Improving Traffic Prediction with Tweet Semantics

Jingrui He*, Wei Shen†, Phani Divakaruni‡, Laura Wynter‡, Rick Lawrence‡
* Computer Science Department, Stevens Institute of Technology, jingrui.he@gmail.com
† @Walmartlabs, Walmart eCommerce, wshen@walmartlabs.com
‡ BAMS, IBM Research, {ricklawr, phanid, lwynter}@us.ibm.com

Abstract

Road traffic prediction is a critical component in modern smart transportation systems. It provides the basis for traffic management agencies to generate proactive traffic operation strategies for alleviating congestion. Existing work on near-term traffic prediction (forecasting horizons in the range of 5 minutes to 1 hour) relies on the past and current traffic conditions. However, once the forecasting horizon is beyond 1 hour, i.e., in longer-term traffic prediction, these techniques do not work well since additional factors other than the past and current traffic conditions start to play important roles.

To address this problem, in this paper, for the first time, we examine whether it is possible to use the rich information in online social media to improve longer-term traffic prediction. To this end, we first analyze the correlation between traffic volume and tweet counts with various granularities. Then we propose an optimization framework to extract traffic indicators based on tweet semantics using a transformation matrix, and incorporate them into traffic prediction via linear regression. Experimental results using traffic and Twitter data originated from the San Francisco Bay area of California demonstrate the effectiveness of our proposed framework.

1 Introduction

With the steadily increasing number of motor vehicles in the United States, road traffic prediction becomes a critical component in modern smart transportation systems. Accurate prediction of both near-term and longer-term traffic conditions can greatly help traffic management agencies generate proactive strategies to alleviate congestion. It can also help road users better plan their trips by avoiding road segments expected to be congested soon. Existing work on road traffic prediction largely focuses on forecasting horizons in the range of 5 minutes to 1 hour by using past and current traffic conditions [Al-Deek *et al.*, 2001; Smith *et al.*, 2002; Kamarianakis and Prastacos, 2003; Min and Wynter, 2011]. The proposed techniques do not generalize well to forecasting horizons beyond 1 hour due to the impact of additional factors, such as scheduled events [Maze *et al.*, 2006; Mahmassani *et al.*, 2009].

With the rapid growth of online social media, more and more people are using Twitter, Facebook, etc to communicate their mood, activities, plans, as well as to exchange news and ideas, which creates a huge repository containing information not accessible from conventional media. In particular, a lot of people are using their mobile devices to access the social media web sites via web applications, hence generating a large number of messages on the go. Many of the messages are related to the current traffic conditions, such as 'Traffic jam on new preedy street, near Parking Plaza Saddar, cars unmoved for last 20 mins', 'Big road block intersection of Rondebult and Commissioner street Boksburg', etc. It is also common for people to announce their travel plans in the near future, such as 'This SUNDAY !!!! We will be playing at Di Piazzas in Long Beach', 'good night! getting up early tomorrow to pack and then off to the airport for our flight @ 5PM', etc.

Motivated by the uniqueness of the information contained in online social media, and the close relationship between traffic and tweets, In this paper, we answer the following question: can we extract tweet-based semantics to help improve longer-term traffic prediction? To answer this question, we first establish the correlation between traffic measurements and tweet counts at various granularity. Then we directly extract semantics from tweets via a sparse matrix, and incorporate the semantics into the auto-regression model used in traffic prediction. Finally, the sparse matrix is obtained by solving an optimization framework, whose goal is to minimize the prediction error in the traffic measurements.

The rest of the paper is organized as follows. In Section 2, we briefly review existing work on traffic prediction and social media aided analysis. Then we study the correlation between traffic measurements and tweet counts in Section 3. It leads to the optimization framework for systematically incorporating tweet semantics in traffic prediction and an iterative algorithm for solving it in Section 4. The experimental results are presented in Section 5. Finally, we conclude the paper in Section 6.

2 Related Work

In this section, we review the existing work from two perspectives, namely traffic prediction and social media aided analysis.

2.1 Traffic Prediction

Road traffic prediction is a critical component in modern smart transportation systems. With an accurate prediction of traffic conditions, traffic management agencies can generate *proactive* traffic operation strategies to alleviate congestion, and road users can plan their trips accordingly ahead of time.

The modeling approaches of traffic prediction can be classified into parametric methods and non-parametric methods. The former category relies primarily on statistical techniques, including historical average and smoothing techniques [Smith and Demetsky, 1997; Williams *et al.*, 1998], auto-regressive moving average models [Ahmed and Cook, 1979; Levin and Tsao, 1980; Al-Deek *et al.*, 2001; Smith *et al.*, 2002; Kamarianakis and Prastacos, 2003; Min and Wynter, 2011], and Kalman filter algorithms [Okutani and Stephanedes, 1984; Guo and Williams, 2010]. The main non-parametric approaches published to date include non-parametric regression [Smith and Demetsky, 1996; Clark, 2003; Huang and Sadek, 2009] and artificial neural networks (ANN) [Clark *et al.*, 1993; Vythoulkas, 1993; Yun *et al.*, 1998; van Lint *et al.*, 2005; Vlahogianni *et al.*, 2005; Khosravi *et al.*, 2011]. These studies rely primarily on traffic data collected from sensors such as loop detectors, GPS devices, cell phones, etc., with forecasting horizons in the range of 5 minutes to 1 hour. Studies on longer-term traffic prediction are rather limited, primarily because additional factors other than the past and current traffic conditions start to play important roles once the forecasting horizon is beyond 1 hour. Only a few researchers and private companies have attempted to analyze and utilize the correlation between traffic data and those external factors such as weather and event schedules [Maze *et al.*, 2006; Mahmassani *et al.*, 2009].

The main focus in this paper is on investigating how Twitter data can be used as an external data source for improving near-term traffic prediction beyond the forecasting horizon of 1 hour. To the best of our knowledge, our study is the first to leverage the rich information in social media to help with traffic prediction.

2.2 Social Media Aided Analysis

As mentioned in the previous section, nowadays, many researchers are trying to exploit the rich information in social media for various purposes. For example, there is a lot of interest in using social media to detect emerging news or events: in [Petrovic *et al.*, 2010], the authors address the problem of detecting new events from a stream of Twitter posts using an algorithm based on locality-sensitive hashing; in [Sankaranarayanan *et al.*, 2009], the authors propose a news processing system called *TwitterStand* to capture tweets that correspond to late breaking news; in [Sakaki *et al.*, 2010], the authors investigate the real-time interaction of events such as earthquakes in Twitter, and propose a probabilistic spatiotemporal model for the target event that can find the center and the trajectory of the event location, etc.

Another line of research is tweet classification for the purpose of information filtering. For example, in [Go *et al.*, 2011], the authors test various algorithms for classifying the sentiment of tweets, such as SVM, Naive Bayes, etc; in [Sriram *et al.*, 2010], the authors use a small set of domain-specific features in addition to the bag-of-word features to classify tweets into a predefined set of classes; etc.

Furthermore, some researchers are extracting information from tweets which might be useful in another domain. In [Bollen *et al.*, 2010], the authors try to answer the question: is the public mood correlated or even predictive of economic indicators? To this end, they first derive from large scale Twitter feeds the collective mood states, and then perform the correlation analysis with the Dow Jones Industrial Average (DJIA) over time. Finally, they show that the accuracy of DJIA predictions can be significantly improved by the inclusion of specific public mood dimensions, such as *Calm*. In [Eisenstein *et al.*, 2010], based on the geo-tagged social media, the authors propose a multi-level generative model that reasons jointly about latent topics and geographical regions. Our proposed work belongs to this direction, and we try to build the correlation between Twitter and a new domain, namely traffic prediction. This is motivated by the presence of a large number of tweets related to traffic conditions.

3 Correlation Study

In this section, we study the correlation between traffic measurements and tweet counts by first introducing the data sets used in this paper, and then presenting the correlation analysis.

3.1 Data Description

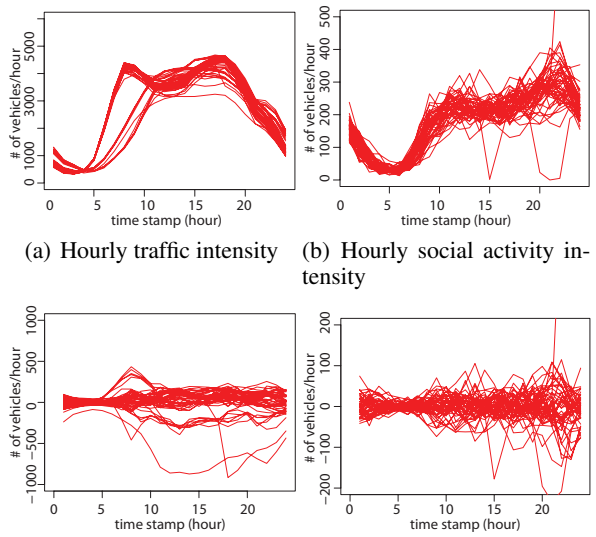
To accommodate the correlation analysis, we need two data sets: one containing traffic measurements, and the other containing tweet information. We generate the traffic data set by collecting measurements from 943 loop detectors covering the San Francisco Bay area between August 3, 2011 and September 30, 2011 using the California Performance Measurement System (PeMS, [1]). This data set contains 1,380,552 entries, each of which records an hourly traffic volume measured at one detector location. We also collected tweet data for the same area during the same time period. The tweets were obtained using the Twitter streaming API with a geo-location filter defining the lat/long bounding box of (-122.75, 36.934, -121.75, 38.369). To avoid spam, we filter out tweets that contain the regular expressions of "http:" or "www.". For each tweet, we collect information of user account, time stamp, content, and the geo-location. This results in a total number of 212,145 tweets from 19,435 distinct users. Note that due to an unexpected data center outage, Twitter data during Sept 2 to Sept 12 were not collected.

3.2 Data Processing

For traffic data, let $\mathbf{v} \in \mathbb{R}^T$ denote the time series of regional level traffic intensity, where T is the total number of time stamps, and its t^{th} element $\mathbf{v}^t$ is the traffic volume (the total number of vehicles passing each detector) averaged over all 943 detectors in time stamp t.

Due to the recurrent nature of traffic, $\mathbf{v}$ typically exhibits periodic fluctuations, as can be seen in Figure 1(a). In traffic prediction, it is common practice to exclude such fluctuations in the prediction models. Therefore, we first estimate

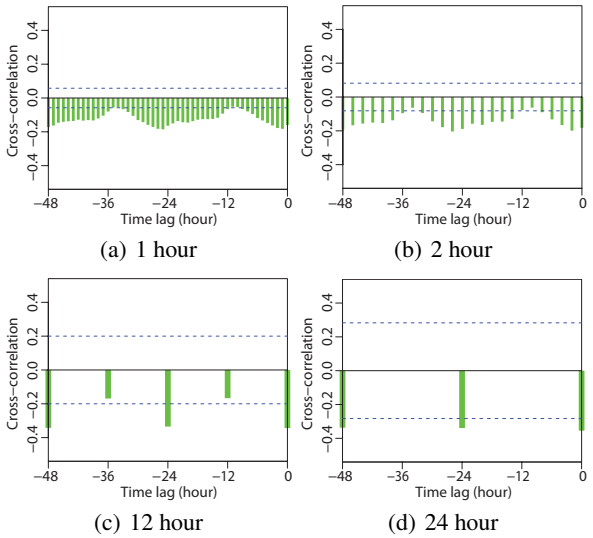
[1] http://pems.dot.ca.gov



(a) Hourly traffic intensity (b) Hourly social activity intensity

(c) De-trended hourly traffic intensity (d) De-trended hourly social activity intensity

Figure 1: Data de-trending



(a) 1 hour (b) 2 hour

(c) 12 hour (d) 24 hour

Figure 2: Cross correlation between de-trended traffic intensity and de-trended social activity intensity

the seasonal variation component and then subtract it from $\mathbf{v}$ to get the de-trended version. More specifically, for each (hour of day, day of week) pair (τ, d) where $\tau = 0, \ldots, 23$ and $d = 0, \ldots, 6$, we define the seasonal variation component $\epsilon_{\tau,d}$ as follows

$$\epsilon_{\tau,d} = \frac{\sum_{\{t|g(t)=(\tau,d)\}} \mathbf{v}^t}{|\{t|g(t) = (\tau, d)\}|}$$

where $g(\cdot)$ is an operator retrieving both the hour of day and day of week indices for a given time stamp t, and $|\{\cdot\}|$ denotes the number of elements in the set.

The de-trended version of regional level traffic intensity is now defined as $\delta\mathbf{v} \in \mathbb{R}^T$, where its t^{th} element $\delta\mathbf{v}^t$ is set as follows.

$$\delta\mathbf{v}^t = \mathbf{v}^t - \epsilon(\tau, d), \text{ s.t. } g(t) = (\tau, d)$$

For Twitter data, let $\mathbf{c} \in \mathbb{R}^T$ denote the time series of social activity intensity measure, whose t^{th} component $\mathbf{c}^t$ is the total number of tweets for time stamp t. Similar to the traffic intensity measure $\mathbf{v}$, $\mathbf{c}$ also exhibits periodic fluctuations. Therefore, we define the de-trended version $\delta\mathbf{c}$ in a similar way as $\delta\mathbf{v}$.

Figure 1 compares the time series of traffic and social media intensities before and after de-trending. Each line in the plot corresponds to the data from one day in the studied time period, and each time stamp corresponds to an hour. The daily recurrent patterns in both the raw traffic and Twitter data can be clearly observed.

3.3 Correlation Analysis

As a first step towards predicting traffic intensity using Twitter data, we test if social activity intensity measure has any correlation with the traffic intensity measure in the same region. Figure 2 shows the cross-correlation results between the current de-trended traffic intensity $\delta\mathbf{v}$ and the de-trended social activity intensity $\delta\mathbf{c}$ in the past 48 hours with time resolutions of 1, 2, 12, and 24 hours respectively. The height of the green bar at time lag $-\Delta t$ represents the correlation between $\delta\mathbf{v}^t$ and $\delta\mathbf{c}^{t-\Delta t}$. The two blue dashed lines mark the 95% confidence intervals of the correlation values.

As can be seen from this figure, the current de-trended traffic intensity and the de-trended social activity intensity in the past 48 hours exhibit statistically significant correlation. Quite interestingly, the correlation seems to be negative for all four time resolutions tested, which implies that when the social activity is less intense than the average level, the traffic activity on the road network is usually more intense than the average level in the near future. In terms of correlation levels, larger time resolutions such as 12 hours and 24 hours tend to have higher absolute values of correlation than smaller ones such as 1 hour and 2 hours. For the data with 12-hour resolution, the correlation for only even time lags is statistically significant.

Furthermore, we test the significance of the correlation between the two time series by adding lagged $\delta\mathbf{c}$ to the original auto-regression model used for traffic prediction [Smith and Demetsky, 1997]. To be specific, we predict $\delta\mathbf{v}$ using the following linear regression model.

$$\delta\mathbf{v}^t = \alpha + \beta_1\delta\mathbf{v}^{t-1} + \beta_2\delta\mathbf{v}^{t-2} + \gamma_1\delta\mathbf{c}^{t-1} + \gamma_2\delta\mathbf{c}^{t-2} \quad (1)$$

where α is the offset, β_1, β_2, γ_1, and γ_2 are coefficients associated with traffic and Twitter data with various lags. We apply this model with time resolutions of 1, 2, 12, and 24 hours respectively, and identify the covariates that are statistically significant. The results are summarized in Table 1.

In this table, the second column shows the estimated value of the coefficients; the third column shows the standard error; the fourth column is the t statistics; and the last column is the p-value. For the time resolutions of 1, 2, 12 hours, there exists at least one lag of $\delta\mathbf{c}$ that is statistically significant.

Furthermore, the coefficients of such covariates are negative, which is consistent with the cross-correlation results shown in Figure 2. For the time resolution of 24 hours, the regression model does not include any statistically significant lags of $\delta\mathbf{c}$.

Table 1: Results of Multiple Linear Regressions

(a) 1-hour time resolution

Coefficients	Value	Std. Err.	t	p-value
α	0.341	1.550	0.220	0.826
β_1	1.142	0.029	39.674	0.000 $*$
β_2	-0.235	0.029	-8.227	0.000 $*$
γ_1	-0.161	0.050	-3.233	0.001 $*$
γ_2	0.046	0.057	0.804	0.421

(b) 2-hour time resolution

Coefficients	Value	Std. Err.	t	p-value
α	0.529	2.873	0.184	0.854
β_1	1.073	0.041	26.246	0.000 $*$
β_2	-0.246	0.040	-6.097	0.000 $*$
γ_1	-0.163	0.051	-3.187	0.002 $*$
γ_2	0.069	0.055	1.244	0.214

(c) 12-hour time resolution

Coefficients	Value	Std. Err.	t	p-value
α	-2.398	8.557	-0.280	0.780
β_1	0.258	0.086	2.985	0.004 $*$
β_2	0.538	0.086	6.231	0.000 $*$
γ_1	0.036	0.042	0.854	0.396
γ_2	-0.109	0.041	-2.633	0.010 $*$

(d) 24-hour time resolution

Coefficients	Value	Std. Err.	t	p-value
α	-1.919	12.470	-0.154	0.878
β_1	0.698	0.160	4.363	0.000 $*$
β_2	-0.035	0.164	-0.215	0.831
γ_1	-0.048	0.043	-1.120	0.269
γ_2	-0.019	0.039	-0.502	0.619

(Note: $*$ means p-value < 0.05)

4 Optimization Framework for Incorporating Tweet Semantics

In the previous section, we established the correlation between de-trended traffic intensity and de-tended social activity intensity. In this section, we propose a general optimization framework, which extends our analysis beyond the social activity intensity, and extracts traffic indicators based on tweet semantics to better predict traffic conditions.

4.1 Traffic Indicators based on Tweet Semantics

During time stamp t, we first map each tweet to the space of stemmed words (stop words removed), which generates a non-negative sparse vector. Putting all such vectors together, and appending an additional column of all 1s, we have the sparse feature matrix $\mathbf{F}^t \in \mathbb{R}_+^{n^t \times (d+1)}$, where n^t is the total number of tweets in time stamp t, and d is the number of stemmed words. Its element $\mathbf{F}^t_{i,j}$ ($i = 1, \ldots, n^t$, $j = 1, \ldots, d$) in the i^{th} row and j^{th} column is positive if and only if the j^{th} word appears in the i^{th} tweet.

Furthermore, let $\mathbf{M} \in \mathbb{R}^{(d+1)\times m}$ denote the transformation matrix, where m is the number of traffic indicators based on tweet semantics. Its element $\mathbf{M}_{j,k}$ ($j = 1, \ldots, d$, $k = 1, \ldots, m$) in the j^{th} row and k^{th} column corresponds to the weight of the j^{th} word in the k^{th} traffic indicator, and its elements in the last row correspond to the offsets of each traffic indicator. For example, suppose that the first traffic indicator only has a large weight for word *today*, then it mainly collects information from tweets related to the activities happening today; suppose that the second traffic indicator only has a large weight for word *airport*, then it focuses on the conditions around the airport, etc. Various traffic indicators are used to depict different aspects of traffic, e.g., according to time, location, etc. Therefore, it is easy to understand that $\mathbf{M}$ is sparse column-wise, which corresponds to each traffic indicator. However, it may not be sparse row-wise, since some words may have positive weights in many traffic indicators, e.g., *traffic*.

Finally, the matrix $\mathbf{S}^t \in \mathbb{R}^{n\times m}$ that consists of the semantic-based traffic indicators for all the tweets in time stamp t is obtained by $\mathbf{S}^t = \mathbf{F}^t \times \mathbf{M}$. Its element $\mathbf{S}^t_{i,k}$ in the i^{th} row and k^{th} column measures the strength of the k^{th} traffic indicator in the i^{th} tweet. Using the previous example, if a tweet mentions the activities around the beach today, then the strength of the first traffic indicator according to the semantics of this tweet is large, whereas the strength of the second traffic indicator (which is related to the conditions around the airport) is small.

4.2 Optimization Problem

Next, we introduce the optimization problem, which finds the optimal transformation matrix $\mathbf{M}$ that minimizes the traffic prediction error. To be specific, we solve the following objective function with respect to $\mathbf{M}$.

$$\min_{\mathbf{M},\alpha,\beta_l,\gamma_l} \sum_{t=\max(r_1,r_2)+1}^{T} (\delta\mathbf{v}^t - \alpha - \sum_{l=1}^{r_1} \beta_l \delta\mathbf{v}^{t-l} - \sum_{l=1}^{r_2} \gamma_l \sum_{k=1}^{m} \mathbf{1}_{1\times n^{t-l}} \mathbf{S}^{t-l}_{:,k})^2 + \lambda \sum_{k=1}^{m} |\mathbf{M}_{:,k}|_1 \quad (2)$$

where λ is a positive parameter that balances between the two terms, r_1 is the maximum time lag associated with traffic data, r_2 is the maximum time lag associated with traffic indicators based on tweet semantics, $\mathbf{1}_{1\times n^{t-l}}$ is a row vector of 1s, $|\cdot|_1$ is the $l1$ norm, $\mathbf{S}^{t-1}_{:,k}$ and $\mathbf{M}_{:,k}$ denotes the k^{th} column of $\mathbf{S}^{t-1}$ and $\mathbf{M}$ respectively.

From Equation 2, we can see that the objective function consists of two terms: the first term measures the prediction error of $\delta\mathbf{v}^t$ using the linear regression model with lags up to r_1 for traffic data and lags up to r_2 for traffic indicators based on tweet semantics; and the second term imposes sparsity on each column of $\mathbf{M}$.

Furthermore, regarding the number of traffic indicators based on tweet semantics, i.e., the number of columns of $\mathbf{M}$, we have the following lemma.

Lemma. $\forall m > 1$, the optimal solution to Equation 2 is equivalent to the optimal solution with $m = 1$.

Proof sketch. For any matrix $\mathbf{M}$ with m columns, we can generate a vector $\mathbf{m}$ by adding all the columns of $\mathbf{M}$ together. The value of the objective function in Equation 2 is the same with $\mathbf{M}$ and $\mathbf{m}$.■

Based on the above lemma, Equation 2 can be simplified as follows.

$$\min_{\mathbf{m},\alpha,\beta_l,\gamma_l} \sum_{t=\max(r_1,r_2)+1}^{T} (\delta\mathbf{v}^t - \alpha - \sum_{l=1}^{r_1} \beta_l \delta\mathbf{v}^{t-l} - \sum_{l=1}^{r_2} \gamma_l \mathbf{1}_{1\times n^{t-l}} \mathbf{F}^{t-l}\mathbf{m})^2 + \lambda|\mathbf{m}|_1 \quad (3)$$

Equation 3 can be solved using the following iterative algorithm. It works as follows. We first initialize $\mathbf{m}$ to be a vector of all zeros, which indicates that no tweet semantics are used. Then, in each iteration, we solve for α, β_l ($l = 1, \ldots, r_1$), γ_l ($l = 1, \ldots, r_2$), and $\mathbf{m}$ in an alternating way.

Algorithm 1 Iterative Algorithm for solving Equation 3

Require: $\mathbf{F}^t$, $\delta\mathbf{v}^t$ ($t = 1, \ldots, T$), r_1, r_2, λ, n_{iter}
Ensure: α, β_l ($l = 1, \ldots, r_1$), γ_l ($l = 1, \ldots, r_2$), $\mathbf{m}$

1: Initialize $\mathbf{m}$ to be a vector of all zeros.
2: **for** $n = 1$ to n_{iter} **do**
3: Fix $\mathbf{m}$, and solve for α, β_l ($l = 1, \ldots, r_1$), and γ_l ($l = 1, \ldots, r_2$) via linear regression.
4: Fix α, β_l ($l = 1, \ldots, r_1$), and γ_l ($l = 1, \ldots, r_2$), and solve for the transformation vector using *glmnet* [Friedman *et al.*, 2010].
5: **end for**

Furthermore, solving the original traffic prediction model based on auto-regression and the model in Equation 1 can be seen as special cases of Equation 3. To see this, if we set $\mathbf{m}$ to be a zero vector, we get the original traffic prediction model without tweet information; on the other hand, if we set $\mathbf{m}$ to be a zero vector except for the last element, which is set to 1, and set $r_1 = r_2 = 2$, we get the model in Equation 1. Therefore, the value of the objective function with the optimal vector $\mathbf{m}$ is at least as good as the original traffic prediction model and the model in Equation 1.

5 Experimental Results

In this section, we test the performance of the proposed framework, and compare with the models based on traffic intensity only, and both traffic intensity and social activity intensity. To be specific, throughout our experiments, we apply the following three models:

1. Model 1: traffic intensity only;
2. Model 2: traffic intensity and social activity intensity (based on the model in Equation 1);
3. Model 3: traffic intensity and tweet semantics (based on the model in Equation 3) using Algorithm 1.

In the third model, we test its performance using two versions of the feature matrix $\mathbf{F}^t$: one with binary values, and the other with tf-idf values. For both versions, the number of columns (e.g., the number of stemmed words) is 161,914, and the number of rows depends on the time stamp.

In Equation 3, ideally, the maximum time lag r_1 for traffic data and r_2 for Twitter data in each of the three models should be tuned by cross-validation. Empirical studies in traffic prediction practice typically suggest optimal values for r_1 ranging from 2 to 6 [Kamarianakis and Prastacos, 2003; Min and Wynter, 2011]. For r_2, the results in Table 1 suggest that the most recent or the second most recent covariate associated with social activity tend to be statistically significant for predicting traffic. Therefore, for the purpose of concept demonstration, in our experiments, each model incorporates exactly the two most recent time lags for both traffic data and Twitter data. In other words, for the first model, $\delta\mathbf{v}^{t-1}$ and $\delta\mathbf{v}^{t-2}$ are used to predict $\delta\mathbf{v}^t$; for the second model, $\delta\mathbf{v}^{t-1}$, $\delta\mathbf{v}^{t-2}$, $\delta\mathbf{c}^{t-1}$, and $\delta\mathbf{c}^{t-2}$ are used; for the third model, $\delta\mathbf{v}^{t-1}$, $\delta\mathbf{v}^{t-2}$, $\mathbf{F}^{t-1}$, $\mathbf{F}^{t-2}$ are used.

The entire data set consisting of both traffic data and Twitter data are partitioned into two parts, with the beginning $(s-1)/s$ ($s = 3, 4, 5, 6, 7$) as the training set and the remaining as the test set. Model estimation and prediction are performed with various time resolutions. For the training data set, the models are estimated through 5-fold cross-validation.

The prediction performance is evaluated by two measures, namely Mean Absolute Percentage Error (MAPE) and Root Mean Square Error (RMSE), which are calculated as follows

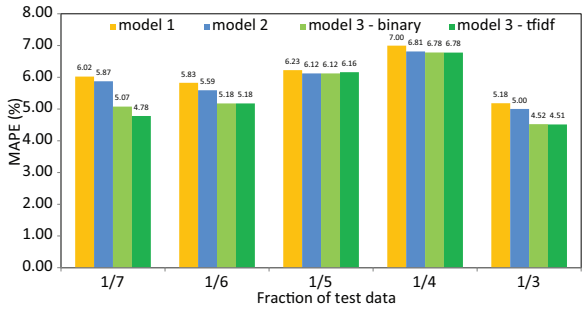
$$\text{MAPE} = \frac{1}{T - \max(r_1, r_2)} \sum_{t=r+1}^{T} \left(\frac{|\delta\mathbf{v}^t - \delta\hat{\mathbf{v}}^t|}{\mathbf{v}^t}\right)$$

$$\text{RMSE} = \sqrt{\frac{1}{T - \max(r_1, r_2)} \sum_{t=r+1}^{T} (\delta\mathbf{v}^t - \delta\hat{\mathbf{v}}^t)^2}$$

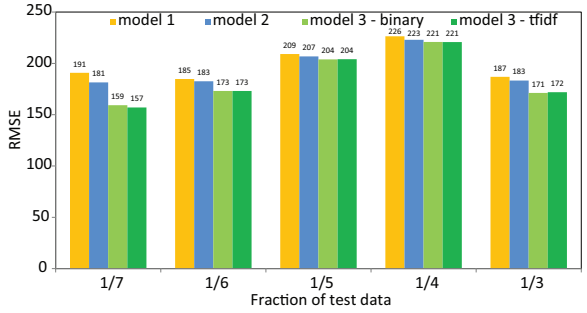
where $\delta\hat{\mathbf{v}}^t$ denotes the estimated value of $\delta\mathbf{v}^t$.

Figure 3 shows the comparison results of the three models in terms of both MAPE and RMSE for time resolution of 12 hours. The results for other time resolutions are similar and hence omitted for brevity. From this figure, we can see that the information in social media indeed helps improve the performance of traffic prediction. To be specific, by including the tweet counts as the covariates, the second model performs better than the first one, which is only using traffic information; by leveraging the traffic indicators based on tweet semantics, the third model further improves the performance in terms of both MAPE and RMSE. Furthermore, the difference between using binary valued and tf-idf valued feature matrices is not significant in most cases. This might be explained by the fact that the presence of certain keywords (instead of the frequency) is enough to characterize the traffic condition.

For illustration purpose, we also show in Figure 4 the profile of predicted values vs. the true values for a sample partition where the last 1/7 of the data is used as the test data.



(a) MAPE



(b) RMSE

Figure 3: Comparison of different models

From this figure, we can see that our proposed model (model 3) tracks the fluctuation in the traffic volume better model 1 (using traffic information only) and model 2 (using both traffic information and tweet counts): in the first 3 intervals, the predicted values using model 3 are closer to the true values than the other two models; and in the remaining 8 intervals, the predicted values using model 3 well approximate the true values.

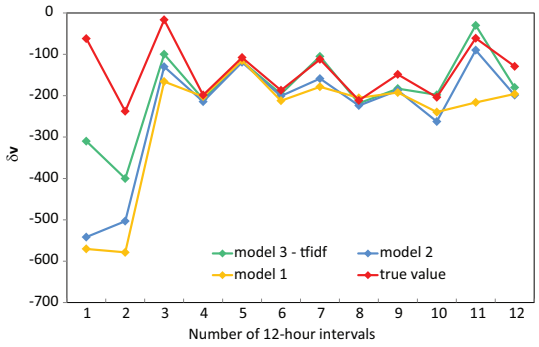


Figure 4: Prediction profile of different models (unit time stamp = 12 hours)

Furthermore, by including the $l1$ norm in the objective function, our proposed framework generates a sparse transformation vector $\mathbf{m}$, which helps us understand the way social media affects traffic. In other words, the non-zero elements in $\mathbf{m}$ correspond to the key words in tweets that indicate traffic conditions, such as ucberkeley, albany (which is a city close to Berkeley in the Bay Area), Friday, giants, etc. Interestingly, the word *giants* is the name for San Francisco baseball team, which indicates that sports-related activities are a key factor in traffic prediction.

6 Conclusion

In this paper, motivated by the fact that people tend to post traffic-related content in social media, we answer the following question: can we leverage such information to improve traffic prediction. To this end, we first perform correlation analysis between traffic measurements and tweet counts, and then propose a general optimization framework to extract traffic indicators based on tweet semantics. Experimental results on traffic data and Twitter data collected from the San Francisco Bay area between August 3, 2011 and September 30, 2011 demonstrate the improved performance of our model over the existing traffic prediction model based on auto-regression.

References

[Ahmed and Cook, 1979] M. S. Ahmed and A. R. Cook. Analysis of freeway traffic time-series data by using box-jenkins techniques. *Transportation Research Board*, 722:1–9, 1979.

[Al-Deek *et al.*, 2001] H. Al-Deek, S. Ishak, and M. Wang. A new short-term traffic prediction and incident detection system on i-4, vol. i. final research report. Technical report, Transportation Systems Institute(TSI), Department of Civil and Environmental Engineering, University of Central Florida, 2001.

[Bollen *et al.*, 2010] Johan Bollen, Huina Mao, and Xiao-Jun Zeng. Twitter mood predicts the stock market. *CoRR*, abs/1010.3003, 2010.

[Clark *et al.*, 1993] S. D. Clark, M. S. Dougherty, and H. R. Kirby. The use of neural networks and time series models for short-term traffic forecasting: a comparative study. *Proceedings of the PTRC 21st Summer Annual Meeting*, 1993.

[Clark, 2003] S. Clark. Traffic prediction using multivariate nonparametric regression. *Journal of Transportation Engineering*, 129:161–168, 2003.

[Eisenstein *et al.*, 2010] Jacob Eisenstein, Brendan O'Connor, Noah A. Smith, and Eric P. Xing. A latent variable model for geographic lexical variation. In *EMNLP*, pages 1277–1287, 2010.

[Friedman *et al.*, 2010] J. H. Friedman, T. Hastie, and R. Tibshirani. Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software*, 33(1):1–22, 2 2010.

[Go *et al.*, 2011] Alec Go, Richa Bhayani, and Lei Huang. Twitter sentiment classification using distant supervision. Technical report, Stanford University, 2011.

[Guo and Williams, 2010] J. Guo and B. M. Williams. Real-time short-term traffic speed level forecasting and uncertainty quantification using layered kalman fiters. *Transportation Research Record*, 2175:28–37, 2010.

[Huang and Sadek, 2009] S. Huang and A. W. Sadek. A novel forecasting approach inspired by human memory: The example of short-term traffic volume forecasting. *Transportation Research Part C*, 17:510–525, 2009.

[Kamarianakis and Prastacos, 2003] Y. Kamarianakis and P. Prastacos. Forecasting traffic flow conditions in an urban network: comparison of multivariate and univariate approaches. *Transportation Research Record*, 1858:74–84, 2003.

[Khosravi *et al.*, 2011] A. Khosravi, E. Mazloumi, S. Nahavandi, D. Creighton, and J. Van Lint. A genetic algorithm-based method for improving quality of travel time prediction intervals (in press). *Transportation Research Part C*, 2011.

[Levin and Tsao, 1980] M. Levin and Y.-D. Tsao. On forecasting freeway occupancies and volumes. *Transportation Research Record*, 773:47–49, 1980.

[Mahmassani *et al.*, 2009] H. S. Mahmassani, J. Dong, J. Kim, R. B. Chen, and B. Park. Incorporating weather impacts in traffic estimation and prediction systems. Technical Report FHWA-JPO-09-065, Northwestern University, 2009.

[Maze *et al.*, 2006] T.H. Maze, M. Agarwal, and G. Burchett. Whether weather matters to traffic demand, traffic safety, and traffic operations and flows. *Transportation Research Record*, 1948:170–176, 2006.

[Min and Wynter, 2011] W. Min and L. Wynter. Real-time road traffic prediction with spatio-temporal correlations. *Transportation Research Part C*, 19:606–616, 2011.

[Okutani and Stephanedes, 1984] I. Okutani and Y. J. Stephanedes. Dynamic prediction of traffic volume through kalman filtering theory. *Transportation Research Part B*, 18:1–11, 1984.

[Petrovic *et al.*, 2010] Sasa Petrovic, Miles Osborne, and Victor Lavrenko. Streaming first story detection with application to twitter. In *HLT-NAACL*, pages 181–189, 2010.

[Sakaki *et al.*, 2010] Takeshi Sakaki, Makoto Okazaki, and Yutaka Matsuo. Earthquake shakes twitter users: real-time event detection by social sensors. In *WWW*, pages 851–860, 2010.

[Sankaranarayanan *et al.*, 2009] Jagan Sankaranarayanan, Hanan Samet, Benjamin E. Teitler, Michael D. Lieberman, and Jon Sperling. Twitterstand: news in tweets. In *GIS*, pages 42–51, 2009.

[Smith and Demetsky, 1996] B. L. Smith and M. J. Demetsky. Multiple-interval freeway traffic flow forecasting. *Transportation Research Record*, 1554:136–141, 1996.

[Smith and Demetsky, 1997] B. L. Smith and M. J. Demetsky. Traffic flow forecasting: comparison of modelling approaches. *Journal of Transportation Engineering*, 123(4):261–266, 1997.

[Smith *et al.*, 2002] B. L. Smith, B. M. Williams, and R. K. Oswalsd. Comparison of parametric and nonparametric models for traffic flow forecasting. *Transportation Research Part C*, 10:303–321, 2002.

[Sriram *et al.*, 2010] Bharath Sriram, Dave Fuhry, Engin Demir, Hakan Ferhatosmanoglu, and Murat Demirbas. Short text classification in twitter to improve information filtering. In *SIGIR*, pages 841–842, 2010.

[van Lint *et al.*, 2005] J. van Lint, S. Hoogendoorn, and H. van Zuylen. Accurate freeway travel time prediction with state-space neural networks under missing data. *Transportation Research Part C*, 13:347–369, 2005.

[Vlahogianni *et al.*, 2005] E. Vlahogianni, M. G. Karlaftis, and J. C. Golias. Optimized and meta-optimized neural networks for short-term traffic flow prediction: A genetic approach. *Transportation Research Part C*, 13:211–234, 2005.

[Vythoulkas, 1993] P. C. Vythoulkas. Alternative approaches to short-term traffic forecasting for use in driver information systems. In *Transportation and Traffic Theory, Proceedings of the 12th International Symposium on Traffic Flow Theory and Transportation*, Berkeley, CA, July 1993.

[Williams *et al.*, 1998] B. M. Williams, P.K. Durvasula, and D. E. Brown. Urban traffic flow prediction: application of seasonal autoregressive integrated moving average and exponential smoothing models. *Transportation Research Record*, 1644:132–144, 1998.

[Yun *et al.*, 1998] S.-Y. Yun, S. Namkoong, J.-H. Rho, S.-W. Shin, and J.-U. Choi. A performance evaluation of neural network models in traffic volume forecasting. *Mathematical Computer Modelling*, 27:293–310, 1998.