traffic detector. Here, we used the least squares optimization with L1-norm regularization (18) to solve this problem.

\[
\min_{\beta} \sum_{k=1}^{K} \sum_{t=1}^{T} (\Delta y_k^{T_{g_{t-}}^g - \alpha} - \sum_{m=1}^{M} \sum_{l=1}^{L} \beta_{ml} \Delta y_{km}^{T_{g_{t-l}}^g - T_{g_{t-l}}^g} - \sum_{p=1}^{P} \chi_{kp} \Delta y_{kp}^{T_{g_{t-r}}^g - T_{g_{t-r}}^g} V_p)^2 + \lambda \sum_{p=1}^{P} |V_p|
\]

Where $K$ is the number of game; $T$ stands for the time steps before game start to be predicted and $T_{g}$ indicates the game starting time; $M$ is total number of neighboring detectors; $L$ denotes the number of time lags of the neighboring detectors; $P$ is the number of tweet features; $r$ represents the number of time lags of tweets relative to
traffic data; $\beta_i$ is the coefficients for traffic variables; $V_p$ is the coefficients for tweet variables; $\alpha$ indicates the intercept of the model. $\lambda > 0$ is the regularization parameter that balances the L1 regularization term and least square term.

By applying different values of $\lambda$, we solved the above optimization problem by CPLEX. Appropriate traffic and tweet features were selected with non-zero coefficients. For the traffic features, the traffic flow variables of neighboring detectors were always chosen by the optimization model. For the tweet features, every target detector had different combination of tweet rate features and tweet sentiment features, shown as Table 4.

### Table 4 Selected tweet features for each detector when $\lambda = 4000$

<table>
<thead>
<tr>
<th>Traffic Flow</th>
<th>Selected Tweet Features</th>
<th>Traffic Flow</th>
<th>Selected Tweet Features</th>
</tr>
</thead>
<tbody>
<tr>
<td>tf_400498_1</td>
<td>no_of_tws_a_sum4</td>
<td>tf_400190_1</td>
<td>no_of_hashtags_a_sum4</td>
</tr>
<tr>
<td></td>
<td>no_of_tws_mentions_a_sum4</td>
<td></td>
<td>no_of_hashtags_w_sum4</td>
</tr>
<tr>
<td></td>
<td>no_of_tws_mentions_w_sum4</td>
<td></td>
<td>no_of_tws_a_sum4</td>
</tr>
<tr>
<td></td>
<td>no_of_tws_urls_w_sum4</td>
<td></td>
<td>no_of_tws_mentions_a_sum4</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>no_of_tws_mentions_w_sum4</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>no_of_tws_urls_a_sum4</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>no_of_tws_urls_w_sum4</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>no_of_tws_w_sum4</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>no_of_users_a_sum4</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>no_of_users_w_sum4</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>polarity_sum_w_sum4</td>
</tr>
<tr>
<td>tf_400955_1</td>
<td>no_of_hashtags_a_sum4</td>
<td>tf_400360_1</td>
<td>no_of_hashtags_a_sum4</td>
</tr>
<tr>
<td></td>
<td>no_of_hashtags_w_sum4</td>
<td></td>
<td>no_of_hashtags_w_sum4</td>
</tr>
<tr>
<td></td>
<td>no_of_tws_mentions_a_sum4</td>
<td></td>
<td>no_of_tws_mentions_w_sum4</td>
</tr>
<tr>
<td></td>
<td>no_of_tws_mentions_w_sum4</td>
<td></td>
<td>no_of_tws_urls_a_sum4</td>
</tr>
<tr>
<td></td>
<td>no_of_tws_urls_w_sum4</td>
<td></td>
<td>no_of_tws_urls_w_sum4</td>
</tr>
<tr>
<td></td>
<td>no_of_tws_w_sum4</td>
<td></td>
<td>no_of_users_a_sum4</td>
</tr>
<tr>
<td></td>
<td>no_of_users_a_sum4</td>
<td></td>
<td>no_of_users_w_sum4</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>polarity_sum_w_sum4</td>
</tr>
</tbody>
</table>

In the Table 4, the independent variables are the volumes of traffic flow (tf) in last 1 hour prior to the games. Because there are 2 sport teams, in here, “w” stands for Golden State Warriors and “a” stands for Oakland Athletics. “sum4” means that the tweet feature is aggregated in last 4 hours before the predicting time.

### Prediction Models
There were many existing prediction approaches for short-term traffic flow, although most of them focus on traffic features only. To compare with the performance under different methods, we examine the following four popular methods to predict short-term event traffic (19) (20) (21):

- Autoregressive Model (AR)
- Neural Networks (NN)
- Support Vector Regression (SVR)
- K-Nearest Neighbor (KNN)

In order to test the performance of social media features in short-term traffic prediction, two models is constructed for every target detector under each method. The model 1 is typically based on the traffic features only, while the model 2 depends on both traffic features and tweet features. By comparing the results between Model 1 and Model 2, we are able to tell whether the social media features improve the overall performance of the accuracy of traffic prediction.

- Model 1: Target traffic flow\sim (Neighboring Traffic flow)
- Model 2: Target traffic flow\sim (Neighboring Traffic flow) + (Tweet Features)

The implementation of those methods is as following. AR model basically implements the linear regression model. The architecture of NN model consists of the input layer, a single hidden layer with 6 neurons and 1 output neuron. In terms of SVR, we used typical linear kernel and epsilon-regression. KNN method takes the straight average of the 10 nearest points in the training set in Euclidean distance to obtain the prediction results.

4. Performance Results

The prediction performance is evaluated by two measures, namely Mean Absolute Percentage Error (MAPE) and Root Mean Square Error (RMSE), defined as follows
Where $K$ stands for the total number of games; $\hat{y}$ denotes the estimated value of traffic flow $y$; $T$ represents total number of time steps to be predicted before the game starts, and $T_g$ indicates the game starting time;

Every prediction model is implemented with above four methods. The training data occupies the 70% of entire dataset, while the remaining 30% was treated as test data. We generated 100 instances of training and test data with 100 different random seeds. Every model runs for 100 times. And the performance is evaluated by the average MAPE and RMSE of 100 experiments.

Figure 7 shows the evaluation results of two models in four regression methods for four target detectors. As one can see, although the MAPEs doesn’t show consistent results across four regression methods and four detectors, the RMSEs of model 2 with tweet features always outperform those of model 1 with only traffic features. Therefore, it is very beneficial to incorporate tweet features in the prediction models. In other words, the tweet features indeed improve the results of short-term traffic prediction. More importantly, such performance improvement exists regardless of forecasting methods.
Figure 7 The MAPE and RMSE for each traffic detector
As indicated from Figure 7, SVR shows overall best results among four methods for game traffic prediction. The average of MAPE and RMSE of SVR are presented in Table 5.

Table 5 Average of MAPE and RMSE for SVR

<table>
<thead>
<tr>
<th></th>
<th>Model 1 Traffic only</th>
<th>Model 2 Traffic Tweet</th>
<th>Error reduction (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>RMSE</td>
<td>385.5252</td>
<td>311.0429</td>
<td>23.946%</td>
</tr>
<tr>
<td>MAPE</td>
<td>0.050437</td>
<td>0.047146</td>
<td>6.9804%</td>
</tr>
</tbody>
</table>

The percentage of error reduction in Table 5 indicates the promising value of social media in event traffic prediction. Because it is very difficult to predict the short-term traffic flow under abnormal conditions, any improvement is significant for traffic management and operations, especially when using no-cost and real-time social media data.

The computational efficiency of short term traffic prediction is critical for practical applications, such as Advanced Traveler Information Systems (ATIS). The traffic flow data usually reports every 30 seconds from detectors, which can be used directly in real-time. The social media data comes from the Twitter Stream application interface in real time but it requires additional data processing to get the tweet rate features and tweet sentiment features. Considering the hourly aggregation and keywords filtration to obtain sport game related tweet, it is feasible to prepare the right social media data in order to get the desirable tweet features in an hour. Estimation of processing time for an hour tweet feature extraction is twenty minutes. As a result, the proposed prediction models can be directly applied for real-time practical applications.

5. Conclusion

In this paper, social media data is incorporated with traffic flow data for with prior-event traffic prediction. The linkages of tweet data and the sport game traffic is analyzed and recognized to build a short-term traffic flow prediction model. We
identify the traffic flow of neighboring detectors as traffic features; the tweet rates and
tweet sentiments as social media features. Four popular forecasting methods,
respectively, autoregressive model, neural networks, support vector regression, and k-
nearest neighbor, are implemented to build the short-term traffic flow prediction
models. We then find that the prediction results with both traffic and tweet features
outperformed those with only traditional traffic features across different detectors.
With support vector regression, average MAPE and RMSE was reduced by 6.98%
and 23.95% after including tweet features in the model. We also argue the efficiency
of computation time for social media data processing in practical real-time
applications.

At the same time, another purpose of research is to demonstrate one way to uncover
the potential great value of big data generated by the social media. In this case, it can
be considered as an effective indicator of public attention and opinions. This
possibility may transfer to other branches of transportation, such as travel behavior
analysis, transportation related choices, transit scheduling and operations, and so on.

Reference


   word of mouth. Journal of the American Society for Information Science and

3. Tumasjan A, Sprenger T, Sandner P, Welpe I. Predicting elections with twitter:

   Text Sentiment to Public Opinion Time Series. 2010.

5. Asur S, Huberman BA. Predicting the Future with Social Media. 2010
   IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent


Godbole N, Srinivasaiah M, Skiena S. Large-Scale Sentiment Analysis for News and Blogs. Proceedings of the International Conference on Weblogs and Social Media (ICWSM); 2007.


