Multivariate Linear Regression of Symptoms-related Tweets for Infectious Gastroenteritis Scale Estimation

Ryo Takeuchi, Hayate Iso, Kaoru Ito, Shoko Wakamiya, Eiji Aramaki
{takeuchi.ryo.tj7, iso.hayate.id3, kito, wakamiya, aramaki}@is.naist.jp

Abstract

To date, various Twitter-based event detection systems have been proposed. Most of their targets, however, share common characteristics. They are seasonal or global events such as earthquakes and flu pandemics. In contrast, this study targets unseasonal and local disease events. Our system investigates the frequencies of disease-related words such as "*nausea*," "*chill*," and "*diarrhea*" and estimates the number of patients using regression of these word frequencies. Experiments conducted using Japanese 47 areas from January 2017 to April 2017 revealed that the detection of small and unseasonal event is extremely difficult (r = 0.13). However, we found that the event scale and the detection performance show high correlation in the specified cases (in the phase of patient increasing or decreasing). The results also suggest that when 150 and more patients appear in a high population area, we can expect that our social sensors detect this outbreak. Based on these results, we can infer that social sensors can reliably detect unseasonal and local disease events under certain conditions, just as they can for seasonal or global events.

1 Introduction

Nowadays, the concept of *social sensors* (Sakaki et al., 2010) has been shown to have great potential feasibility for various practical applications. Particularly, disease detection is a core target of social sensor based studies. To date, detection has been demonstrated for influenza (Aramaki et al., 2011; Paul et al., 2014; Lampos et al., 2015; Iso et al., 2016; Wakamiya et al., 2016; Zhang et al., 2017; Lampos et al., 2017), E.Coli (Diaz-Aviles and Stewart, 2012), and H1N1-type flu (Culotta, 2013; Lampos and Cristianini, 2010).

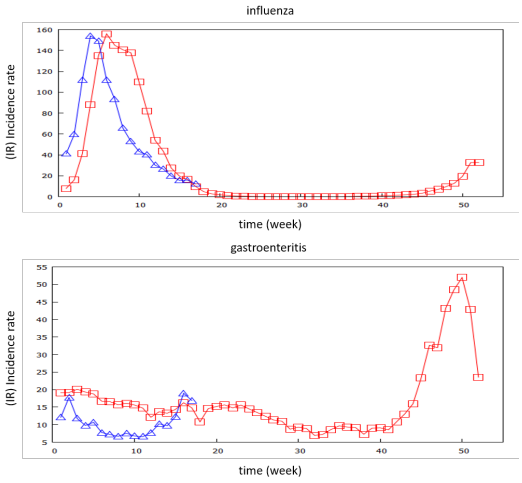


Figure 1: Seasonal global event (a) vs. Unseasonal local event (b). The X-axis shows the timeline (weekly based). The Y-axis shows the incidence rate (IR), corresponding to the patient number per area during the latest two years. Thin line with red square markers show the incidence rates of 2016. The line with blue triangle markers show the incidence rates of 2017. (a) Seasonal global event (*Influenza* in Japan). The *influenza* portrays a single big peak. (b) Unseasonal local event *Gastroenteritis* in the same area as (a). The *Gastroenteritis* shows numerous small peaks. It is difficult to detect Peak periods.

In this field, infectious diseases have drawn much attention mainly for the following two reasons. First, from a practical perspective, infectious disease prevention is a crucially important mission for a nation because infectious diseases, especially influenza, cause many deaths and spread rapidly. Next, from the perspective of informatics, epidemics of these diseases are suitable targets because some epidemics have the following characteristics that make them easy to ascertain from

Proceedings of the International Workshop on Digital Disease Detection using Social Media 2017 (DDDSM-2017), pages 18–25,
Taipei, Taiwan, November 27, 2017.

social media:

1. **Seasonal Event**: some epidemics are seasonal diseases that have basically one big peak during one year (e.g. influenza).

2. **Large Scale Event**: some epidemics infect thousands of people. Accordingly, the scale of information related to the disease in twitter also becomes large (e.g. more than 100,000 flu-related Japanese tweets per day).

Compared with previous works, this study tackles a more challenging task: detection of outbreaks of *infectious gastroenteritis* (In the rest of the paper, we simply call it *gastroenteritis*). Outbreaks of gastroenteritis are often caused by viruses such as *Norovirus* and *Campylobacter*. Symptoms include some combinations of various hard complaints, diarrhea, vomiting, and abdominal pain, fever, and dehydration, which typically last less than two weeks. These features of gastroenteritis make the task more difficult: unlike the flu, the name of a particular disease agent is rarely tweeted. The increased number of patients must be estimated with tweets related to several symptoms.

Although gastroenteritis is sometimes called *stomach flu*, the gastroenteritis characteristics in social respects show quite a contrast to the flu.

1. **Unseasonal Event**: An outbreak of gastroenteritis is not seasonal. It can burst at any time of a year. Moreover, there can be many peaks during a single year.

2. **Local Event**: The scale of the gastroenteritis varies, starting from a smaller event involving a couple of patients to a larger event involving thousands of patients.

A comparison of influenza and gastroenteritis is presented in Figure 1. These characteristics also make it difficult to apply a method intended for influenza detection to gastroenteritis detection.

This study investigates the estimation performance for smaller events rather than previous targets. The results reveal that the event size is a core factor affecting the social sensor performance. From experimentally obtained results, small events (related to about 150 people) were detected with high accuracy (the correlation ratio between social sensor estimation and the actual value is 0.8).

This result contributes to social sensor reliability. This paper is the first reporting the overall relation between social sensor performance and its factors. Although detection of small and unseasonal events is difficult, the sensor can be applied in specified situations.

2 Related Work

Detection of infectious diseases is an important part of national health control. Detection tasks are classifiable into two types: (1) Seasonal infection for diseases such as influenza, and (2) Unseasonal infection such as food poisoning (infectious gastroenteritis) and bio-terror attacks.

For the earliest possible detection, most countries have infection prevention centers: The U.S. has the Centers for Disease Control and Prevention (CDC). The E.U. has its European Influenza Surveillance Scheme (EISS). Japan has its Infection Disease Surveillance Center (IDSC). For each of them, surveillance systems rely on virology and clinical data. For instance, the IDSC gathers influenza patient data from 5,000 clinics and releases summary reports. Such manual systems typically have a 1–2 week reporting lag, which is sometimes pointed out as a major flaw.

In an attempt to provide earlier infectious detection, various new approaches have been proposed to date, such as telephone triage based estimation (Espino et al., 2003) and over the counter drug sales based estimation (Magruder, 2003).

The first web-based infectious disease surveillance was Google Flu Trends (GFT), which uses the Google query log dataset to predict the number of flu patients (Ginsberg et al., 2009). Although GFT has illustrated the effectiveness of web-based surveillance, the Google query log is not a public dataset.

Recent advances of the Web-based infectious disease surveillance depend mainly on open datasets such as those of Twitter (Zhang et al., 2017; Lampos et al., 2017; Iso et al., 2016; Wakamiya et al., 2016; Paul et al., 2014).

Zhang et al. (2017) use several indicator information resources and report the prediction performance obtained for the U.S., Italy, and Spain. Lampos et al. (2017) use word embedding (Mikolov et al., 2013) for enriching the feature selection of the flu model and thereby increase the inference performance. In Japan, the first successful system is that of Aramaki et al. (2011). They

classify whether a user is infected by the flu or not for each tweet that includes a flu-related word. Wakamiya et al. (2016) examines the popularity difference between urban and rural cities for finer-grained infectious disease surveillance.

A state-of-the-art system for use with a Japanese infectious disease model by Iso et al. (2016) uses a time lag for improving nowcasting and for extending the forecasting model. However, they merely examine the prevalence rate throughout Japan; they do not consider the scale of user popularity.

This paper presents an examination of Twitter data through various scales of events, from infection of a few people to an epidemic affecting thousands of people, to detect Twitter-based detection performance.

3 Method

3.1 Extracting Tweets by Patients

To detect outbreaks of gastroenteritis with tweets, we estimate the number of patients.

First, the system collects Japanese tweets via Twitter API[1]. Then we select keyword sets of the following three typical patient complaints: "*nausea*", "*chill*", and "*diarrhea*". This keyword sets are selected in preliminary experiments that use 11 major complaints (Chester et al., 2011). Using the tweet corpus collected in the previous step, we built a classifier that judges whether a given tweet is sent by a patient (positive) or not (negative). This task is a sentence binary classification. We used a SVM-based classifier under the bag-of-words (BOW) representation (Cortes and Vapnik, 1995; Joachims, 1998). Then we split a Japanese sentence into a sequence of words using a Japanese morphological analyzer, MeCab[2] (ver.0.98) with IPADic (ver.2.7.0) (Kudo et al., 2004). The polynomial kernel (d=2) is used as the kernel function. To build the training set, a human annotator assigned either a positive or negative label. For the labeling process, we followed conditions used in our previous study (Aramaki et al., 2011). Table 1 presents samples of tweets with labels.

Finally, we classified tweets into areas for area-based disease surveillance. The area is resolved based on metadata attached to a tweet as follows:

[1] https://dev.twitter.com/overview/api
[2] http://taku910.github.io/mecab/

Table 1: Samples of labeled tweets

Tweet	keyword	P/N
When I got out of the bath I felt chilly. So I am wearing long sleeves and long pants, but now it's hot ('-'). I changed clothes ('-') It might be a cold ...	chill	P
I feel nauseous... I thought it resulted from coccyx pain, but I wonder. if I caught a cold.	nausea	P
I have diarrhea. I am going to a public restroom.	diarrhea	P
I really hate mantis. I hate them more than pigeons. I feel chilly when I think about it.	chill	N
whole-body exposure: 1 Gy nausea, Death year exposure is 10 Gy 1 ms	nausea	N
Meanwhile, Chiba prefecture announced on January 1 that 39 people that 39 people in Ichikawa City, Ibaraki prefecture, had group food poisoning complaining of symptoms such as diarrhea.	diarrhea	N

The tweet on the table are Japanese translations of English.

GPS Information: A tweet includes GPS data if a user allows the use of the location function. However, most users turn off this function for privacy reasons. Currently, the ratio of tweets with GPS information is only 0.46% (=35,635/7,666,201) in our dataset.

Profile Information: Several users include an address in a profile. We regard the user as near the profile address. The ratio of tweets with profile location is 26.2% (=2,010,605/7,666,201). To disambiguate the location names, we use a Geocoding service provided by Google Maps[3].

We removed the tweets without inferred geolocation for the study.

3.2 Linear Regression Analysis of Patient Numbers

Next we investigate the relation between the number of infected people and the number estimated using positive tweets. We use the number of infected people reported from the National Institute of Infectious Diseases (NIID) [4]. The number of infected people in each area is reported per sentinel weekly. To remove the population bias in areas, we calculate the **Incidence Rate (IR)** of people in an area during a week as follows.

$$IR_{repo}(a,t) = \frac{pat_{a,t}}{pop_a} \times 10^k \quad (1)$$

In that equation, $pat_{a,t}$ is the total number of all patients reported in the specified area a within the week index t, pop_a is the area's population, and k is a constant for correcting the value. In the experiment, k is set to 5.

[3] https://developers.google.com/maps/documentation/\geocoding/start
[4] https://www.niid.go.jp/niid/en/

Then, we estimate the linear association between the IR_{repo} and the estimated one, IR_{est}, by application of multivariate linear regression as

$$IR_{est} = b_a^{s1} x_a^{s1} + b_a^{s2} x_a^{s2} + b_a^{s3} x_a^{s3} + b_p, \quad (2)$$

where x_a^s represents the number of positive tweets containing the specific word s. In addition, b_a^s and b_p are variables to be estimated.

4 Experiment

4.1 Setting

For this experiment, we estimated the Incidence Rate from positive tweets with the exploratory variables derived in Section 3.2. The experimental data consist of a training set and a test set. The data are shown in Table 2. For training, we used 1,720,325 tweets for 52 weeks from March 19, 2016 through December 31, 2016 for each area (47 areas in Japan).

Table 2: Dataset statistics

keyword	training	test
`nausea`	560,620 (53%)	594,443 (50%)
`chill`	378,652 (37%)	498,748 (35%)
`diarrhea`	781,053 (74%)	493,693 (69%)
Total	1,720,325 (34%)	1,586,884 (51%)

In brackets represents proportion of positive tweet

For clinical research, we estimated IR_{est} from the test set for 16 weeks from January 1, 2017 through April 19, 2017.

4.2 Results

The overall result is shown in Figure 2. Figure 3a and Figure 3b present details of results from two areas. Figure 3a presents a moderate example in *Nagano* area, where tweet-based estimation highly correlates with the reported values. In contrast, Figure 3b corresponding to *Tokyo* area reveals the weakness of our approach: the estimated value differs greatly from the reported value.

The difference between the two areas reflects the scale of the event. In Figure 3a, the reported values have one large peak (starting from 20 people to 30 people). In contrast, Figure 3b shows a slight increase in the reported values (from 13 to 15 during 15 weeks). From these results, the estimation of small events is difficult, causing numerous false-positive results.

5 Discussion

5.1 Event scale and Estimation Performance

Results reveal that Twitter-based estimation is often adversely affected by small events, yielding poor performance overall. However, in the case of large-scale events, the social sensor usually works well. For that reason, we investigated the relation between the (Sensor) Estimation Performance (EP) and the Event Scale (ES). For this work, we define these indicators as explained below.

Estimation Performance (EP): It is necessary to ascertain how accurately social sensors can estimate an event. We define this indicator as the correlation between IR_{repo} and IR_{est}.

Event Scale (ES): Fundamentally, the higher EP should be obtained when the epidemic scale (ES) is larger. We therefore assessed the correlation between the ES and the EP. We simply define ES based on the difference of IR in a time window as

$$ES = \max_{t \in T} IR(t) - \min_{t \in T} IR(t), \quad (3)$$

where T stands for a time window in the target timeline. $IR(t)$ is a function indicating IR at the week index t.

As for a time window, we divided the test set every 4 weeks (one window) and calculated IR. Correlation between EP and ES is shown in Figure 4. From Figure 4, correlation between EP and ES revealed poor performance (only 0.13). This result suggests that no overall correlation exists between EP and ES.

5.2 Discussion based on Epidemic Pattern

Not only the Event Scale (ES) but also event pattern would affect EP. We considered that event pattern classified by epidemic phase, such as the beginning and the end:

Increasing $\nearrow$ is the phase of the beginning of the epidemic during which the IR increases during the target time window.

Decreasing $\searrow$ is the phase of the end of the epidemic during which the IR decreases during the target time window.

Peak $\wedge$ is the phase of the epidemic peak during which the maximum of the IR is observed.

Between $\vee$ is intermediate of two epidemic peaks.

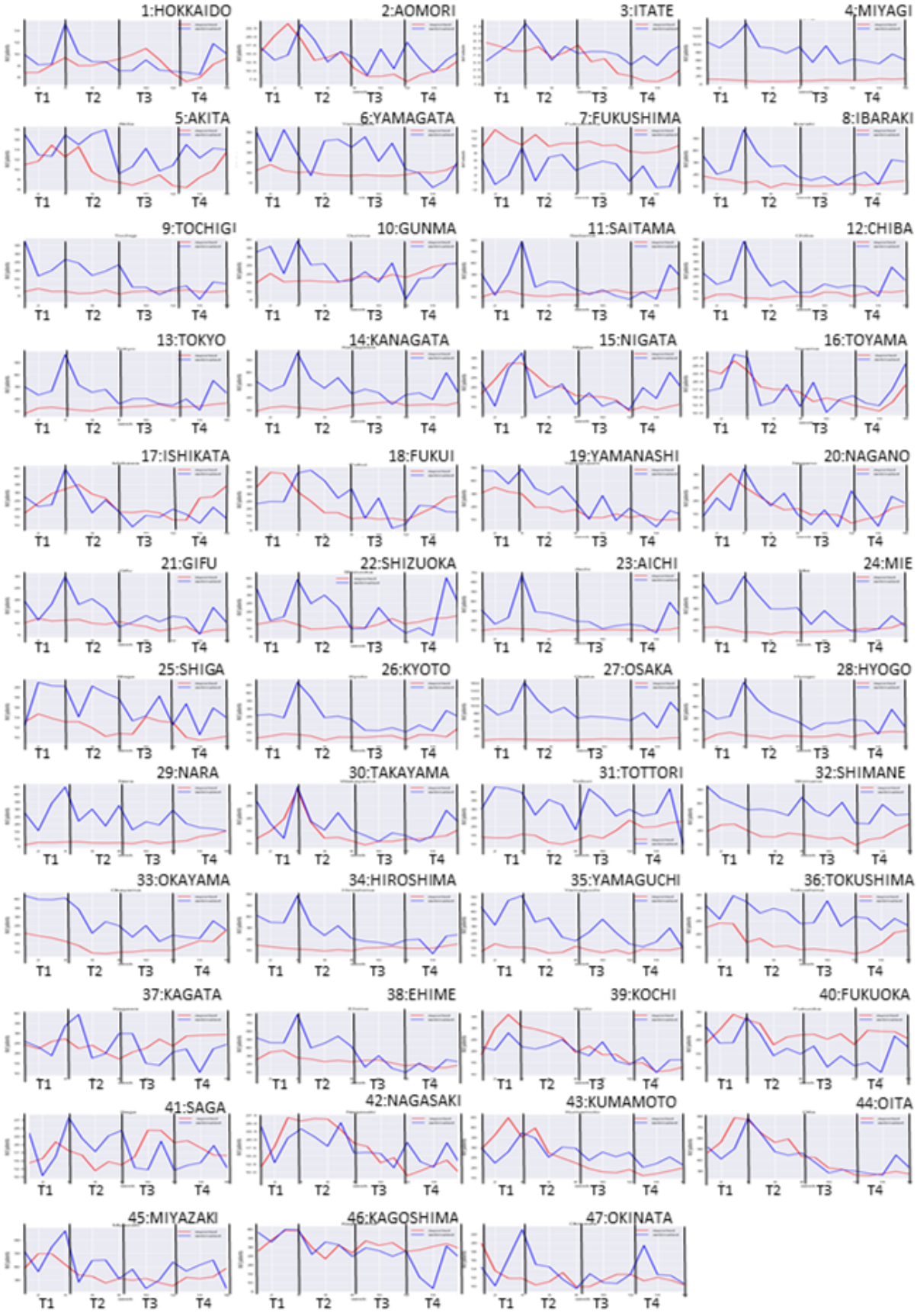
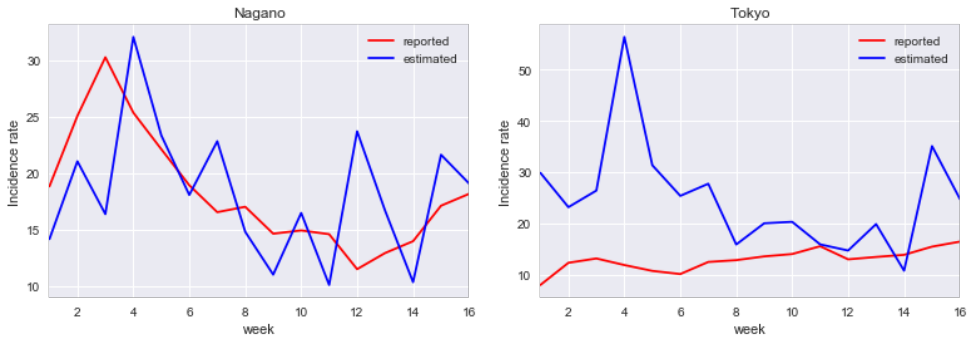


Figure 2: Result for each area. The X-axis indicates the time line (week). The Y-axis indicates the Incidence Rate (IR). *T* indicates the time window (4 weeks).



(a) Example: Nagano (b) Example: Tokyo

Figure 3: Representative results in two areas: (a) Nagano and (b) Tokyo. The red line shows IR_{repo}. The blue one shows IR_{est}.

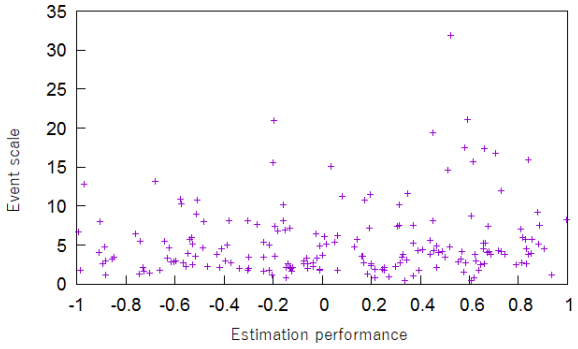


Figure 4: Relation between EP (X-axis) and ES (Y-axis). This revealed poor performance (r = 0.13).

The detailed definition is presented in Table 3. The table presents a window for which $IR_b - IR_i > 0$ and $IR_b - IR_e > 0$ (represented as $IR_b - IR_i > 0$ and $IR_b - IR_e > 0$) is regarded as the *increasing* Pattern.

The results are presented in Table 4, indicating correlation between the EP and ES for each Pattern. As the table shows, the performance showed divergence in each Pattern. For instance, the *decreasing* Pattern showed high correlation ($r = 0.305$). In contrast, the *between* Pattern shows quite poor performance (less than 0 correlation).

5.3 Discussion based on Area Population

The number of tweets is related to the population Therefore, we inferred that the EP is affected by the population of each area. We classified each window by four types based on population. We defined the four types as explained below.

Super High population area (SHP) is area with population of 2.5 million or more

High population area (HP) is area with population of 1.5 million to 2.5 million

Low population area (LP) is area with population of 1 million to 1.5 million

Super Low population area (SLP) is area with population of 1 million or less

Table 5 shows the correlation between the EP and ES in each population area. From Table 5, in high population area (1.5 million to 2.5 million), weak correlation was found ($r = 0.214$). Furthermore, correlation between EP and ES is related to population.

5.4 Combination of Factors

As described above, we introduced three factors that affected the estimation performance (EP): (1) event scale (ES), (2) event pattern, and (3) area population. In this section, we combined the above findings, and investigated the correlation coefficient between the performance and the (1) ES in each factor, (2) event pattern (four types) and (3) area population (four types). The 16 obtained combinations are shown in Table 6.

From Table 6, when the epidemic decreases greatly in areas with low population, the performance tends to be high. Especially, this trend is significant for *decreasing* pattern in low and super low population areas. In contrast, *peak* and *between* pattern show poor correlation.

From practical viewpoints, the *increasing* pattern is important because catching the increase or decrease of patients contributes to prevention. Figure 5 presents the relation between the EP and ES in *increasing* pattern in the high population area, which shows moderate correlation ($r = 0.378$). In

Table 3: Patterns of Epidemic Phase. Each pattern is classified by three phase: IR_b, IR_i and IR_e. The IR_b is the IR at the beginning of the target window. The IR_i is the average IR of all weeks, except for the beginning and the end. In the experiments, the window size is 4 weeks. Consequently, the IR_i is the average IR of the second and third weeks. The IR_e is the IR at the end of the target window.

Pattern	$IR_b - IR_i$	$IR_b - IR_e$	of samples
Increasing ↗	+	+	55
Decreasing ↘	−	−	45
Peak ∧	+	−	56
Between ∨	−	+	32

Table 4: Correlation between EP and ES in each pattern

Pattern	correlation
Increasing ↗	0.098
Decreasing ↘	0.305
Peak ∧	-0.024
Between ∨	0.058

Table 5: Correlation between EP and ES in each area population

Population	correlation
SHP area	0.125
HP area	0.214
LP area	0.185
SLP area	0.059

the figure, moderate performance ($r > 0.8$ in the X-axis) is obtained when the ES is greater than 7 in the Y-axis. It corresponds to a greater than 150 patient increase in a month.

From this result, we can estimate the borderline of the reliable warning that is 150 patient increasing or decreasing in the high population area.

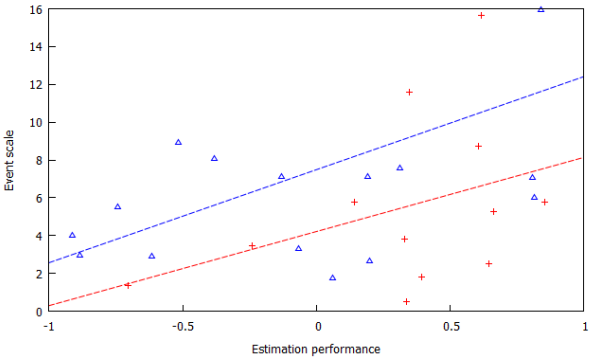


Figure 5: Relation between EP (X-axis) and ES (Y-axis) in the Increasing Pattern in the High Population area (plot with blue triangle) and the Decreasing Pattern in the Low Population area (plot with red cross). This situation for which performance depends on the scale.

Table 6: Correlation between EP and ES in the combination of event pattern and population areas

	SHP area	HP area	LP area	SLP area
Increasing ↗	0.255	0.378	0.01	-0.112
Decreasing ↘	-0.678	0.164	**0.550****	**0.538***
Peak ∧	0.144	0.063	-0.394	0.411
Between ∨	0.494	0.411	-0.409	0.179

Bold font indicates significant correlation (** is $p < 0.05$, * is $p < 0.10$).

5.5 Practical Contribution and Future Direction

To date, social sensors have demonstrated their potential feasibility for various event detections. However, the practical application is rarely launched. One reason is the lack of reliability of social sensors. In other words, we can never fully trust social sensor-based information.

Results of this study demonstrated that the event scale and the estimation performance of social sensor are related. We think this finding is practically important because this characteristic provides important information for the following two use cases:

1. In cases where a really big epidemic occurs, we can believe that the system must detect the clue of the epidemic.

2. In contrast, in cases where the system estimation is normal, we can at least infer that the current situation is not crisis.

From a practical viewpoint, these features that can engage such safety are important. Based on these results, we are developing a surveillance service supported by Infectious Disease Surveillance Center (IDSC). In the near future, we would like to report a case of a system-running experience.

This report describes our attempt at the detection of small and unseasonal disease events.

The method employs the regression of disease-related word frequencies. Results of the experiment, based on Japanese 47 areas from January 2017 to April 2017, suggests that the detection of small events is difficult (r = 0.13). Although the overall performance is poor, the event scale (change in the number of patients) and the detection performance size show correlation (the phase of epidemic in high population area shows a correlation ratio of $r = 0.38$). We think this finding is practically important because it enables realization of a practical system that is useful in the following two use cases. (1) If a truly large epidemic occurs, we can infer that the system must detect it, (2) In other words, the system estimation is low, we can at least infer that the current situation is not so severe. These characteristics are fundamentally important for use in protecting public safety.

In future work, we plan to apply other classification algorithms and compare the performance. Furthermore, we will examine the indicator to represent the ES more effectively.

Acknowledgements

This work was supported by Japan Agency for Medical Research and Development (Grant Number: 16768699) and JST ACT-I.

References

Eiji Aramaki, Sachiko Maskawa, and Mizuki Morita. 2011. Twitter catches the flu: Detecting influenza epidemics using twitter. In Proc. of EMNLP. pp. 1568–1576.

Tammy L Stuart Chester, Marsha Taylor, Jat Sandhu, Sara Forsting, Andrea Ellis, Rob Stirling, and Eleni Galanis. 2011. Use of a web forum and an online questionnaire in the detection and investigation of an outbreak. Online Journal of Public Health Informatics 3(1).

Corinna Cortes and Vladimir Vapnik. 1995. Support-vector networks. Machine Learning 20(3):273–297.

Aron Culotta. 2013. Lightweight methods to estimate influenza rates and alcohol sales volume from twitter messages. Lang. Resour. Eval. 47(1):217–238.

Ernesto Diaz-Aviles and Avaré Stewart. 2012. Tracking twitter for epidemic intelligence: Case study: Ehec/hus outbreak in germany, 2011. In Proc. of WebSci. pp. 82–85.

J. Espino, W. Hogan, and M. Wagner. 2003. Telephone triage: A timely data source for surveillance of influenza-like diseases. In Proc. of AMIA Annual Symposium. pp. 215–219.

Jeremy Ginsberg, Matthew H Mohebbi, Rajan S Patel, Lynnette Brammer, Mark S Smolinski, and Larry Brilliant. 2009. Detecting influenza epidemics using search engine query data. Nature 457(7232):1012–1014.

Hayate Iso, Shoko Wakamiya, and Eiji Aramaki. 2016. Forecasting word model: Twitter-based influenza surveillance and prediction. In Proc. of COLING. pp. 76–86.

Thorsten Joachims. 1998. Text categorization with support vector machines: Learning with many relevant features. In proc. of ECML pp. 137–142.

Taku Kudo, Kaoru Yamamoto, and Yuji Matsumoto. 2004. Applying conditional random fields to japanese morphological analysis. In Proc. of EMNLP. volume 4, pp. 230–237.

Vasileios Lampos and Nello Cristianini. 2010. Tracking the flu pandemic by monitoring the social web. In Proc. of CIP. pp. 411–416.

Vasileios Lampos, Andrew C Miller, Steve Crossan, and Christian Stefansen. 2015. Advances in nowcasting influenza-like illness rates using search query logs. Scientific Reports 5.

Vasileios Lampos, Bin Zou, and Ingemar Johansson Cox. 2017. Enhancing feature selection using word embeddings: The case of flu surveillance. In Proc. of WWW. International World Wide Web Conferences Steering Committee, Republic and Canton of Geneva, Switzerland, pp. 695–704.

S. Magruder. 2003. Evaluation of over-the-counter pharmaceutical sales as a possible early warning indicator of human disease. In Johns Hopkins University APL Technical Digest (24).

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. In Proc. of NIRS. pp. 3111–3119.

Michael J Paul, Mark Dredze, and David Broniatowski. 2014. Twitter improves influenza forecasting. PLoS Currents Outbreaks .

Takeshi Sakaki, Makoto Okazaki, and Yutaka Matsuo. 2010. Earthquake shakes twitter users: Real-time event detection by social sensors. In Proc. of WWW. pp. 851–860.

Shoko Wakamiya, Yukiko Kawai, and Eiji Aramaki. 2016. After the boom no one tweets: microblog-based influenza detection incorporating indirect information. In Proc. of EDB. pp. 17–25.

Qian Zhang, Nicola Perra, Daniela Perrotta, Michele Tizzoni, Daniela Paolotti, and Alessandro Vespignani. 2017. Forecasting seasonal influenza fusing digital indicators and a mechanistic disease model. In Proc. of WWW. pp. 311–319.