Crowd Forecasting at Venues with Microblog Posts Referring to Future Events

Ryotaro Tsukada*, Haosen Zhan*, Shonosuke Ishiwatari*§, Masashi Toyoda*,
Kazutoshi Umemoto*, Haichuan Shang*†, and Koji Zettsu†
* The University of Tokyo, Tokyo, Japan
† National Institute of Information and Communications Technology, Tokyo, Japan
Email: {tsukada, hos, ishiwatari, toyoda, umemoto, shang}@tkl.iis.u-tokyo.ac.jp, zettsu@nict.go.jp

***Abstract*—Large events with many attendees cause congestion in the traffic network around the venue. To avoid accidents or delays due to this kind of unexpected congestion, it is important to predict the level of congestion in advance of the event. This study aimed to forecast congestion triggered by large events. However, historical congestion information alone is insufficient to forecast congestion at large venues when non-recurrent events are held there. To address this problem, we utilize microblog posts that refer to future events as an indicator of event attendance. We propose a regression model that is trained with microblog posts and historical congestion information to accurately forecast congestion at large venues. Experiments on next 24-hour congestion forecasting using real-world traffic and Twitter data demonstrate that our model reduces the prediction errors over those of the baseline models (autoregressive and long short term memory) by 20% – 50%.**

***Index Terms*—human mobility, microblogs**

I. Introduction

Large events such as baseball games or concerts attract huge crowds of people, causing congestion around their venues. Such congestion has various negative impacts not only on the event attendees themselves, but also on passers-by. For example, if a train is more crowded than usual when it is packed with people returning from a concert, passengers may feel physically and mentally stressed, and the increased time spent aboard the train results in economic losses. If tourists are not informed in advance about congestion at sightseeing spots, they may be dissatisfied with the trip if their plans are disrupted by the congestion. In addition, congestion can even cause fatal accidents. On 2014 New Year's Eve celebration event in Shanghai, the crowd became uncontrollable and 35 people died in a stampede [1].

Congestion prediction plays a key role in solving these problems. For instance, if the time, place, and degree of future congestion can be predicted, train passengers can take a different line where trains are less crowded, and tourists can make plans so that they can avoid congestion. Against this background, previous studies [2], [3] have proposed frameworks that utilize data collected from GPS-equipped devices and predict congestion on the city scale. However, the predictions of these methods are limited to a few hours in advance because the longer the forecast horizon is, the greater the effect of external factors (e.g., large events) becomes [4]. Thus, the forecasting should be over a long enough period to give people a time to take measures against congestion.

In this study, we tackle the problem of longer-term forecasting of congestion. However, as we show later (Section III), it is difficult to forecast congestion in the vicinity of venues where non-recurrent events are held. To address this problem, we focus on the fact that some microblog posts refer to future events. Such posts are valuable for automatically extracting the time, place, and type of the target events from the microblog posts. For example, if there is a post that says "Protest in front of the National Diet Building on April 14!", we can infer from it that a mass demonstration will take place in front of the National Diet Building on April 14. Based on this idea, we propose to predict congestion in the future from both historical congestion information and microblog posts referring to future events (Section IV). We demostrated the superiority of our model relative to baseline models in experiments on real-world traffic and Twitter data (Section V).

II. Related Work

A. Congestion Prediction Using Location Data

Location data collected from GPS-equipped mobile phones or cars are widely used as means of congestion prediction at venues. Fan et al. [2] pointed out that the previous studies on congestion prediction treated congestion triggered by large events as outliers. To tackle this problem, they proposed an online version of the Markov chain model trained with GPS-based short-term human movement data to forecast city-scale human movements during large events.

A more direct approach to forecasting future human movements is to use query logs from transit or map apps. Konishi et al. [5] utilized the route-search query logs from a transit app and forecasted congestion in railway networks due to large events such as fireworks. Liao et al. [3] proposed a deep learning method based on the fact that the number of the search queries for a venue on map apps increases just before the event.

The existing methods directly associate these location data with spatio-temporal points in a city. In contrast, our model can combine these spatio-temporal data with the microblog posts that are only weakly associated with venues, which does not depend on specific services.

§Currently, he works for Mantra Inc.

B. Event Extraction from Microblog Posts

Microblog services such as Twitter have been widely used as social sensors for capturing information on real-world events.

Yamada et al. [6] extracted the local event information from Twitter posts to help tourists make trip plans. They normalized venue names to collect posts containing event information and extracted the event name and duration from those posts. Jatowt et al. [7] proposed a visual analytics framework for future and past events based on time-referring expressions in microblog posts.

Inspired by these studies, we utilize microblog posts that contain both normalized venue names and time-referring expressions as an indicator of future attendance.

C. Congestion Estimation and Prediction Using Microblog Posts

Some studies attempt to utilize microblog posts to estimate or predict congestion in the real world.

Onishi and Nakashima [8] analyzed the mutual interaction between congestion in the real world and the number of microblog posts in the virtual world. They tried to explain the mutual interaction between the congestion and the number of microblog posts through the parameters of the model trained with real-world data. However, their model is limited to estimating the current congestion by using the number of microblog posts and cannot be applied to future congestion prediction, which we study in this paper.

He et al. [9] pointed out that there are posts that mention traffic information and proposed a method to predict the future traffic volume for longer periods. Their method utilizes posts that are created within the target area of the prediction. However, it is no longer available, as Twitter removed the function of geotagging in 2019.[1]

To address this issue, we extend these approaches and propose a general method that does not depend on the geotagging function.

III. Preliminary Experiment

As we describe in Section II, using historical congestion information is a typical approach to congestion prediction. In this section, we show that forecasts based on this information alone are far from satisfactory, especially at venues.

A. Dataset

We used "Konzatsu-Tokei ®" Data[2]. It consists of estimated numbers of people in square (250 m $\times$ 250 m) grids of Japan that were aggregated every hour from Sep. 2015 until Nov. 2018. We used the last three months of data for testing and the remaining for training. We focused on the 1,500 most crowded grids in Tokyo as of Aug. 2018. Those grids contain large venues such as baseball stadiums and concert halls, and transportation hubs such as main terminals and arterial roads.

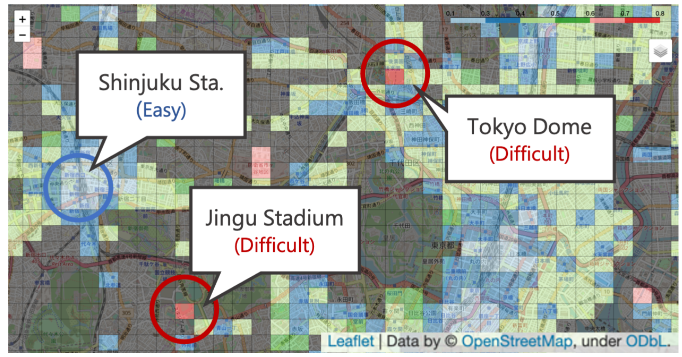


Fig. 1. Prediction errors for the 1,500 most crowded grids in Tokyo. Grids with higher WAPEs are shown in red and those with lower WAPEs are in blue.
"Konzatsu-Tokei ®" ©ZENRIN DataCom CO., LTD.

B. Prediction Method

We trained an autoregressive (AR) model for each grid. It makes a prediction $\hat{X}_t$ for a time step t by linear regression using actual values over the last week $\{X_{t-24\times7},\ X_{t-24\times7+1},\ \cdots,\ X_{t-1}\}$. In the rest of this paper, the time interval is set to 60 minutes.

C. Evaluation Metric

To make the results of different grids comparable, we use the weighted absolute percentage error (WAPE):

$$\text{WAPE} = \frac{1}{N}\sum_{t}^{N}\left|\frac{\hat{X}_t - X_t}{\tilde{X}}\right|, \tag{1}$$

where X_t and $\hat{X}_t$ are respectively the actual and predicted values at a time step t, N is the total number of time steps evaluated, and $\tilde{X}$ is the average of the actual values.

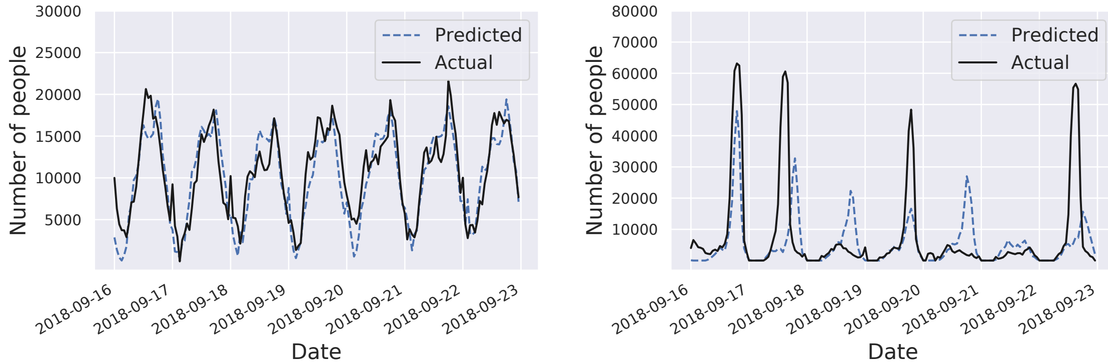
D. Prediction Results and Discussion

The prediction results are shown in Fig. 1, where grids with higher WAPEs (i.e., difficult to predict) are shown in red and those with lower WAPEs (i.e., easy to predict) are in blue. The grids with lower WAPEs are distributed around main terminals such as Tokyo Station and Shinjuku Station. On the other hand, the grids with higher WAPEs are distributed around large venues such as Tokyo Dome and Jingu Baseball Stadium.

To further investigate the prediction error variance, we focused on the two grids that contain Shinjuku Station and Tokyo Dome. Fig. 2 shows the actual and predicted time series of each of the two grids for the week from Sep. 16, 2018. A recurrent pattern during commuting hours is observed at Shinjuku Station. The AR model captured this pattern, resulting in high performance. However, at Tokyo Dome,

[1] https://twitter.com/TwitterSupport/status/1141039841993355264

[2] "Konzatsu-Tokei ®" Data consists of people flow data collected by individual location data sent from mobile phones, given the users' consent, through applications* provided by NTT DOCOMO, INC. The data are processed collectively and statistically in order to conceal private information. The original location data are GPS data (latitude and longitude) sent in about every a minimum period of 5 minutes and do not include enough information to identify individuals. * Applications such as "docomo map navi" service (map navi / local guide).



(a) Shinjuku Station (b) Tokyo Dome

Fig. 2. Predicted and actual number of people within two grids in Tokyo for a week from Sep. 16, 2018. "Konzatsu-Tokei ®" ©ZENRIN DataCom CO., LTD.

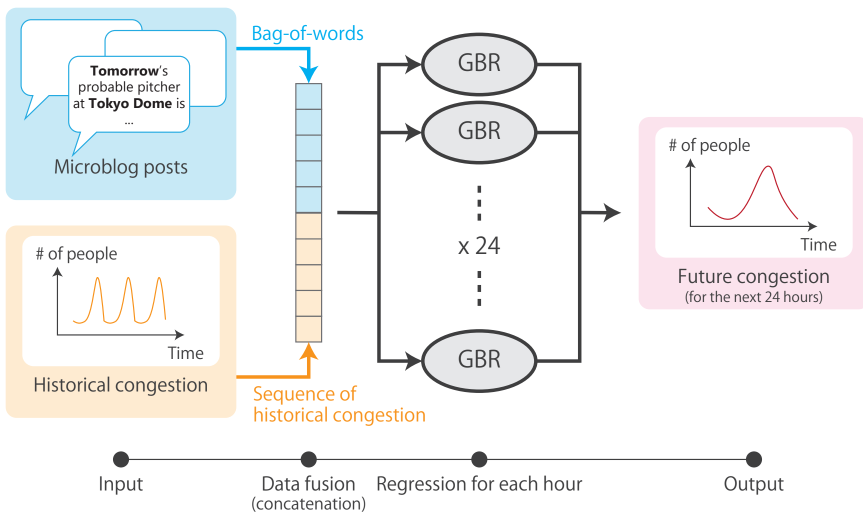


Fig. 3. Overview of our model.

congestion irregularly occurs due to the events held there. The AR model failed to adapt to these surges, resulting in poor performance. The causes of the surges were baseball games held at Tokyo Dome on Sep. 16, 17, and 19.

IV. Proposed Method

As we showed in the previous section, historical congestion information alone is insufficient to forecast congestion in the vicinity of large venues. This is because the model is not aware of the days when events with many attendees (e.g., baseball games and concerts) are held.

A simple solution to this problem is to use the event schedules published by venue managers or event organizers. However, the number of venues for which official schedules are published is limited. To make predictions for various types of venues and events, it is desirable to be able to automatically collect information about future events without relying on such schedules.

In this study, we focus on the fact that information about future events is posted on microblogs. To be specific, we propose a method that utilizes heterogeneous data consisting of microblog posts about events and historical congestion information on venues. Fig. 3 shows an overview of our model. In what follows, we explain each component of our model.

A. Microblog Posts as an Event Indicator

As an indicator of future congestion at large venues, we utilize microblog posts that contain time-referring expressions and a venue name. We consider that a post which is useful for predicting congestion on a future date d at a venue v should meet all of the following conditions:

- It contains a time-referring expression to d
- It contains the venue name v
- It was created before d

We eliminate duplicates (e.g., auto-generated posts by bots) from these posts. Then, we concatenate the unique posts into a single document and use the bag-of-words representation of it. The bag-of-words representation can handle any number of posts and it is invariant to the permutation of posts. This property is suited to this task because the numbers of posts

TABLE I
DESCRIPTION OF VENUES.

Type	Venue	Capacity
Sports venue	Nissan Stadium	72,327
	Tokyo Dome	55,000
	Ajinomoto Stadium	50,000
	Jingu Baseball Stadium	35,133
	Yokohama Stadium	30,000
	Chichibunomiya Rugby Field	24,871
Concert hall	Pacifico Yokohama	18,000
	Yokohama Arena	17,000
	Nippon Budokan	14,471

are different for each target day and the order of input posts are not related to the crowd size on the target day.

B. Hourly Prediction by Microblog Posts and Historical Congestion Information

To capture the short-term trend in the time series, we simply concatenate the bag-of-words vector with hourly data on the number of people for $24 \times n$ time steps and feed the concatenated vector into a regression model. Here, n is a hyperparameter that determines the number of days for which the historical time series is used.

Our model makes a prediction by regression for each hour, taking as input the concatenated bag-of-words and time-series vector, and outputting the number of people at the corresponding hour. We use the gradient boosting regression (GBR) model [10]. GBR is based on the theory of gradient boosting learning, which is a kind of ensemble method. It is used with decision trees as weak learners. An advantage of GBR is that it can handle mixed-type data. We set the hyperparameters of GBR to the default values of scikit-learn's implementation [11].

V. EXPERIMENT

We examined the effectiveness of our model in experiments with real-world data. Our experiments were designed to answer the following questions:

Q1 Can our model identify event days? (Section V-B)
Q2 How accurately can our model forecast the crowd size in different settings? (Section V-C)
Q3 What are the important features? (Section V-D)

A. Experimental Settings

Dataset. Our dataset consisted of spatio-temporal population data and microblog posting data.

For the spatio-temporal population data, we used "Konzatsu-Tokei ®" Data, which is described in Section III, collected from Dec. 2014 to Nov. 2018. We used the last 12 months for testing (Dec. 2017 – Nov. 2018) and the remaining for training (Dec. 2014 – Nov. 2017). As shown in Table I, we chose nine venues in Tokyo and Kanagawa, Japan, from two different event categories (i.e., sports venues and concert halls). These were the nine largest venues by capacity as of Dec. 2014. For each venue, we chose a grid that covers the venue and

TABLE II
EVENT THRESHOLD (MEDIAN OF DAILY PEAK CROWD SIZES PLUS 5,000) AND NUMBER OF EVENT DAYS IN TEST DATA DETERMINED BY THE THRESHOLD.

Venue	Threshold	# event days
Nissan Stadium	6,161.5	10
Tokyo Dome	19,308.0	174
Ajinomoto Stadium	5,612.0	35
Jingu Baseball Stadium	7,379.5	103
Yokohama Stadium	9,448.5	80
Chichibunomiya Rugby Field	8,839.0	19
Pacifico Yokohama	11,649.5	50
Yokohama Arena	10,060.0	43
Nippon Budokan	8,226.0	67

considered the number of people in the grid as the crowd size around the venue. To capture the weekly patterns of the historical time series, the hyperparameter n (Section IV-B) was set to $n = 7$, unless otherwise mentioned. There are cyclic trends in the time series data. They result from daily or weekly commuting and interfere with the crowd-size prediction. Thus, it is important to remove these trends from the time-series data analysis [12]. We removed them from both the input and target time series data by simply subtracting the historical averages for each hour of the day.

For the microblog posting data, we used posts that were extracted from our Twitter archive. Our archive has been maintained since Mar. 2011 by continuously crawling with the Twitter API. It consists of timelines from about 2.5 million public users. Our crawling started with 30 famous Japanese users, and the set of users has been repeatedly extended by following retweets and mentions in their timelines. To expand the coverage of event-related posts meeting the matching conditions (Section IV-A), we created a dictionary of synonyms that maps variants of venue names to a formal name using Wikipedia's redirect data. After tokenizing the matched posts by MeCab [13] and removing stop words, we used the 5,000 most frequent words to obtain the bag-of-words vector.

Baselines. We compared our method with two baselines: AR (same as in our preliminary experiment in Section III) and long short term memory (LSTM) [14]. For LSTM, we followed the default hyperparameters of PyTorch's implementation [15]. Note that these baselines did not use the microblog posts. They predicted the crowd size for the next 24 hours in an autoregressive manner (taking the previous output as input).

Evaluation Metrics. We evaluated the effectiveness of our model in two scenarios: a coarse-grained one and a fine-grained one.

The coarse-grained evaluation determined whether a given model can identify event days. As mentioned in Section IV, however, the official event schedule is not always available for many venues. Thus, we predefined a threshold for each venue and regarded the date as an event day if the peak crowd size exceeded the threshold and a non-event day otherwise. The threshold was defined as $m + 5000$, where m is the median of the daily peaks of the crowd sizes in the training data. Table II

TABLE III
PERFORMANCE (PRECISION AND RECALL) OF EVENT-DAY DETECTION.

	AR		LSTM		Ours	
Venue	Precision	Recall	Precision	Recall	Precision	Recall
Nissan Stadium	0.00	0.00	0.00	0.00	0.73	0.80
Tokyo Dome	0.67	0.61	0.80	0.38	0.86	0.91
Ajinomoto Stadium	0.00	0.00	1.00	0.03	0.96	0.71
Jingu Baseball Stadium	0.61	0.48	0.73	0.21	0.75	0.86
Yokohama Stadium	0.65	0.51	0.87	0.16	0.90	0.91
Chichibunomiya Rugby Field	0.00	0.00	0.00	0.00	0.82	0.47
Pacifico Yokohama	0.59	0.26	0.54	0.14	0.70	0.60
Yokohama Arena	0.75	0.14	0.00	0.00	0.55	0.40
Nippon Budokan	0.62	0.12	0.67	0.03	0.56	0.37

TABLE IV
PERFORMANCE (PRECISION AND RECALL) OF NON-EVENT-DAY DETECTION.

	AR		LSTM		Ours	
Venue	Precision	Recall	Precision	Recall	Precision	Recall
Nissan Stadium	0.97	1.00	0.97	1.00	0.99	0.99
Tokyo Dome	0.67	0.72	0.62	0.91	0.91	0.87
Ajinomoto Stadium	0.90	1.00	0.91	1.00	0.97	1.00
Jingu Baseball Stadium	0.81	0.88	0.76	0.97	0.94	0.89
Yokohama Stadium	0.87	0.92	0.81	0.99	0.98	0.97
Chichibunomiya Rugby Field	0.95	1.00	0.95	1.00	0.97	0.99
Pacifico Yokohama	0.89	0.97	0.88	0.98	0.94	0.96
Yokohama Arena	0.90	0.99	0.88	1.00	0.92	0.96
Nippon Budokan	0.83	0.98	0.82	1.00	0.87	0.93

“Konzatsu-Tokei ®” ©ZENRIN DataCom CO., LTD.

shows the threshold and the number of event days in the test data for each venue. Based on this criterion, we evaluated the results with precision and recall as a binary classification problem.

The fine-grained evaluation assessed the degree of predicted congestion. For this scenario, we used the mean absolute error (MAE):

$$\mathrm{MAE} = \frac{1}{N}\sum_{t}^{N} |\hat{X}_t - X_t| , \tag{2}$$

where X_t and $\hat{X}_t$ are respectively the actual and predicted value at a time step t, and N is the total number of time steps evaluated.

B. Q1: Event-or-Not Prediction

First, we report the results of the coarse-grained evaluation. Table III shows the performance of event-day detection of each method, where the event days are considered to be positive examples. The recall values of the AR model and the LSTM model were nearly zero for many venues. This implies that they failed to predict the presence of almost all of the events. As well, the precision and recall values of the AR / LSTM model were zero for some venues. These uncommon low values are partly explained by the small numbers of event days at these venues (Table II). Since few events took place at these venues, the input time series of the last week was likely to be a sequence of days when few people gathered and it did not contain any useful clues to forecast the time series on the event days. In contrast, our model consistently achieved much higher performance. By utilizing microblog posts referring to future events as an additional clue, our model successfully distinguished event days from non-event days.

Comparing the results for the different types of venues, it can be seen that our model achieved higher performance on baseball stadiums, namely Tokyo Dome, Jingu Baseball Stadium, and Yokohama Stadium, while it had lower performance on concert halls like Yokohama Arena and Nippon Budokan. A possible reason for this difference is that different types of events are held at each venue. For example, baseball games are predominantly held at baseball stadiums. The microblog posts that refer to the games typically contain characteristic words, such as the team name whose home stadium is the target venue. This led to our method having higher performance at sports venues. At concert halls, however, the posts that refer to concerts contain various proper nouns, such as performer names. Consequently, our model had poorer performance on concert halls. Later, we will present our feature importance analysis, whose results support this explanation.

To confirm that our model could predict the crowd size on non-event days that composed the majority of the test data, we report evaluation results focusing on non-event days. Table IV shows the detection performance of non-event days of each method, where the non-event days are considered to be positive examples. Our model again had high performance for most venues. This shows its practicality.

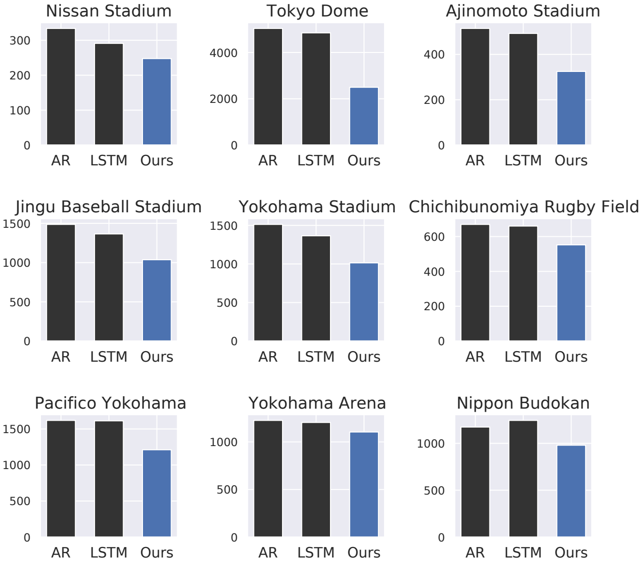


Fig. 4. Prediction error (MAE) on the test data ($n = 7$).
"Konzatsu-Tokei ®" ©ZENRIN DataCom CO., LTD.

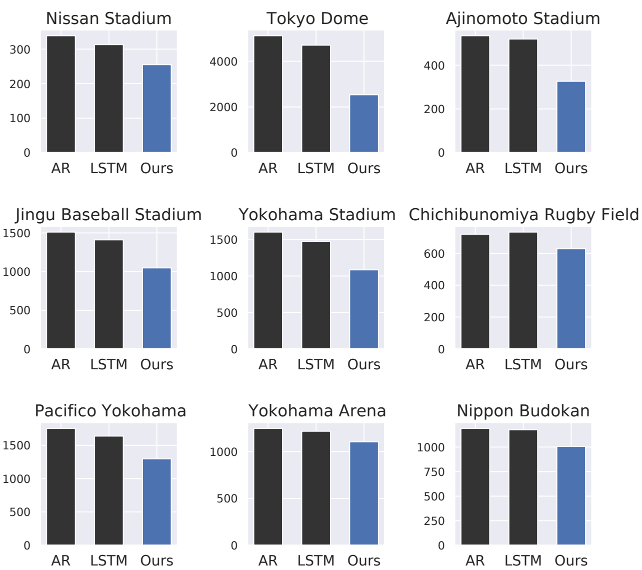


Fig. 5. Prediction error (MAE) on the test data ($n = 1$).
"Konzatsu-Tokei ®" ©ZENRIN DataCom CO., LTD.

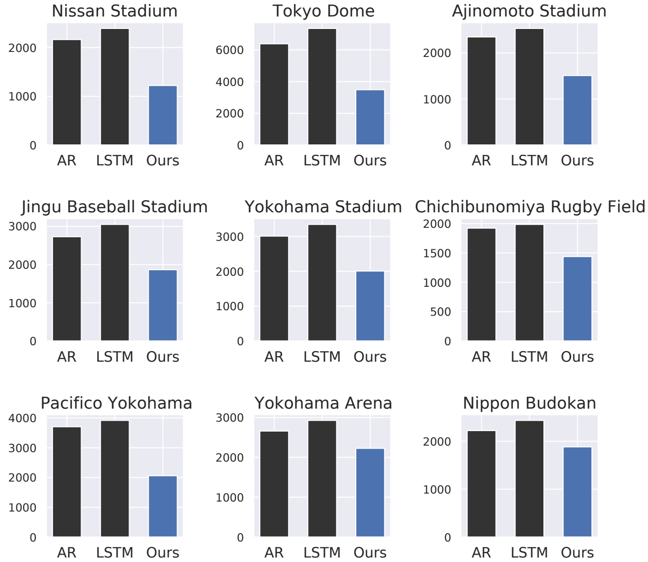


Fig. 6. Prediction error (MAE) during event days on the test data.
"Konzatsu-Tokei ®" ©ZENRIN DataCom CO., LTD.

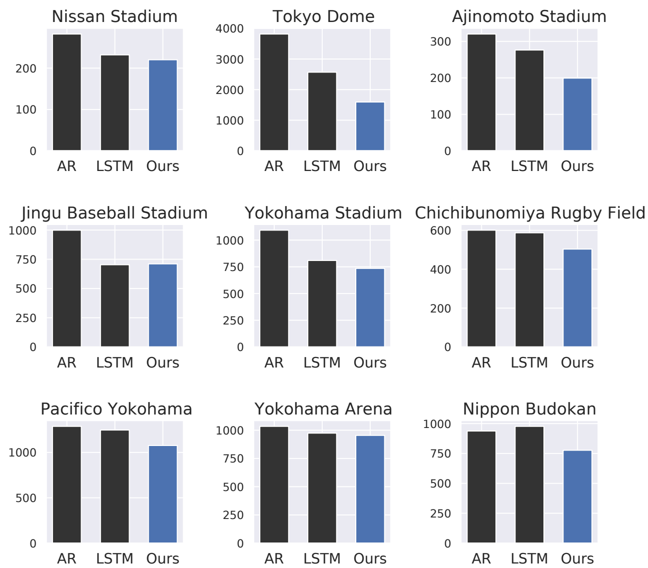


Fig. 7. Prediction error (MAE) during non-event days on the test data.
"Konzatsu-Tokei ®" ©ZENRIN DataCom CO., LTD.

C. Q2: Crowd Size Prediction

Next, we report the fine-grained results. Fig. 4 shows prediction performance for crowd size measured on all of the test data. Our model had lower errors than the baselines did for all venues. The improvement was larger for sports venues than concert halls, which is consistent with the event-or-not classification experiment reported in Section V-B.

To further investigate the ability of our model, we conducted the same prediction task under different settings.

Length of Historical Congestion Information. In this experiment, we changed the value of the hyperparameter n (Section IV-B), i.e., the number of days for which the historical time series was used. Fig. 5 shows the result for $n = 1$. By comparing this result with the result for $n = 7$ (Fig. 4), we can see that the model performed slightly better when the last seven days of the time series vector was fed to it.

Event Day and Non-Event Day. Since most of the test data consisted of non-event days, the evaluation on the whole test data leads to an underestimation of errors. Therefore, we evaluated the prediction performance on event days (Fig. 6) and non-event days (Fig. 7) separately. Although it was more challenging to predict the crowd size during event days, our

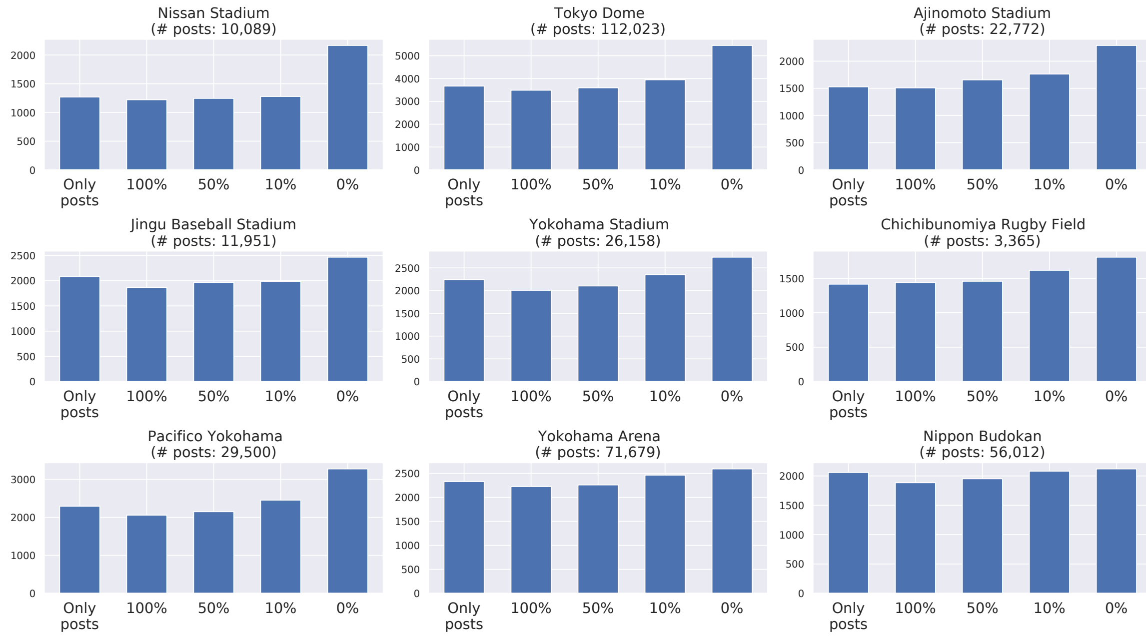


Fig. 8. Prediction error (MAE) during event days on the test data with a reduced number of posts.
"Konzatsu-Tokei ®" ©ZENRIN DataCom CO., LTD.

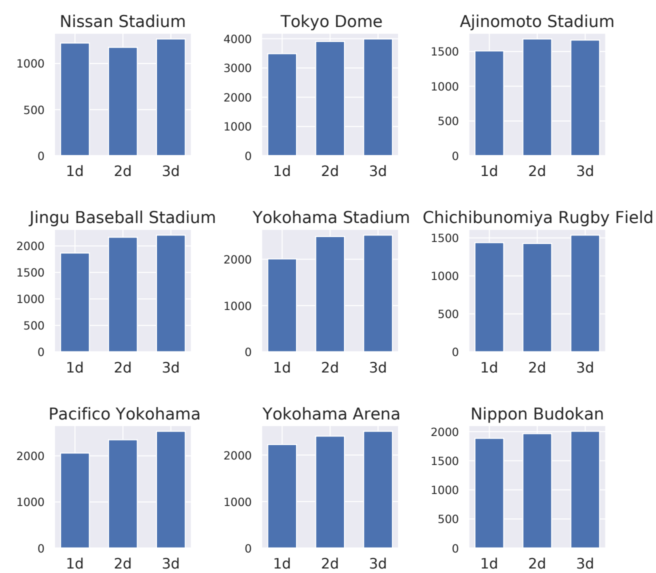


Fig. 9. Prediction error (MAE) during event days on the test data with different forecast horizons.
"Konzatsu-Tokei ®" ©ZENRIN DataCom CO., LTD.

model again outperformed the baseline models for all venues. The difference was smaller on non-event days. Nevertheless, our method performed the best on all except one venue.

Forecast Horizon. To test if our model can predict congestion for longer forecast horizons, we changed the amount of available information prior to prediction. Fig. 9 shows the results for different forecast horizons. For the d-day ahead prediction, only the posts that were posted d or more days before the target day were used. Similarly, only the historical time series d or more days before the target day was used.

For all venues, the longer forecast horizon resulted in a higher error. In particular, the gap between one-day and two-day ahead prediction was the largest. This suggests that the important posts that provide clues to future events tend to be posted on the eve of the events. Another possibility is that the one-day-ahead prediction setting was able to leverage the historical time series to identify the increase in the number of people gathering at the venue on the eve of the event while the earlier prediction settings could not do so.

Number of Posts. To check if the number of posts has an impact on prediction performance, we randomly sampled posts in the training and test data at a constant rate and trained the model with the sampled data. Fig. 8 shows MAE with the reduced number of posts. The experiment where 100% of the posts were used is the same as the experiment shown in Fig. 6. When 0% of the posts were used, only the historical time series vector was fed to the model. Conversely, "Only posts" indicates when only the bag-of-words vector was fed to the model.

It can be seen from the table that the more posts there were, the lower the errors became. The errors were highest when the posts were not used at all, suggesting that the microblog posts contain information that is useful for crowd-size prediction. Another interesting observation is that the "100%" condition performed better than the "Only posts" condition for most venues. The results presented here demonstrate that historical congestion information and event-related posts have complementary roles in the future congestion prediction. Therefore, our idea of using these two types of information is promising.

TABLE V
TEN MOST IMPORTANT FEATURES FOR FIVE PREDICTION TIME SLOTS AT TOKYO DOME
([h @ d DAY(S) AGO] REPRESENTS THE CROWD SIZE OF THE VENUE d DAY(S) AGO AT h). NOTE THAT THESE FEATURES ARE TRANSLATED.

12 PM	2 PM	4 PM	6 PM	8 PM
14	14	14	Tokyo Dome	18
[10 AM @ 1 day ago]	kyojin	18	18	Giants
[3 PM @ 1 day ago]	18	probable pitcher	Giants	Tokyo Dome
[11 AM @ 1 day ago]	Giants	Giants	winning	14
day	probable pitcher	kyojin	Paul McCartney	ticket
hometown	[11 AM @ 1 day ago]	[4 PM @ 1 day ago]	seat	Paul McCartney
[8 PM @ 1 day ago]	[4 PM @ 1 day ago]	Tokyo Dome	tomorrow	Pasela
kyojin	[10 AM @ 1 day ago]	[10 AM @ 1 day ago]	14	[8 PM @ 2 days ago]
[11 AM @ 2 days ago]	[8 PM @ 1 day ago]	tomorrow	exchange	winning
schedule	[2 PM @ 1 day ago]	[8 PM @ 1 day ago]	kinki	Southern All Stars

TABLE VI
TEN MOST IMPORTANT FEATURES FOR FIVE PREDICTION TIME SLOTS AT NIPPON BUDOKAN
([h @ d DAY(S) AGO] REPRESENTS THE CROWD SIZE OF THE VENUE d DAY(S) AGO AT h). NOTE THAT THESE FEATURES ARE TRANSLATED.

12 PM	2 PM	4 PM	6 PM	8 PM
[12 PM @ 1 day ago]	JSDF Marching Festival	Seiko Matsuda	ticket	[8 PM @ 1 day ago]
August 15	[1 PM @ 1 day ago]	[8 PM @ 3 days ago]	those who	ticket
[1 PM @ 1 day ago]	[12 PM @ 1 day ago]	[5 PM @ 1 day ago]	seat	[7 PM @ 6 days ago]
akb48	full of	Wakayama	[8 PM @ 1 day ago]	piece
All Japan Championship	tournament	[6 PM @ 1 day ago]	gate	[8 PM @ 6 days ago]
[10 AM @ 1 day ago]	Taemin	those who	[8 PM @ 6 days ago]	seat
[7 AM @ 3 days ago]	[12 PM @ 3 days ago]	very	alfee	eric
tournament	stage	hall	July 9	book
[11 AM @ 2 days ago]	Seiko Matsuda	[7 AM @ 4 days ago]	concert	[8 AM @ 4 days ago]
[11 AM @ 1 day ago]	All Japan Championship	concert	[7 PM @ 1 day ago]	adjustment

D. Q3: Feature Importance

Finally, we report features useful for predicting future congestion around venues. Our model (GBR) is based on a decision tree, and thus, the importance of each feature can be computed [16]. Thus, for each hour, we analyzed words in the microblog posts or slots of the historical time series that were significant for predicting the crowd size at that hour. Tables V and VI show the ten most important features for five prediction time slots at Tokyo Dome and Nippon Budokan, respectively.

At Tokyo Dome, the word "Giants" (a professional baseball team whose home stadium is Tokyo Dome) ranks high for most hours. This word characterizes the baseball games, which compose the majority of events held there. The words "14" and "18" are important for hours from 12 p.m. to 6 p.m. These words represent the start time (e.g., "The game starts at 18:00...") of baseball games or concerts.

At Nippon Budokan, the sports-related phrases "All Japan Championship" and "tournament" rank high for the early hours. These sports tournaments usually start in the morning. Thus, our model focuses on these words to identify the event type and the start time. For the late hours, concert-related words, "concert", "ticket", and "seat", rank high. These words are used to identify concerts. However, specific performer names (in analogy with the word "Giants" at Tokyo Dome) do not rank high for the late hours, which means our model did not focus on such words for concert halls. This is because there are numerous performers and it is rare that a performer repeatedly holds concerts at the same venue.

This analysis partly explains why the prediction errors at concert halls were higher than those at baseball stadiums.

VI. CONCLUSION

We tackled the problem of forecasting congestion around large venues. First, we showed that historical congestion information alone is insufficient for forecasting, as it cannot capture the surge of people caused by non-recurrent events. To find clues about events, we leveraged microblog posts mentioning the target venues and the target days as additional features for training a congestion prediction model. Experimental results on real-world traffic and Twitter data demonstrated the superiority of our model over the baseline models.

The experiments showed that, our model had lower performance when few posts referred to the future events. Thus, in the future, we will consider using knowledge acquired from the venues mentioned in many posts to improve the prediction of crowd sizes at venues mentioned in few posts. There is still room for improvement of the matching conditions (Section IV-A). Some of the posts that meet these conditions may be noisy. For example, a post that says "The DVD of our Tokyo Dome concert will be released tomorrow!" meets these conditions, but it does not suggest any information about the size of the crowd at Tokyo Dome tomorrow. On the other hand, posts that do not meet these conditions can also be useful. Additionally, we plan to devise a sophisticated method of data fusion to pick up important information from noisy posts and apply our model to more venues of various event types.

Acknowledgment

This work was supported by Commissioned Research (201) on Smart Sustainable Mobility Platform Project of the National Institute of Information and Communications.

References

[1] CNN, "'I failed to protect you' – Details emerge of victims in deadly Shanghai stampede." https://edition.cnn.com/2015/01/02/world/asia/china-shanghai-new-years-stampede/, 2015.

[2] Z. Fan, X. Song, R. Shibasaki, and R. Adachi, "CityMomentum: An online approach for crowd behavior prediction at a citywide level," in *Proceedings of the 2015 ACM International Joint Conference on Pervasive and Ubiquitous Computing*, pp. 559–569, ACM, 2015.

[3] B. Liao, J. Zhang, C. Wu, D. McIlwraith, T. Chen, S. Yang, Y. Guo, and F. Wu, "Deep sequence learning with auxiliary information for traffic prediction," in *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pp. 537–546, ACM, 2018.

[4] J. Zhang, Y. Zheng, and D. Qi, "Deep spatio-temporal residual networks for citywide crowd flows prediction," in *Proceedings of the 31st AAAI Conference on Artificial Intelligence*, pp. 1655–1661, AAAI Press, 2017.

[5] T. Konishi, M. Maruyama, K. Tsubouchi, and M. Shimosaka, "CityProphet: City-scale irregularity prediction using transit app logs," in *Proceedings of the 2016 ACM International Joint Conference on Pervasive and Ubiquitous Computing*, pp. 752–757, ACM, 2016.

[6] W. Yamada, D. Torii, H. Kikuchi, H. Inamura, K. Ochiai, and K. Ohta, "Extracting local event information from micro-blogs for trip planning," in *Proceedings of the 8th International Conference on Mobile Computing and Ubiquitous Networking*, pp. 7–12, IEEE Computer Society, 2015.

[7] A. Jatowt, É. Antoine, Y. Kawai, and T. Akiyama, "Mapping temporal horizons: Analysis of collective future and past related attention in Twitter," in *Proceedings of the 24th International Conference on World Wide Web*, pp. 484–494, IW3C2, 2015.

[8] M. Onishi and S. Nakashima, "Mutual interaction model between the number of people in real space and the number of tweets in virtual space," in *Proceedings of the 23rd International Conference on Pattern Recognition*, pp. 2073–2078, IEEE Computer Society, 2016.

[9] J. He, W. Shen, P. Divakaruni, L. Wynter, and R. Lawrence, "Improving traffic prediction with tweet semantics," in *Proceedings of the 23rd International Joint Conference on Artificial Intelligence*, pp. 1387–1393, AAAI Press, 2013.

[10] J. H. Friedman, "Greedy function approximation: A gradient boosting machine," *Annals of Statistics*, vol. 29, no. 5, pp. 1189–1232, 2001.

[11] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and Édouard Duchesnay, "Scikit-learn: Machine learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.

[12] Z. Wu, N. E. Huang, S. R. Long, and C.-K. Peng, "On the trend, detrending, and variability of nonlinear and nonstationary time series," *Proceedings of the National Academy of Sciences*, vol. 104, no. 38, pp. 14889–14894, 2007.

[13] T. Kudo, K. Yamamoto, and Y. Matsumoto, "Applying conditional random fields to Japanese morphological analysis," in *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing*, pp. 230–237, ACL, 2004.

[14] F. A. Gers, J. Schmidhuber, and F. Cummins, "Learning to forget: Continual prediction with LSTM," in *Proceedings of the 9th International Conference on Artificial Neural Networks*, vol. 2, pp. 850–855, IEE, 1999.

[15] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, A. Desmaison, A. Kopf, E. Yang, Z. DeVito, M. Raison, A. Tejani, S. Chilamkurthy, B. Steiner, L. Fang, J. Bai, and S. Chintala, "PyTorch: An imperative style, high-performance deep learning library," in *Advances in Neural Information Processing Systems*, vol. 32, pp. 8026–8037, Curran Associates, 2019.

[16] L. Breiman, J. Friedman, C. J. Stone, and R. A. Olshen, *Classification and regression trees*. CRC Press, 1984.