Twitter-informed Crowd Flow Prediction

Gary Goh, Jing Yu Koh
Singapore University of Technology and Design
8 Somapah Road, Singapore 487372
e-mail: {gary_goh, jingyu_koh}@mymail.sutd.edu.sg

Yue Zhang*
Department of Engineering, Westlake University
18 Shi Long Shan Road, Hangzhou 310024
e-mail: yue.zhang@wias.org.cn

***Abstract*— We explore the usage of real-time tweets to explain for non-recurring large-scale spatio-temporal crowd movement. The aim is to evaluate the usefulness of tweets to improve the performance of city-wide crowd flow prediction. We conduct experiments in the context of Singapore city to investigate our proposition by extending upon an existing crowd flow prediction model. Implemented using a deep-neural-network-based approach, an end-to-end predictive model is configured to take in tweets as additional inputs to forecast the future flow of crowds in an urban environment. We extract various features from tweets, such as tweet counts, tweet tenses and sentiments as additional signals to the predictive model. From the experimental results, we show that some models are able to improve the prediction accuracy, and share our insights on how tweets are related to crowd flows.**

Keywords*— *Spatio-temporal prediction; deep learning; twitter-informed.

I. INTRODUCTION

Fine-grained crowd flow prediction within a city is valuable for traffic control management and could improve travelling experience and public safety. Crowd flows refers to traffic flows that are aggregated spatially over a region in a city, and temporally over a time interval. Accurate knowledge of future crowd flows could lead to better travel time estimations and more optimized route selection during navigation for general users [1]. It could also facilitate governments and/or urban city planners to strategize and enforce targeted traffic control measures in advance to curb the level of congestion in the city, and could potentially be very useful in averting overcrowding situations in specific regions. For example, in September 2017, a huge crowd of people gathered together at a train station in Mumbai on a rainy morning rush hour when four trains arrived simultaneously, resulting in a tragic stampeded that killed 23 people. 36 people also died in a stampede during Shanghai's 2015 New Year's Eve celebration. These tragedies could have been avoided or at least mitigated if authorities are advised with future crowd flow predictions and take early preventive measures, such as setting up blockades, broadcasting warnings, or conducting evacuations.

Many studies have been conducted on the traffic flows prediction problem, and recent work achieves relatively reasonable accuracies [2-5]. These approaches focus on capturing patterns from historical observations of traffic flows to predict future observations. Due to the nature that traffic flows are largely periodical, such as the predictable peaks in the morning and evening rush hours, relying on past observations is generally effective. However, the poor predictive performance arises when there are non-recurring events that can influence large-scale crowd movement, which cannot be inferred from historical data. Examples of such events include, traffic incidents, road closures, road works, sports events, musical concerts, celebratory events, or any other events that cause sudden interests in particular regions such as the sudden congregation of "Pokemon Go" players in specific random spawn locations to catch rare in-game creatures. These events can be rare and only affect small regions in short time intervals, yet it is especially during these critical periods, that accentuates the need for more accurate crowd flow predictions so that the relevant authorities can respond timelier to the situation.

We consider utilizing real-time texts from the Internet, which can contain information on such critical non-recurring events, by feeding them as additional inputs to an existing crowd flow prediction baseline model. More specifically, we focus on tweets to represent non-recurring crowd flows influencing information, as it has been demonstrated that Twitter is able to react to news events more quickly when compared with traditional media [6]. It presents a huge well of untapped freely available information, which explains the extensive research attention on tweets information extraction in recent years. In this paper, we aim to address the following research questions:

1. Can tweets be useful for crowd flow prediction?
2. How are tweets related to traffic / crowd flows?

We analytically answer the above questions through conducting empirical experiments in the context of Singapore, experimenting with two traffic flow datasets – vehicular inductive loop detectors' signals across the city to measure the number of vehicles passing on roads, and mobile phone signals to measure the number of people moving from point to point. We employ the Spatio-temporal Residual Network (ST-ResNet) crowd flow prediction model, as proposed by Zhang, Zheng and Qi [5], as our baseline. ST-ResNet is an end-to-end deep neural network predictive model built to forecast the citywide crowd flows, which has been shown useful for predicting crowd flows in Beijing and New York, but not yet for Singapore. Its design is also highly flexible and allows easy integration of additional inputs.

Incorporating tweets as inputs to an end-to-end structured prediction model is challenging due to the large presence of

* The work is done while the third author worked at SUTD.

noise found in unstructured text. Some tweets may be irrelevant to the search keywords used to extract them. In addition, efficiency is a crucial factor for real-time metropolitan-level crowd flow prediction. Thus, we explore several efficient linguistic features such as tweet counts, tweet tenses and tweet sentiments, which are relatively insensitive to noise, and extend upon our baseline model to receive tweet information as additional inputs. Results over four years of data suggest that tweet information is indeed relevant to traffic, significantly reducing prediction errors for both road traffic and mobile phone signals. We additionally find that people tend to tweet more nearing relevant traffic-influencing events, which suggests that tweets are good indicators to crowd flows. To our knowledge, we are the first to investigate the usage of tweets to the crowd flow prediction task. Regrettably, we are unable to release the traffic flow datasets due to confidentiality issues. However, our code for the extended model and the tweets dataset are publicly available[1], and we strongly encourage readers to reproduce our results in the context of other cities.

The rest of the paper is organized as follows. Section II introduces our problem statement formally along with the datasets we used in our experiments. Section III summarizes the internal structures of ST-ResNet and how it is configured to take tweets inputs. The experiment settings are detailed in Section IV, along with the results and our discussions. Section V presents related work in this field. Finally, Section VI concludes the paper.

II. Problem Statement

There are numerous ways to define spatial regions, but for the context of our study, we use an evenly-spaced grid map of dimensions I×J to partition a city, where each grid cell denotes a spatial region. The size of the map is based on the limits of the latitudes and longitudes. Each of the grid cell contains two types of crowd flows: inflows and outflows, as illustrated in Fig. 1. Together, crowd flows for every region within a time interval can be represented by a 3-dimensional image-like matrix with 2 channels; one for inflow and the other for outflow.

At the t^{th} time interval, the crowd flows in all $I \times J$ regions is denoted as a tensor $\boldsymbol{X}_t \in \mathbb{R}^{2 \times I \times J}$ where $(\boldsymbol{X}_t)_{0,i,j} = x_t^{in,i,j}$ denotes the inflows and $(\boldsymbol{X}_t)_{1,i,j} = x_t^{out,i,j}$ denotes the outflows. The crowd flow prediction problem becomes a rolling horizon time series prediction task, where the aim is to predict the next time interval's image. Formally, the crowd flow prediction problem is defined as: given historical observations $\{\boldsymbol{X}_t | t = 0, \ldots, n\text{-}1\}$ and any additional inputs from external factors, predict $\boldsymbol{X}_n$.

For our experiments, we used two different sets of citywide traffic flow data from Singapore to construct the crowd flow tensors. Their metadata is tabulated in Table I. We use multiple datasets from different time spans so as to validate the performance of the predictive model, since cross validation is not feasible for the rolling horizon prediction problem.

[1] https://github.com/garygsw/twitter-crowd-flow-prediction

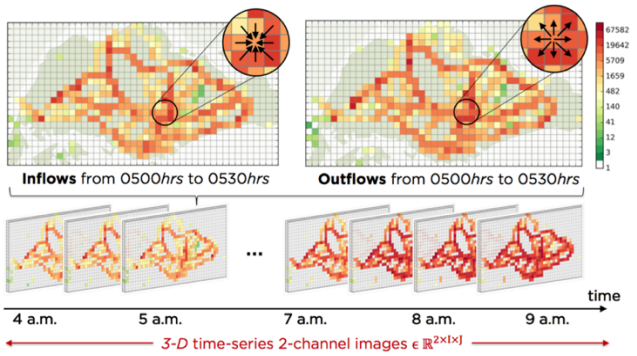


Figure 1. Visual representation of spatio-temporal crowd flows with a time series of 3D 2-channel images.

TABLE I. Crowd Flow Datasets Description

Dataset	VLD	MPS
Data type	Vehicular counts	Origin-destination pairs
Time span	Set 1: 1/3/2013 – 30/6/2013 Set 2: 1/9/2014 – 31/12/2014 Set 3: 1/12/2015 – 31/3/2016	1/8/2017 – 30/11/2017
Time interval	30 minutes	1 hour
# of time intervals (T)	5,856	1,464
Grid map size (I, J)	(89, 49)	(90, 54)

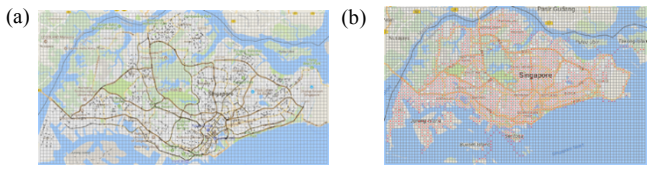


Figure 2. Singapore map overlaid with the grid map. (a) lines denote road links with installed VLDs; (b) points denote traversable grids.

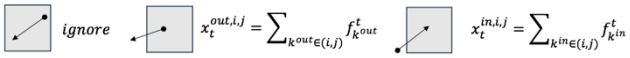


Figure 3. Aggregation to type of flows depending on direction and position of VLD. Here, f_k^t represents the flow in road link k time interval t.

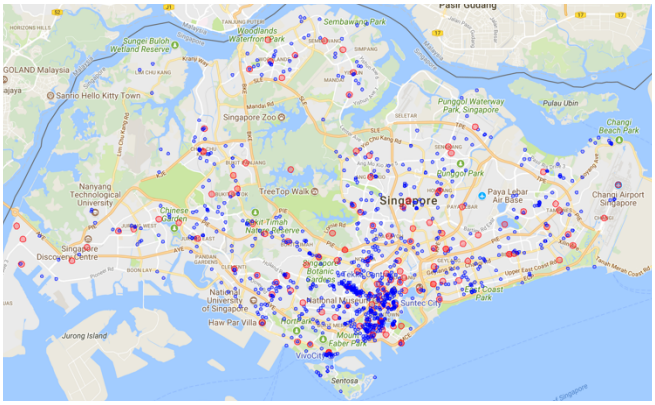


Figure 4. Singapore map overlaid with selected locations of with tweet mentions extracted. Red points denote train stations; others are in blue.

In addition, we collected weather and public holidays datasets, as well as the extracted tweets information from a set of tweets. All of these datasets corresponds to the time span of the traffic flow datasets. The dataset preparation process is as follows:

A. Vehicular Loop Detectors Datasets

We aggregate signals from over 57,000 vehicular inductive loop detectors (VLDs) that are installed across major road intersections and expressways in Singapore, as shown in Fig. 2a. These sensors count the number of vehicles that passes through each point.

Let $\mathbb{Q}$ be a collection of VLDs signals at the t^{th} time interval. For grid cell (i,j) that refers to the region in the i^{th} row and the j^{th} column, the aggregated inflows and outflows at the time interval t are defined respectively as:

$$x_t^{in,i,j} = \sum_{q \in \mathbb{Q}} |q|, \quad q_s \notin (i,j) \wedge q_e \in (i,j) \tag{1}$$

$$x_t^{out,i,j} = \sum_{q \in \mathbb{Q}} |q|, \quad q_s \in (i,j) \wedge q_e \notin (i,j) \tag{2}$$

where $|q|$ is the value of the VLD signal in $\mathbb{Q}$; $q_s \to q_e$ represents a VLD signal starting from point q_s and ending with point q_e; $q_i \in (i,j)$ means that point q_i lies within grid cell (i,j), and vice versa (also, see Fig. 3).

B. Mobile Phone Signals

We aggregate mobile phone signals (MPSs), where the location shifts of mobile phone users are estimated based on the closest cellular tower connected to their phones. As the origin and destination points could be situated quite far away, there is a need to infer their most likely trajectories in their path through the grid in order to construct the crowd flow tensors. A map of traversable regions is marked out so as to ensure all points are reachable by any point, as shown in Fig. 2b. Subsequently, a breadth-first search shortest path algorithm is used to infer the trajectories for every origin-destination pairs.

Let $\mathbb{P}$ be a collection of trajectories at the t^{th} time interval. For grid cell (i,j), the aggregated inflows and outflows at the time interval t are defined respectively as:

$$x_t^{in,i,j} = \sum_{Tr \in \mathbb{P}} |p|, \quad \forall\, k > 1, \quad g_{k-1} \notin (i,j) \wedge g_k \in (i,j) \tag{3}$$

$$x_t^{out,i,j} = \sum_{Tr \in \mathbb{P}} |p|, \quad \forall\, k > 1, \quad g_{k-1} \in (i,j) \wedge g_k \notin (i,j) \tag{4}$$

where $Tr: g_1 \to g_2 \to \cdots \to g_{|Tr|}$ is a trajectory in $\mathbb{P}$, and g_k are points along the trajectory; $|p|$ is the number of people travelling on trajectory Tr.

C. External Data

Weather information is scrapped from a website which provides historical hourly weather information (`www.timeanddate.com`). The sub-factors include temperature, wind speeds, and one-hot-vectors to represent one of the 8 different weather conditions – sunny, cloudy, overcast, rain, light rain, heavy rain, fog and haze. The public holidays and weekends can be inferred from the calendar and is encode by a binary vector that correspond to every time interval.

TABLE II. EXAMPLES OF TWEET SEARCH KEYWORDS USED

Train stations	Points of interests	Estate names
Braddell	Esplanade Theatre	Serangoon
Dhoby Ghaut	Marina Bay	Tampines
Outram Park	Singapore Indoor Stadium	Clementi

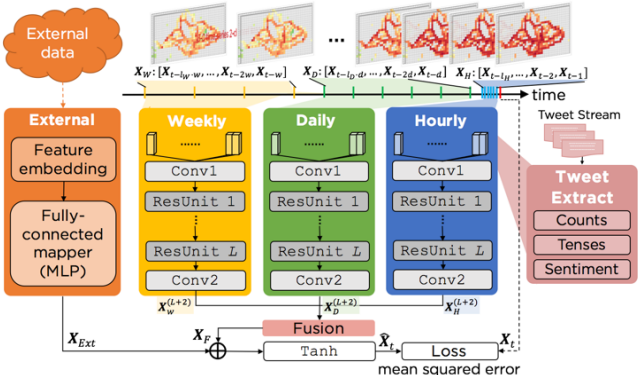


Figure 5. ST-ResNet overall architecture with extended tweet extract component.

D. Tweets Data

Our tweets are collected based on a set of 369 search keywords that are appropriately selected to cover key regions within Singapore, extracting a total of 1.28 million tweets. Several keywords are shown in Table II. The full list of search keywords is attached together with our code.

These keywords are based on the names of locations with high capacity to hold large crowds, as well as names of regions which are representative of localized regions, such as train stations names, estate names, towns, campuses etc. The spatial distribution of these locations are shown in Fig. 4. The tweets are aggregated spatially based on the latitudes and longitudes of the location of its search keyword, and temporally based on its creation time.

Here we have a $T \times I \times J$ matrix containing a set of tweets relevant to each grid cell (i,j), where T is the total number of time intervals in our datasets.

III. MODEL

An overview of the ST-ResNet model's architecture is shown in Fig. 5. The aim is to predict the next crowd flow at time t. The original model comprises of four components. The first three components in the middle possess the exact same structure, but each of them models the spatio-temporal correlations from historical observations at different time granularities – weekly, daily and hourly. The fourth component, shown on the left of Fig. 5, considers external factors that affect the crowd flows across the entire city, such as the weather, day of the week and public holidays. We extend the model by adding a fifth component, which extracts a set of features from a tweet stream and appends them to the inputs of the hourly component before feeding it through the neural networks. The following sub-sections describes each component more attentively, but we refer readers to Zhang, Zheng and Qi's paper [5] for more details.

A. Historical Observations Component

The weekly, daily and hourly components takes in an ordered concatenated series of crowd flow matrices from past observations but at three different lengths of time intervals apart denoted by W, D and H respectively. The size of each of the sequence is parameterized by l_W, l_D and l_H. The idea behind the choice of these time granularities originates from how traffic patterns are usually found in these time intervals. These two main elements form the internal structure of this component:

Convolutions. The idea is to stack multiple convolutional neural networks (CNNs) to explore spatio-temporal correlations between nearby and distant inflows and outflows across different historical time intervals. Each convolutional mapping is denoted by *Conv*, and is defined as:

$$\boldsymbol{X}^{(i+1)} = f\left(\boldsymbol{W}^{(i)} * \boldsymbol{X}^{(i)} + \boldsymbol{b}^{(i)}\right) \quad (5)$$

where $*$ denotes the convolution function; f is the rectified linear unit (ReLU) activation function; $\boldsymbol{W}^{(i)}$, $\boldsymbol{b}^{i}$ are learnable parameters; and i is the index of each stack.

Residual Units. To be able to stack multiple CNNs without incurring training degradation, L number of residual units are used, where each unit denoted by ResUnit. Each residual unit defines the mapping as follows:

$$\boldsymbol{X}^{(l+1)} = \boldsymbol{X}^{l} + \mathcal{F}\left(\boldsymbol{X}^{(l)}; \boldsymbol{\theta}^{(l)}\right), \qquad l = 1, \ldots, L \quad (6)$$

where $\mathcal{F}$ is the residual function, where it contains two stacks of convolution with ReLU, and $\boldsymbol{\theta}^{(l)}$ includes all learnable parameters in the l^{th} residual unit.

Following the original implementation of the model, batch normalization is also added before applying ReLU. The final part of each component is joined by a last convolution layer *Conv2* where the dimensions of the outputs will match the original dimensions of the crowd flow tensor. The outputs for each component are denoted by $\boldsymbol{X}_H^{(L+2)}$, $\boldsymbol{X}_D^{(L+2)}$ and $\boldsymbol{X}_W^{(L+2)}$.

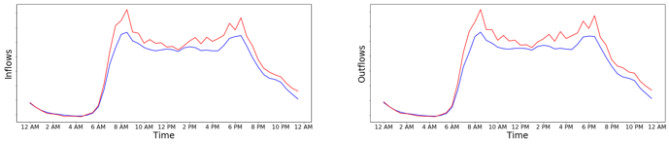
B. External Component

The external component considers exogenous knowledge which have been shown to be crucial in influencing future crowd flows. Crowd flows during public holidays or weekends can be considerably different compared to flows during normal work days or weekdays. Weather also plays an important factor in determining the behavior of crowd flows. Lagged weather conditions (i.e. weather at t-1) is used to forecast the crowd flows at time interval t.

C. Integrating Tweet Features

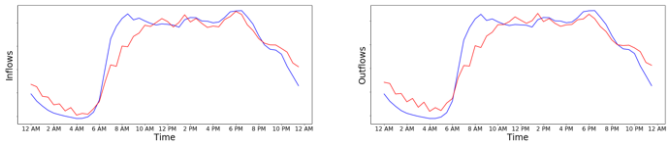
In this component, we extract features from real-time tweets that might be relevant in helping explain for non-recurring crowd flows by introducing two new parameters, denoted by $(lag^{-}, lead^{+})$ which represent a time interval window of tweets to be included in the inputs. In particular, lag^{-} refers to the number of time intervals before t during which the tweet features are extracted, while $lead^{+}$ refers to the number of time intervals from t onwards. The size of the source tweet information window is thus $lag^{-} + lead^{+}$. For real-time prediction, it would be infeasible to consider any $lead^{+} > 0$ since it will be unknown at time interval t-1. However, we included $lead^{+}$ in our experiments as we intend to analyze potential relationships that future tweets might have with future crowd flows. The features that are extracted from tweets include:

Time	Tweet text
17:38	Now at Jalan Besar stadium alr!!
17:41	Now going to jalan besar for ball picker
19:12	Finally the jam cleared. Now en route to Jalan Besar !!
19:34	Kickoff at the Jalan Besar Stadium - Albirex Niigata (S) 0-0 Home United! #SLeague
19:50	A bit late but thrilled to be watching #ALB v #HUFC @Jalan Besar Stadium



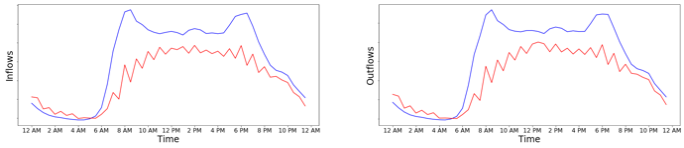
(a) VLDs' crowd flows (bottom) in nearby regions during a present large-scale event on 1 March 2013 around 7.30pm and a sample of the relevant tweets (top). Crowd flows are higher than expected.

Time	Tweet text
16:30	@Ai_Arakawa FALL OUT BOY IS COMING TO SG. AUG 6 @FORT CANNING . TICKETS AT SISTIC. HOSTED BY @LiveEmpire #FOBinSG
16:31	@qatarairways FALL OUT BOY IS COMING TO SG. AUG 6 @FORT CANNING . TICKETS AT SISTIC. HOSTED BY @LiveEmpire #FOBinSG
…	… x 121



(b) VLDs' crowd flows (bottom) in nearby regions on 29 June 2013 and a sample of the irrelevant tweets mentions of a far future event on 6 August 2013 (top). Crowd flows in that region on that day is comparable to the daily average.

Time	Tweet text
10:58	Wah f*ck the road near paya lebar mrt flood like shit its fucking knee deep in water !!!
10:59	wow rain till gt flood at paya
10:59	Oh my gosh... Paya Lebar is flooded. Literally.. Like shin-level.
11:09	Water level falls below 90%. High Flood Risk.11:09:13 #SgFlood,,,#SgFlood



(c) VLDs' Crowd flows (bottom) in nearby regions during a negatively-potrayed event on 28 April 2013 around 11am and the relevant tweets (top). Crowd flows are lower than expected.

Figure 6. Case studies to support the choice of tweet features extracted. *Note: Blue lines represent daily mean flows, while the red lines represent specific flows on the day.*

Tweet counts. The simplest way to represent the level of interests in a particular location is to simply track the counts of tweets that refers to the region. Based on our hypothesis, if the tweet counts from a particular grid cell is high at some specific time interval, crowd flows from nearby cells should be higher than normal. The intuition is that if a particular region gains interests as measured by tweet counts, crowds are more likely to flow there. For example, as shown in Fig. 6a, there is a spike in the counts of tweets that has specific mentions of "Jalan Besar Stadium", which turns out to be referencing a large-scale soccer match with higher levels of induced crowd flows into the region throughout the day. This feature is denoted as $\boldsymbol{T}_C \in \mathbb{R}^{I \times J}$.

Tweet tenses. Although the tweets are collected on a real-time basis, the tweets might not refer to an event that is happening at present time. People might tweet about some event that has already happened in the past, or will only happen in the future. For example, as shown in Fig. 6b, there was a huge spike of tweets with mentions of "Fort Canning" referring to a far future concert event that has no relation to the present, and did not contribute to any anomalies in the crowd flows in the corresponding nearby regions. Hence, solely basing on tweet counts to measure current interests level in a region without considering the time dimension might not be the most accurate. Tenses information is derived from Parts-of-speech (POS) tags of the root verb. We take the tags <VBD> and <VBN> for past tense, <VBG>, <VBZ>, and <VBP> for present tense, and <MD> for future tense, and take their counts. We use the Stanford POS tagger to obtain the tags [7]. This feature is denoted as $\boldsymbol{T}_T \in \mathbb{R}^{3 \times I \times J}$.

Tweet sentiment. So far, we assumed that a high spike of interest at a particular location induces large crowd flows around it. However, this assumption may not hold especially when the interests are negative, and instead may suggest the opposite by reducing crowd flows in the region. Examples of such events include a last-minute cancellation of an event or outbreak of a disastrous event. For example, an undesirable major flooding incident happened in Paya Lebar, and was discussed heavily on Twitter. However, this increased spike of tweet counts did not induce additional crowd flows into the region but instead did the opposite (see Fig. 6c). This is intuitive as people are less inclined to travel in regions that are negatively portrayed. Inversely, we also expect crowd flows in positively interpreted regions to surge. Thus, we also explore adding sentiment as an extra source of information to help measure the degree of such scenarios. Tweet sentiment information is extracted using a simple counting of positive and negative words based on a manually annotated sentiment lexicon by Hu and Liu [8]. This feature is denoted as $\boldsymbol{T}_S \in \mathbb{R}^{2 \times I \times J}$.

For each of the sub-features, a sequence of matrices that corresponds to the tweets information time window is prepared, and aggregated via simple summation to obtain a single matrix. Finally, they are concatenated with the input sequence in the hourly component to allow the model to also explore dependencies between tweets and the crowd flow matrices. We choose to merge the tweet features with the hourly component's inputs because our underlying intention is to use tweets to model short-term effects to crowd flows. These three features are specially designed to be simple, insensitive to noise, and highly efficient to extract from tweets, as opposed to using more complex methods such as neural networks to represent tweet information, which is important for real-time prediction to be effective.

D. Data Fusion

Finally, the outputs from the components, $\boldsymbol{X}_H^{(L+2)}$, $\boldsymbol{X}_D^{(L+2)}$, $\boldsymbol{X}_W^{(L+2)}$, and $\boldsymbol{X}_{Ext}$ are fused together to produce a prediction tensor $\bar{\boldsymbol{X}}_t$. The fusion is performed in two steps as follows:

Parametric-matrix-based fusion. The first step fused the first three historical observations components via a parametric-matrix-based method to form $\boldsymbol{X}_F$ with the following mapping:

$$\boldsymbol{X}_F = \boldsymbol{W}_W \circ \boldsymbol{X}_W^{(L+2)} + \boldsymbol{W}_D \circ \boldsymbol{X}_D^{(L+2)} + \boldsymbol{W}_H \circ \boldsymbol{X}_H^{(L+2)} \tag{7}$$

where $\circ$ denotes the element-wise multiplication operator; $\boldsymbol{W}_W$, $\boldsymbol{W}_D$ and $\boldsymbol{W}_H$ are learnable parameters that fine-tune the level of effect from the weekly, daily and hourly component on each grid cell, allowing the model to specify the level of effect from each past observation component on every region locally.

Fusion with external component. The second step simply adds up the output from the first step with the external component output, and applies a hyperbolic tangent function to the sum to transform the output values to be in the range $[-1, 1]$. The function for this step is as follows:

$$\widehat{\boldsymbol{X}}_t = tanh(\boldsymbol{X}_F + \boldsymbol{X}_{Ext}) \tag{8}$$

The model is then trained to minimize the mean squared error (MSE) between the predicted flow matrix and the true flow matrix. The MSE loss function is defined as:

$$\mathcal{L}(\theta) = \sum_t^T \left\| \boldsymbol{X}_t - \widehat{\boldsymbol{X}}_t \right\|_2^2 \tag{9}$$

where θ are all the learnable parameters in the model. Note that since not all grid cells contain crowd flows due to non-traversable regions, we modify the loss function by adding a mask to only calculate loss for specific regions where crowd flows exist.

IV. EXPERIMENTS

A. Settings

Baselines. Apart from ST-ResNet as our main baseline, we also compare the results with two other simple baselines – historical average and persistence model. The historical average model predicts the next inflow and outflow values by using the average of the past observations in the same grid cell, and corresponding time interval in the week. The persistence model simply takes the most recent observation of the crowd flows as the next time interval prediction.

Preprocessing. We use the Min-Max normalization to scale the crowd flows values and the extracted tweets features values into the range of $[-1, 1]$, and $[0, 1]$for the wind speeds and temperature.

TABLE III. COMPARISON OF THE RESULTS AMONGST THE BASELINES AND EXTENDED MODELS [a]

Model	Dataset				Average
	VLD1	VLD2	VLD3	MPS	
ST-ResNet (*Main baseline*)	**3.1278**	**3.4302**	**3.4586**	**2.2520**	**3.0672**
Historical average	5.2428	5.6838	5.0124	2.3585	4.5744
Persistence model	4.3864	4.2329	4.9451	4.9921	4.6391
ST-ResNet + Tweet Counts	**3.1073**	3.2965	3.2345	2.2369	**2.9688**
ST-ResNet + Tweet Tenses	3.1113	3.4238	3.2664	2.2581†	3.0149
ST-ResNet + Tweet Counts + Tenses	3.1459	3.3231	**3.2294**	2.2271	2.9814
ST-ResNet + Tweet Sentiment	3.1300	3.4255†	3.4609†	2.2441	3.0651
ST-ResNet + Tweet Counts + Sentiment	3.1984	3.2578	3.2498	2.3321†	3.0095
ST-ResNet + Tweet Counts + Tenses + Sentiment	3.1578	**3.2455**	3.3409	**2.2072**	2.9879

[a] *Underlined – did not beat baseline; Bold – best score for dataset; † – statistical insignificant.*

Hyperparameters. We use Keras with Tensorflow as the backend to implement our models. The training is done via back-propagation (Adam), with a fixed learning rate of 0.0002. The other hyperparameters in the model are set as follows: $L = 2$, $l_H = 4$, $l_D = 1$, and $l_W = 1$. Batch size used is 32. The convolutions use 64 filters with kernel size of 3×3, while *Conv2* uses 2 filters with the same kernel size. The last four weeks (i.e. 28 days) in each dataset is selected to be the test set, while the rest is the train set. From the train set, 10% validation is used as the development test set, and the remainder 90% is used to train the model in the development phase. Until an early-stop is reached or up to 500 epochs, the development phase ends, and the training continues on with the full train set evaluated with the original test set for 100 epochs.

B. Results and Discussion

The results of our extended models (i.e. with tweets information) with the tweet information time window fixed at $(2^-, 0^+)$, and the baseline models (i.e. without tweets information) are shown in Table III. The error values reported in Table III are the normalized root mean square error (RMSE) in percentages.

For every experiment in Table III, except for the main baseline, we also conduct a paired t-test using the error reduction values from the main baseline and each contender model, denoted by b_i and c_i respectively, to test for statistical difference between each matched prediction point i. Assuming that the difference in the error reduction values are normally distributed, the null hypothesis is $H_0: b_i = c_i$ and alternative hypothesis is $H_1: b_i > c_i$. Those results marked with † are found to be statistically insignificant which support that claim that it is no better than the main baseline.

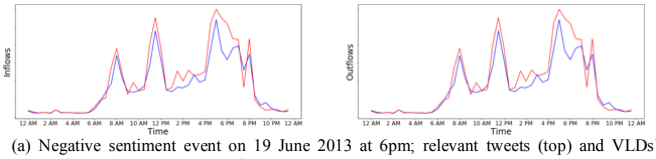
Comparisons between baselines. We observe that all of the models outperform the simple baselines, historical average and persistence model, by a notable margin, signifying the effectiveness of ST-ResNet. With error values just a little above 3%, it indicates that our main baseline is highly competitive. It also implies that crowd flows in Singapore are easy to predict.

Effect of tweet counts. Extended models with tweet counts are able to reduce the errors by 3.28% on average. The error reduction is small yet statistically significant margin, and this observation is consistent for all datasets. This shows that tweets are indeed highly relevant to crowd flow, and crucial to improve prediction performance.

Through our experiments, we also investigate whether information from tweets are indeed unique from historical traffic patterns, so as to assess the value of tweet counts to the crowd flow prediction problem. We addressed the following questions: do people tweet only during peak hours when communing to and from work, and does the tweet counts duplicate historical traffic pattern? If so, this might qualify any derived information from tweets as any patterns found in tweets would only be a replication of those found in traffic flows. Our findings from Fig. 8 show a strong pattern that suggests that the total tweet counts diminish around the sleeping hours (2am-5am), slowly increase along the day (from 5am onwards), and peaks during the night (11pm), and it contains minimal correlation with the peak hour traffic patterns. Thus, we conclude that the tweet counts contain some useful information that does not overlap with the historical crowd flows components.

Effect of tweet tenses and sentiment. We observe that models with tweet tenses and/or sentiment have shown conditional effectiveness across different datasets. Our experiments reveal a combination of datasets which fetched positive outcomes and others that performed below the main baseline (underlined in Table III). Upon investigation, we attribute the cause to some degree of feature misrepresentation from their original intentions, resulting in the treatment of these features as noise. For instance, as shown in Fig. 7a, a high negative sentiment count is detected within the tweets when a public train broke down during an evening rush hour. We originally expected lower crowd flows near the affected region, as explained in Section III.C, however, the crowd flows increased instead. This can be probably explained by a sudden surge of demand in private cars, taxi and busses in the region which may have contributed to the increased crowd flows. We also provide a few examples of tweets that are labelled as either past or future tense in Fig. 7b. Similarly, as per earlier discussed, we deemed these tweets as irrelevant since they are not in present tense. However, these tweets should be considered relevant as they refer to some near-recent or near-future activities which can be useful for prediction.

Time	Tweet text
18:23	walao mrt breakdown at serangoon omg zzzz
18:29	F*cking train stalling at serangoon
18:41	Train broke down at ne line. Now waiting for bus at serangoon . But bus stop overcrowded.
18:50	Human traffic at serangoon is insane
18:57	You gotta be f*cking kidding me. The entire serangoon mrt breakdown



(a) Negative sentiment event on 19 June 2013 at 6pm; relevant tweets (top) and VLDs' crowd flows (bottom). Crowd flows are higher than expected. Note: Blue lines represent daily mean flows, while the red lines represent specific flows on the day.

Tense	Tweet text
Future	Weekend drive in the morning is the bestttt. I can go from hougang to jurong in 10 mins hahaha
	So there's this dude in paya lebar square dancing while holding a cigarette.
	Dinner with Dad @Simpang Bedok . I can say this is the one of the best Mee Goreng I've tasted in my…
Past	Just landed safely back at Changi .
	I told the cab driver I want to go to Bedok . He told me ""sorry I don't take children"
	Just posted a photo @Fort Canning Hill Park
	Finally watched it and it was a great movie. #guardiansofthegalaxy @Golden Village @Yishun

(b) Traffic-relevant tweets yet is in future or past tense.

Figure 7. Case studies to show feature misrepresentations.

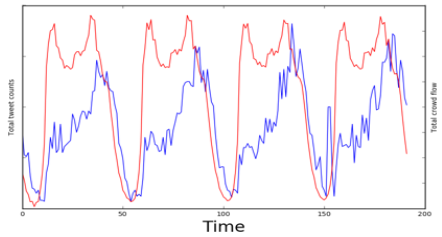


Figure 8. Total tweet counts vs. Total VLDs' crowd flows from 1 March 2013 to 7 March 2013. *Blue line denotes total tweet counts. Red line denotes total crowd flows.*

TABLE IV. SENSITIVITY ANALYSIS OF VARIOUS TWEET INFORMATION TIME WINDOWS USING THE ST-RESNET + TWEET COUNTS MODEL[a]

Lag⁻	Lead⁺			
	0	1	2	3
0	3.3389	3.2240	3.2765	3.2038
1	3.2474	3.2168	3.2045	3.2660
2	3.2128	3.2594	3.2547	3.2712
3	3.2092	**3.1993**	3.2267	3.2608

[a] Values in RMSE

Effect of tweet time interval window. We vary the tweet information time window $(lag^{-}, lead^{+})$ and report the results in Table IV, where it lists the average RMSE amongst all VLD's dataset for the corresponding time interval window. The optimal time interval window is empirically determined to be $(3,1)$. Intuitively, the larger the window size, the better the results of the prediction should be, since more tweets are used. However, this pattern is only observed with higher lag^{-} which ends to achieve better result, as compared to when higher $lead^{+}$ is used. This suggests that one limitation in using tweets in our setting, which is that tweets do not contain information about future events if these events are unforeseeable in the present, as such types of events can only induce people to tweet about it when it happens then.

Reliability of twitter as data source. It is well known that tweets contain large amount of noise as evidenced in examples shown in Fig 6 and Fig. 7. This accentuates the challenge in mining useful information from tweets, and also risk noise being introduced to the predictive model. Furthermore, the tweets that we collected only cover certain grids in the map as we acknowledge the sparsity of their spatial coverage, as shown in Fig. 3. Even though the tweets are collected based on locations that are deemed to be representative of well-known clustered regions, this process also relies on local experts to handpick these locations which could be non-trivial for bigger cities. One premise of this approach also rely on the general usage of twitter adopted by the general population. The accessibility of mobile technologies and to social media might not be as widespread as a highly connected city like Singapore.

V. RELATED WORK

There are several related work that work on the traffic flow prediction. Lv et al. [9] also proposed a deep learning approach by using stacked auto-encoders to model the non-linear spatial and temporal correlations in the traffic data. On the contrary, our work differs from them as we focus on the easier problem of predicting aggregated traffic flows (or crowd flows) across an entire city, rather than at individual road links. Similar to our motivations, He et al. [10] also used twitter to improve traffic prediction linear regression model. In their approach, an optimization process is designed to transform the semantics of the tweets into a salient traffic-indicator matrix. We believe that similar approaches to encode the semantics of tweets into the prediction model can be perform in our context and we leave it for future work.

In the literature, there have been several works that utilize social media as a source of data to study traffic mobility patterns, which are well summarized in [11, 12].

Some works propose to use tweets, instead of traditional physical sensors such as VLDs and road cameras, to monitor real-time traffic. Carvalho et al. [13] argue that the latter methods are expensive, require high maintenance, provide poor spatial coverage and usually inaccessible by the public, all of the disadvantages which the former is able to overcome. Wang et al. [14] further capitalize on geo-tagged tweets to determine congestions along a specific highway in London. However, it is also well known only a small proportion of tweets contain geotags and thus they might not fully represent the level of social activity at their respective regions. Semwal et al. [15] and Shekhar et al. [16] conducted correlations of traffic-related tweets with traffic congestions, instead of traffic flows. Besides, the key challenge as identified in most studies above, is to reliably classify tweets

as being traffic relevant, especially due to the large presence of noise in unstructured natural language text. With the added complexity of spatio-temporal granularities, it only hinders efforts to improve the accuracy of the tweet relevance filtering process. Furthermore, existing studies are often conducted in smaller scales, within specific space and time constraints. Our work is closely related, as we also adopt tweets to represent additional traffic information, but do away with the complex irrelevant tweet filtering step and instead use simple features that are more insensitive to noise.

There has been a line of previous work that use tweets for predictions but applied in other domains such as movies' box office prediction [17], DDoS attacks prediction [18], and presidential election results prediction [19]. While these domains do not face spatio-temporal challenges as pointed out by Zhang et al. [5], the prediction task in their domain is inherently more difficult as patterns are not as predictable due to the presence of more complex interacting human behaviors, and lack of reliable data indicators. Nevertheless, they demonstrate that tweets are good source of information to predict crowd behaviors, and suggest that tweets are indeed informative on telling human behaviors in different aspects of social life ranging from entertainment and politics due to its nature to contain rich real-time information. Thus, our findings are consistent with such existing investigations.

VI. Conclusion

In this paper, we explored the effectiveness of using tweets to crowd flow prediction by extending upon an existing state-of-the-art crowd flow prediction model known as ST-ResNet, adding various linguistic features from real-time tweets. These features include tweet counts, tweet tenses' counts, and tweet sentiments' counts. Through our empirical experiments with two different datasets used to represent traffic flows in Singapore, we found that tweets are indeed useful to improving the prediction accuracy up to 3.28% on average, and tested to be statistically significant. We also shared several ways how tweets are related to crowd flows, with respect to the tweet features extracted and the choice of time interval window chosen.

The development of our framework to use tweets as additional source of information for crowd flow prediction is still very much work in progress. For future work, it could be a good direction to look deeper into the contextual meaning that can be found in the tweets using more advanced natural language processing methods, while also considering efficiency for real-time prediction. Multilingual text processing can also be useful, especially in cities like Singapore, where non-English languages such as Malay, Chinese and Singlish are widely used, resulting in several misclassifications during feature extraction.

Acknowledgment

We thank our anonymous reviewers for their constructive comments which helped to improve the paper. The research is supported by the National Research Foundation, Prime Minister's Office, Singapore, under its CREATE programme, Singapore-MIT Alliance for Research and Technology (SMART) Future Urban Mobility (FM) IRG.

References

[1] X. Niu, Y. Zhu, Q. Cao, X. Zhang, W. Xie, and K. Zheng, "An online-traffic-prediction based route finding mechanism for smart city," *International Journal of Distributed Sensor Networks,* vol. 11, p. 970256, 2015.

[2] A. Abadi, T. Rajabioun, and P. A. Ioannou, "Traffic flow prediction for road transportation networks with limited traffic data," *IEEE transactions on intelligent transportation systems,* vol. 16, pp. 653-662, 2015.

[3] M. Ni, Q. He, and J. Gao, "Using social media to predict traffic flow under special event conditions," in *The 93rd Annual Meeting of Transportation Research Board*, 2014.

[4] Y. Xu, Q.-J. Kong, R. Klette, and Y. Liu, "Accurate and interpretable bayesian mars for traffic flow prediction," *IEEE Transactions on Intelligent Transportation Systems,* vol. 15, pp. 2457-2469, 2014.

[5] J. Zhang, Y. Zheng, and D. Qi, "Deep Spatio-Temporal Residual Networks for Citywide Crowd Flows Prediction," in *AAAI*, 2017, pp. 1655-1661.

[6] S. Petrović, M. Osborne, and V. Lavrenko, "Streaming first story detection with application to twitter," in Human language technologies: The 2010 annual conference of the north american chapter of the association for computational linguistics, 2010, pp. 181-189.

[7] K. Toutanova and C. D. Manning, "Enriching the knowledge sources used in a maximum entropy part-of-speech tagger," in Proceedings of the 2000 Joint SIGDAT conference on Empirical methods in natural language processing and very large corpora: held in conjunction with the 38th Annual Meeting of the Association for Computational Linguistics-Volume 13, 2000, pp. 63-70.

[8] M. Hu and B. Liu, "Mining and summarizing customer reviews," in Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining, 2004, pp. 168-177.

[9] Y. Lv, Y. Duan, W. Kang, Z. Li, F. Wang, *Traffic flow prediction with big data: A deep learning approach*. IEEE Trans. Intelligent Transportation Systems. 2015 Apr 1;16(2):865-73.

[10] J. He, W. Shen, P. Divakaruni, L. Wynter, R. Lawrence, Improving Traffic Prediction with Tweet Semantics. InIJCAI 2013 Aug 3 (pp. 1387-1393).

[11] Y. Lv, Y. Chen, X. Zhang, Y. Duan, N. Li. Social media based transportation research: the state of the work and the networking. *IEEE/CAA Journal of Automatica Sinica*. 2017 Jan;4(1) pp. 19-26.

[12] S. Xu, S. Li, R. Wen. Sensing and detecting traffic events using geosocial media data: A review. Computers, Environment and Urban Systems. 2018 Jun 30.

[13] S. Carvalho, L. Sarmento, and R. J. F. Rossetti, "Real-Time Sensing of Traffic Information in Twitter Messages," 2012.

[14] D. Wang, A. Al-Rubaie, J. Davies, and S. S. i. Clarke, "Real Time Road Traffic Monitoring Alert based on Incremental Learning from Tweets," presented at the Evolving and Autonomous Learning Systems (EALS), 2014 IEEE Symposium on, 2014.

[15] D. Semwal, S. Patil, S. Galhotra, A. Arora, N. Unny, *STAR: Real-time Spatio-Temporal Analysis and Prediction of Traffic Insights using Social Media.* InProceedings of the 2nd IKDD Conference on Data Sciences 2015 Mar 20 (p. 7). ACM.

[16] H. Shekhar, S. Setty, U. Mudenagudi, Vehicular traffic analysis from social media data. In *Advances in Computing, Communications and Informatics* (ICACCI), 2016 International Conference on 2016 Sep 21 (pp. 1628-1634). IEEE.

[17] S. Asur and B. A. Huberman, "Predicting the future with social media," in Proceedings of the 2010 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology-Volume 01, 2010, pp. 492-499.

[18] Z. Wang and Y. Zhang, "DDoS event forecasting using Twitter data," in *Proceedings of the 26th International Joint Conference on Artificial Intelligence*, 2017, pp. 4151-4157.

[19] A. Tumasjan, T. O. Sprenger, P. G. Sandner, and I. M. Welpe, "Predicting elections with twitter: What 140 characters reveal about political sentiment," *Icwsm,* vol. 10, pp. 178-185, 2010.