Understanding the Diversity of Tweets in the Time of Outbreaks

Nattiya Kanhabua
L3S Research Center
Leibniz Universität Hannover
Hannover, Germany
kanhabua@L3S.de

Wolfgang Nejdl
L3S Research Center
Leibniz Universität Hannover
Hannover, Germany
nejdl@L3S.de

ABSTRACT

A microblogging service like Twitter continues to surge in importance as a means of sharing information in social networks. In the medical domain, several works have shown the potential of detecting public health events (i.e., infectious disease outbreaks) using Twitter messages or *tweets*. Given its real-time nature, Twitter can enhance *early outbreak warning* for public health authorities in order that a rapid response can take place. Most of previous works on detecting outbreaks in Twitter simply analyze tweets matched disease names and/or locations of interests. However, the effectiveness of such method is limited for two main reasons. First, disease names are highly ambiguous, i.e., referring slangs or non health-related contexts. Second, the characteristics of infectious diseases are highly dynamic in time and place, namely, strongly time-dependent and vary greatly among different regions. In this paper, we propose to analyze the *temporal diversity* of tweets during the known periods of real-world outbreaks in order to gain insight into a temporary focus on specific events. More precisely, our objective is to understand whether the temporal diversity of tweets can be used as indicators of outbreak events, and to which extent. We employ an efficient algorithm based on sampling to compute the diversity statistics of tweets at particular time. To this end, we conduct experiments by correlating temporal diversity with the estimated event magnitude of 14 real-world outbreak events manually created as ground truth. Our analysis shows that correlation results are diverse among different outbreaks, which can reflect the characteristics (severity and duration) of outbreaks.

Categories and Subject Descriptors

H.3.3 [**Information Storage and Retrieval**]: Information filtering; J.3 [**Life and Medical Sciences**]: Health; Medical information systems

General Terms

Algorithms, Experimentation, Performance

Keywords

Web Observatory; Twitter; Outbreak Events; Event Detection; Temporal Diversity

WWW 2013 Companion, May 13–17, 2013, Rio de Janeiro, Brazil.
ACM 978-1-4503-2038-2/13/05.

1. INTRODUCTION

Twitter or a microblogging service is regarded as a valuable resource for real-time Web information. Numerous research works have shown the usefulness for real-time Web applications in many domains. For instance, Twitter messages (or *tweets*) can be used for detecting natural disasters [17, 23], analyzing political persuasion [14, 22], and predicting financial time series [16]. In a medical domain, Twitter has been used for tracking influenza outbreaks [1, 5, 10, 15, 20]. The reason for this is that Twitter is capable of transmitting information faster than traditional media channels, such as, news or official outbreak reports [7]. Given its timeliness property, detecting outbreak events in Twitter can provide early warnings for public health officials in order to prevent and/or mitigate the impact of relevant events. Most of previous works on detecting outbreaks in Twitter consider an outbreak event as a temporal anomaly found in time series data that occurs when an infectious disease or its impact is above an expected level, for a particular time and place. For a given outbreak, time series data used for analysis is created by *keyword-matching* [1, 10, 12, 15], namely, computing the aggregated counts of tweets that contain a medical condition (i.e., infectious disease names) from location corresponding to the outbreak event.

However, the effectiveness of such method relying solely on keyword-matching is limited due to two main reasons:

Highly ambiguous and noisy data. Time series data created from tweets are noisy, highly ambiguous and sparse. Noise and ambiguity are caused by spurious events in which an entity is correctly detected, but *its role is not*. Examples of tweets irrelevant to the medical domain are shown in Table 1. Incomplete or sparse time series data implies that instances of an event are missing or under-reported. This may occur due to: 1) the presence of processing errors - an acronyms or abbreviations not recognized as medical conditions, 2) the fact that people who are actually suffering do not tweet, 3) the tweets which contain these mentions have not been collected by the system, i.e., based on the imbalance between the type of tweets collected (e.g., personal versus news tweets), and 4) the minimum required entity types are not present. Sparse time series data refers specifically to low aggregation counts, which impact the anomaly detection algorithm.

Temporal and spatial dynamics of diseases. The characteristics of infectious diseases are highly dynamic in time and space, and their behaviors vary greatly among different regions and the time periods of the year. Some dis-

eases can be rare or aperiodic, while others occur more periodically. In addition, various diseases have different transmission rates and levels of prevalence within a region. For example, cholera infections vary greatly in frequency, severity, and duration. On the one hand, in some regions historically, only sporadic outbreaks occur in areas, i.e., parts of South America and Africa. On the other hand, even in areas where cholera infections are endemic (the South Asian countries of Bangladesh and India) the epidemic levels change dramatically from one year to the next.

Given the imperfect time series data, several questions arise when detecting outbreaks under this condition. *Can we actually trust outbreak events that have been detected for early warning?* Or, *to which extent the detected outbreak events can be reliable?* While previous works have already addressed the problem of noisy and incomplete data in Twitter, most of existing works focus on building classifiers for detecting self-reported illness [4, 18], syndromes [3] and ailments [15]. Unfortunately, such methods supervised learning techniques are time-consuming and expensive, i.e., require manually labeling a set of relevant tweets used for training a classification model. Moreover, the detection of *novel* and *aperiodic* outbreak events requires adaptive approaches which take into account feature change over time.

In addition to noise and incompleteness, none of previous works have explicitly addressed the temporal and spatial dynamics of diseases for detecting outbreaks in Twitter. More precisely, previous works only focus on a limited number of infectious diseases, e.g., influenza or dengue fever, and only countries with a high density of Twitter users were subjects of the study (e.g., United Sates, United Kingdom or Brazil).

In this work, we seek a new feature that surges signals for outbreak events in order to go beyond exploiting the aggregated counts of tweets matched specific disease names and locations. Thus, we propose to analyze the diversity metrics of tweets over time, so-called *temporal diversity*. The diversity statistics are computed as the sum of similarities of all object pairs in a collection of tweets. We adapt a method for measuring diversity statistics proposed by Deng et al. in [6]. The diversity statistics can capture a broad spectrum of topics, communities and knowledge that are evolving over time. In particular, analyzing temporal diversity can shed light on two aspects. First, an increase of content diversity over time indicates that a community is broadening its area of interest. Second, negative peaks in diversity can additionally reveal a temporary focus on specific events. Understanding the dynamics of real-time Web contents generated in social networks during real-world events can enhance *an application to Web Observatories*. To the best of our knowledge, we are the first to study temporal diversity in Twitter. Moreover, our analysis covers general diseases that are not only seasonal, but also sporadic diseases occurring in low tweet-density areas like Kenya or Bangladesh. The contributions of this work are listed as follows:

- We conduct the first study of analyzing temporal diversity in Twitter.
- We propose a method to extract topic dynamics for outbreak events, which will be used as an estimate of real-world outbreak statistics.
- We perform a comparative study by correlating the temporal diversity with the estimated statistics of 14 real-world outbreaks manually created.

Table 1: Examples of tweets and their respective categories *irrelevant* to the medical domain.

Literature	*A two hour train journey, Love In the Time of Cholera ...*
Music	*Dengue Fever's "Uku," Mixed by Paul Dreux Smith Universal Audio...*
Marketing	*Exclusive distributor of high quality #HIV/AIDS Blood & Urine and #Hepatitis #Self -testers.*
General	*Identification of genotype 4 Hepatitis E virus binding proteins on swine liver cells: Hepatitis E virus...*
Negative	*i dont have sniffles and no real coughing..well its coughing but not like an influenza cough.*
Joke	*Thought I had Bieber Fever. Ends up I just had a combo of the mumps, mono, measles & the hershey squ...*

The rest of the paper is structured as follows. In Section 2, we explain how to create outbreak time series data as well as how to estimate real-world outbreak statistics used for the analysis. In Section 3, we present a method to calculate temporal diversity metrics. In Section 4, we explain our experimental settings by outlining our dataset and real-world outbreak ground truth, as well as discuss results. In Section 5, we present related work. Finally, in Section 6, we conclude the paper and outline the future work.

2. OUTBREAK TIME SERIES DATA

Previous work [5] usually employs case/victim statistics, e.g., Influenza-like-Illness (ILI) rates, which are publicly available for a common disease like influenza. However, we aim at studying several real-world outbreak events for *general diseases* that are not only seasonal, but also sporadic diseases that occur in low tweet density areas. Thus, we propose to employ the topic dynamics of outbreaks as *an estimate of event magnitude*. The process is composed of two main steps: 1) create time series data for a given outbreak, and 2) identify topic dynamics using an unsupervised clustering technique. In this section, we first explain how to create time series data for an outbreak event. Then, we describe a method for identifying topic dynamics for the event.

2.1 Matching Tweets

We create time series data by matching tweets with keywords relevant for a given outbreak event, namely, disease name and location. This simplified assumption is based on a minimum requirement by domain experts to assess public health events. More precisely, a tweet will be matched with the outbreak when it contains the co-occurring of two entity types: *medical condition* and *geographic expression*, which are corresponding to the disease name and location of the outbreak event, respectively.

The location information of each tweet can be identified by choosing from the following sources ordered by the degree of relevance: 1) text-contained location (geographic expressions), 2) geo-location information (latitude and longitude), and 3) user profile location. Note that, there could be more than one geographic expression mentioned in a tweet, and

Table 2: List of negative keywords associated to some infectious diseases (provided by *MedISys*).

anthrax	concert, fly, song, castle, film, novel, book, music, band
botulism	plastic+surgery, cosmetic, dermatologic, soccer, film, book, football, show, concert, band, movie, music
cholera	love, album, band, music, concert, music, song, book
ebola	film, book, football, show, concert, band, movie, music
mumps	multifrontal+massively, programming+system,processing+system, concert, film, novel, book, music, band, movie,
norovirus	film, book, football, show, concert, band, movie, music

we take into account all geographic expressions found in the contents. We note that a location implicitly inferred from a tweet language (provided as an attribute by the Twitter API) is not suitable since this information is unreliable.

For each location determined, we will normalize it into three granularities of geographic concept hierarchy: country, continent and latitude. The intuition of considering different geographic granularities is that public attentions might depend on their geographic distance from an outbreak event. For example, people tend to talk or share their opinions about an ongoing outbreak in a neighborhood country because they are concerning that the outbreak can spread into their country. Consequently, we consider not only a *country-level* location, but also *continent-level* and *latitude-level* locations. Thus, locations with fine-grained granularities (e.g., cities, provinces and states) will be normalized by mapping them into coarser granularities (e.g., country and latitude levels) using a geo-tagger tool.

2.2 Identifying Topic Dynamics

After obtaining time series data, we will identify topics mentioned in the data stream using a clustering technique. First of all, we have to apply filtering in order to remove tweets *irrelevant* to the medical domain. Our filtering step exploits a list of negative keywords associated to diseases from two resources freely-available on the Web: 1) MedISys[1] providing a list of negative keywords created by medical experts, and 2) Urban Dictionary[2] - a Web-based dictionary of slang, ethnic culture words or phrases. Examples of negative keywords from are MedISys displayed in Table 2.

In the next step, we apply an unsupervised clustering method over the data stream of outbreak-related tweets. The intuition of clustering is to group tweets into topics in order to track the topic evolution of outbreak events over time, denoted *outbreak-related topic dynamics*. Our steps for identifying outbreak-related topics for a particular outbreak are described as follows:

[1] http://medusa.jrc.it/medisys/homeedition/en/home.html

[2] http://www.urbandictionary.com/

Table 3: Topics about the 2011 botulism outbreak in France from 7-Sept-11 (*top*) and 8-Sept-11 (*bottom*).

Cluster Name	Representative Tweet
Tapenade	*tapenade linked eight botulism cases france*
Recall	*botulism france the official recall with photo*
Serious	*france people seriously ill after botulism life support infected tapenade products produced cavaillon provence*
Food Agency	*warning botulism outbreak france the food standards agency warning people not consume certain brand*
Botulism France	*botulism toll climbs france company not inadequate processing*
Life Support	*people sick and life support with botulism france the culprit olive tapenade from unlicensed vendor*

- Take the time series data of tweets relevant to a particular outbreak event as input.
- For each time period, perform clustering by topic using an unsupervised algorithm.
- Filter result topics by determining the quality and number of tweets in a cluster with respect to specific thresholds.
- Output the aggregated counts of topics in each time period as outbreak-related topic time series.

Examples of topics mentioned about the 2011 botulism outbreak in France are shown in Table 3 chosen from 7 September 2011 and 8 September 2011. Each topic is represented by a cluster name and a representative tweet. To this end, we will employ the identified topics of outbreaks as *an estimate of event magnitude* and correlate with temporal diversity explained in the next section.

3. TEMPORAL DIVERSITY

Our motivation in leveraging temporal diversity in Twitter is inspired by the existing work on measuring diversity metrics to a textual document collection [6], where a diversity metric is the averaged similarity of all pairs of documents in the collection. Several measurements can be used for computing the pair-wise similarity between two documents, e.g., cosine similarity and Jaccard coefficient. In this work, we employ Jaccard coefficient because of its simplicity and high feasibility for a large scale collection, such as, Twitter data. Given a pair of tweets, the pair-wise similarity using Jaccard coefficient is computed as follows:

$$JC(d_i, d_j) = \frac{|O_{d_i} \cap O_{d_j}|}{|O_{d_i} \cup O_{d_j}|} \quad (1)$$

where O_{d_i} is a set of objects representing a tweet d_i, e.g., top-k terms extracted from a tweet and ranked by tf-idf. In addition to term-based features, we also represent a tweet with entity-based features, such as, location entities (country, continent and latitude). $|O_{d_i} \cap O_{d_j}|$ and $|O_{d_i} \cup O_{d_j}|$ are

the size of intersection and union of representative objects of tweets d_i and d_j, respectively. The higher the value of Jaccard coefficient, the higher the similarity of a tweet pair.

Temporal diversity can be computed as *Refined Diversity Jaccard Index* [6], which is in fact the average Jaccard similarity of all tweet pairs in a set of tweets D_{t_k} at particular time t_k. The diversity score of a set of tweets at time t_k can be computed as follows:

$$div(t_k) = \frac{2}{n \cdot (n-1)} \sum_{i<j} JC(d_i, d_j) \quad (2)$$

where $1 \leq i < j \leq n$ and $n = |D_{t_k}|$ is the number of tweets at particular time t_k, i.e., a single day. In addition, a sliding time window ω will be applied in order to smoothing any two adjacent days. The novelty of our work is that we represent a tweet using both term-based and entity-based features, namely, *top-k terms* and *location entities*. Hence, two types of diversity metric will be determined for each tweet representation, such as, div_{term} and div_{entity}. We combine both diversity metrics using a mixture model defined as follows:

$$div_{mix}(t_k) = \alpha \cdot div_{entity}(t_k) + (1-\alpha) \cdot div_{term}(t_k) \quad (3)$$

where α underlines the importance of both metrics. The values of α and ω will be empirically determined.

Since all-pair comparison is a well-known time complexity problem, we employ an efficient method for computing diversity based on sampling, such as, the SampleDJ Algorithm [6]. The approach selects a random sample of input by sampling r tweets uniformly at random (without replacement) from a collection of n documents. Then, the sum of similarities of all document pairs in the sample is computed thus scaling the diversity index of the sample. More precisely, the method samples from all possible pairs, computes the similarities of those r pairs and scales the result according to r and n as the final estimate. The detailed description of the algorithm can be referred to [6].

4. EXPERIMENTS

In this section, we explain experimental settings and discuss the results of our experiments.

4.1 Experimental Settings

Document Collections. We collected tweets from the year 2011 using more than 1,200 terms (i.e., disease names, synonyms, pathogens and symptoms) provided by the medical domain experts. Our document collection consists of 112,134,136 tweets collected in about one year. In addition to Twitter data, we also used ProMED-mail[3], a global reporting system providing information about outbreaks of infectious diseases. We gathered 3,056 ProMED-mail reports from the year 2011 as outbreak ground truth. All documents and tweets were annotated with locations, medical conditions and temporal expressions using a series of language processing tools, including OpenNLP[4] for tokenization, sentence splitting and part-of-speech tagging, HeidelTime [19] for temporal expression extraction, and OpenCalais [5] for named entity recognition. To index tweets, the Apache Lucene [6] search engine version 2.9.3 was used. We employ the Yahoo! PlaceFinder API[7] as a geo-tagger tool.

Table 4: List of real-world outbreaks represented by ID, disease name, country and event period.

ID	Disease	Country	Event Period
1	anthrax	Bangladesh	[11-May,18-Jun]
2	anthrax	India	[03-Jun,22-Jun]
3	botulism	Finland	[17-Oct,01-Nov]
4	botulism	France	[01-Sept,10-Sept]
5	cholera	Kenya	[11-Nov,03-Dec]
6	ebola	Uganda	[13-May,30-Jun]
7	ehec	Germany	[05-May,30-Jun]
8	leptospirosis	Denmark	[02-Jul,23-Jul]
9	leptospirosis	Philippines	[27-Jun,15-Jul]
10	mumps	Canada	[10-Jun,17-Aug]
11	mumps	United States	[27-Sept,11-Oct]
12	norovirus	France	[16-Jul,25-Jul]
13	rubella	Fiji	[26-Jul,09-Aug]
14	rubella	New Zealand	[15-Aug,19-Aug]

Outbreak Ground Truth. To validate the usefulness of Twitter data, the real-world outbreak statistics are required. An important aspect of our work is that we consider the duration of each outbreak by analyzing temporal expressions in a ProMED-mail document, unlike aforementioned work [2] that assumes the publication date of a document as the estimated event time of an outbreak. Particularly, we determined the starting date of a disease by looking at *the first* ProMED-mail post, and the ending date was related to *the last* ProMED-mail publication for that disease-location.

The reason for this is that information in ProMED-mail undergo moderation, so there is often a delay between the time of the actual outbreak and the publication date of the related report. However it is worth noting that this strategy gives us a good confidence only on the beginning date of the outbreak. Indeed, the absence of further ProMED-mail posts does not necessarily mean an end of the outbreak, but just that there was no significant news in which it was reported. The detailed description of ground truth creation was described in our work [11]. Finally, we identified 14 different outbreaks occurring in 2011 listed in Table 4.

Parameter Settings. The parameters for computing diversity metrics were: $k = 100$ terms, $\omega = 7$ days, and the mixture parameter α is varied in a range of $\{0.0, 0.1, \ldots, 1.0\}$. The RandomDJ Algorithm takes two parameters, i.e., an error threshold and a confidence, which are set to 0.05 and 0.9 respectively. For clustering tweets, we used the Lingo clustering algorithm provided by Carrot2 - an Open Source Search Results Clustering Engine [8] with default parameters.

Metrics. Correlation coefficient [24] was used to measure the statistical correlation between the temporal diversity scores and the outbreak-related topics over time, which ranges from 1 for perfectly correlated results, through 0 when there is no correlation, to -1 when the results are perfectly correlated negatively.

[3] http://www.promedmail.org
[4] http://opennlp.apache.org/
[5] http://www.opencalais.com/
[6] http://lucene.apache.org/
[7] http://developer.yahoo.com/geo/placefinder/
[8] http://project.carrot2.org/

4.2 Results

Figure 1 and Figure 2 illustrate the topics over time and temporal diversity of different infectious diseases. The illustrations show the temporal development for a particular infectious disease without differentiating among outbreaks occurring in different locations. The reason for this is that we want to study the *global trends* for all infectious diseases. As can be observed, outbreak-related topics show similar trends during the known time periods of real-world outbreak events (cf. Table 4). For almost all of diseases, we notices interesting, strongly time-dependent patterns in topics, except those for *anthrax* and *rubella* where no clear pattern can be found.

The results of temporal diversity vary greatly, which reflects how the language (i.e., terms and location entities) are used differently among diseases over time. In order to gain more insight, we calculate the correlation coefficient between two time series, namely, temporal diversity and outbreak-related topics over time. The correlation results are displayed in Table 5, where the highest (absolute) correlation value of each outbreak is marked in bold. Note that, the diversity metric used for correlating is a mixture value (div_{mix}) of the diversities computed using top-k terms and location entities, which are corresponding to div_{term} and div_{entity} respectively. The results of the mixture model are computed at different values of the mixture parameter α.

Our observation is that the entity-based diversity metric is highly correlated to the real-world outbreak topics for some diseases, such as, *mumps*, *ebola*, *botulism* and *ehec*. On the other hands, the term-based diversity metric exhibits to some extend the correlation with the real-world outbreak topics for infectious diseases like *cholera*, *anthrax* and *rubella*. The disease *leptospirosis* receives high correlation coefficients for the combination of div_{term} and div_{entity}, which can reflect the *leptospirosis* outbreaks in Denmark and Philippines. However, no clear assumption can be concluded about selecting the optimal value for α.

To conclude, we see high correlation between outbreak-related topics and temporal diversity for most cases. Our plan for future is to improve the calculation of diversity metrics in two ways: 1) better representation of tweets, and 2) employ a semantic-based similarity measurement.

5. RELATED WORK

In this section, we present related work on analyzing and mining Twitter data towards applications in the generic and medical domains.

Social media search and retrieval. Sakaki et al. [17] use Twitter to detect a real-time event, such as, Earthquake. They propose a probabilistic spatio-temporal model for the target event that can find the center and the trajectory of the event location. More precisely, they model the probability of an event occurrence at time t using an exponential distribution in a homogeneous Poisson process and the location of the event is estimated using the Kalman filter and particle filter algorithms. Nagmoti et.al [13] employ implicit information for improving ranking. They exploit the social network created by the ranking highly active authors (TweetRank) and their followers (FollowerRank) to re-rank the top-k tweets of the TABS microblog search engine. An slightly different approach is taken by Dong et al. [8], who exploit Twitter data to support recency ranking for general web search results. A comprehensive survey paper outlining other challenges related to search and retrieval of microblogs can be found in [9].

Medical-domain applications. In the medical domain, there has been a recent surge in detecting public health related tweets for early warning. The focus has been on building classifiers for detecting self-reported illness [4, 18], syndromes [3] and ailments [15]. Others have focused on validating the timeliness of Twitter by correlating tweets with real-world statistics such as Influenza-like-Illness (ILI) rates [5]. These works have shown that for a common disease like influenza, twitter can be used as a proxy for predicting flu outbreaks [20] and detecting illness with a population [1].

6. CONCLUSIONS AND FUTURE WORK

In this paper, we proposed an approach to computing diversity metrics in Twitter. In addition, we presented a method to extract topic dynamics using an unsupervised clustering technique. We studied how the temporal diversity metrics correlate to the real-world outbreak statistics, namely, the topic dynamic of an outbreak event. As a plan for future work, we seek to improve the diversity calculation in two ways: 1) a new representation for tweets, e.g., using other types of entities, and 2) employ a semantic-based similarity measurement [21]. Another possible direction to improve outbreak detection is to study a new feature, namely, opinion dynamics over time. More precisely, we can combine both temporal diversities and opinion dynamics in order to investigate how they reflect real-world events, such as, disease outbreaks. To this end, we want to validate the timeliness of Twitter for an early outbreak detection task by comparing to the *earliest known time* of real-world outbreak events. In order that, we will exploit a standard time series analysis method, namely, cross-correlation coefficient for estimating how variables are related at different time lags.

7. REFERENCES

[1] E. Aramaki, S. Maskawa, and M. Morita. Twitter catches the flu: Detecting influenza epidemics using twitter. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, EMNLP '11, 2011.

[2] N. Collier. What's Unusual in Online Disease Outbreak News? *Journal of Biomedical Semantics*, 1(1):2, 2010.

[3] N. Collier and S. Doan. Syndromic classification of twitter messages. *CoRR*, abs/1110.3094, 2011.

[4] N. Collier, N. T. Son, and N. M. Nguyen. Omg u got flu? analysis of shared health messages for bio-surveillance. *CoRR*, abs/1110.3089, 2011.

[5] A. Culotta. Towards detecting influenza epidemics by analyzing twitter messages. In *Proceedings of the First Workshop on Social Media Analytics*, SOMA '10, 2010.

[6] F. Deng, S. Siersdorfer, and S. Zerr. Efficient jaccard-based diversity analysis of large document collections. In *Proceedings of the 21st ACM International Conference on Information and Knowledge Management*, CIKM '12, 2012.

[7] E. Diaz-Aviles, A. Stewart, E. Velasco, K. Denecke, and W. Nejdl. Epidemic intelligence for the crowd, by

Table 5: Correlation coefficients of outbreak topic dynamics and temporal diversity at different values of α.

Outbreak	α=0	α=0.1	α=0.2	α=0.3	α=0.4	α=0.5	α=0.6	α=0.7	α=0.8	α=0.9	α=1.0
1	-0.0973	-0.1373	-0.1684	-0.1914	-0.2078	-0.2195	-0.2278	-0.2337	-0.2380	-0.2412	**-0.2435**
2	-0.1591	-0.2669	-0.3419	-0.3696	**-0.3701**	-0.3614	-0.3508	-0.3409	-0.3322	-0.3248	-0.3184
3	-0.1816	-0.1521	-0.1236	-0.0966	-0.0715	-0.0485	-0.0277	-0.0088	0.0082	0.0235	**0.0372**
4	**0.8285**	0.8137	0.7298	0.5752	0.3969	0.2403	0.1185	0.0272	-0.0415	-0.0939	-0.1349
5	-0.0663	-0.1951	-0.2982	-0.3795	-0.4437	-0.4946	-0.5355	-0.5688	-0.5961	-0.6189	**-0.6382**
6	**0.9134**	0.9010	0.8804	0.8508	0.8123	0.7655	0.7120	0.6536	0.5926	0.5310	0.4705
7	0.7218	0.7273	0.7305	**0.7319**	0.7316	0.7302	0.7278	0.7247	0.7210	0.7169	0.7125
8	0.8255	0.8538	**0.8689**	0.8556	0.7986	0.6957	0.5649	0.4317	0.3127	0.2133	0.1327
9	0.1943	0.2449	0.3018	0.3580	0.4021	**0.4242**	0.4237	0.4083	0.3866	0.3639	0.3427
10	0.4653	0.6425	0.7555	0.8233	0.8638	0.8884	0.9036	0.9132	0.9194	0.9233	**0.9258**
11	**0.9789**	0.9768	0.9631	0.9532	0.9462	0.9412	0.9374	0.9344	0.9320	0.9300	0.9284
12	**0.5398**	0.5070	0.4623	0.4009	0.3171	0.2064	0.0702	-0.0792	-0.2215	-0.3405	-0.4310
13	-0.1398	-0.1498	-0.1629	-0.1762	-0.1702	-0.1163	-0.0526	-0.0111	0.0137	0.0293	**0.0398**
14	0.5138	**0.5144**	0.5134	0.5088	0.4955	0.4623	0.3850	0.2361	0.0413	-0.1229	-0.2300

the crowd. In *International AAAI Conference on Weblogs and Social Media*, ICWSM '12, 2012.

[8] A. Dong, R. Zhang, P. Kolari, J. Bai, F. Diaz, Y. Chang, Z. Zheng, and H. Zha. Time is of the essence: improving recency ranking using twitter data. In *Proceedings of the 19th international conference on World wide web*, WWW '10, 2010.

[9] M. Efron. Information search and retrieval in microblogs. *Journal of the American Society for Information Science and Technology (JASIST)*, 62(6):996–1008, 2011.

[10] J. Gomide, A. Veloso, W. Meira, V. Almeida, F. Benevenuto, F. Ferraz, and M. Teixeira. Dengue surveillance based on a computational model of spatio-temporal locality of twitter. In *Proceedings of the 3rd ACM Web Science Conference*, WebSci '11, 2011.

[11] N. Kanhabua, S. Romano, and A. Stewart. Identifying relevant temporal expressions for real-world events. In *Proceedings of the SIGIR 2012 Workshop on Time-aware Information Access (TAIA '12)*, 2012.

[12] N. Kanhabua, S. Romano, A. Stewart, and W. Nejdl. Supporting temporal analytics for health-related events in microblogs. In *Proceedings of the 21st ACM international conference on Information and knowledge management*, CIKM '12, 2012.

[13] R. Nagmoti, A. Teredesai, and M. De Cock. Ranking approaches for microblog search. In *Proceedings of 2010 IEEE/WIC/ACM International Conference on Web Intelligence*, WI '10, 2010.

[14] B. O'Connor, R. Balasubramanyan, B. R. Routledge, and N. A. Smith. From tweets to polls: Linking text sentiment to public opinion time series. In *Proceedings of the Fourth International Conference on Weblogs and Social Media*, ICWSM '10, 2010.

[15] M. J. Paul and M. Dredze. You are what you tweet: Analyzing twitter for public health. In *Proceedings of the Fifth International Conference on Weblogs and Social Media*, ICWSM '11, 2011.

[16] E. J. Ruiz, V. Hristidis, C. Castillo, A. Gionis, and A. Jaimes. Correlating financial time series with micro-blogging activity. In *Proceedings of the fifth ACM international conference on Web search and data mining*, WSDM '12, 2012.

[17] T. Sakaki, M. Okazaki, and Y. Matsuo. Earthquake shakes twitter users: real-time event detection by social sensors. In *Proceedings of the 19th international conference on World wide web*, WWW '10, 2010.

[18] M. Sofean, A. Stewart, K. Denecke, and M. Smith. Medical case-driven classification of microblogs: Characteristics and annotation. In *Proceedings of SIGHIT International Health Informatics Symposium*, IHI '12, 2012.

[19] J. Strötgen and M. Gertz. Heideltime: High quality rule-based extraction and normalization of temporal expressions. In *Proceedings of the 5th International Workshop on Semantic Evaluation (SemEval '10)*, 2010.

[20] M. Szomszor, P. Kostkova, and E. de Quincey. #swineflu: Twitter predicts swine flu outbreak in 2009. In *Proceedings of the Third International Conference on Electronic Healthcare*, eHealth '10, 2010.

[21] G. Tsatsaronis, I. Varlamis, and K. Nørvåg. Semafor: semantic document indexing using semantic forests. In *Proceedings of the 21st ACM International Conference on Information and Knowledge Management*, CIKM '12, 2012.

[22] A. Tumasjan, T. O. Sprenger, P. G. Sandner, and I. M. Welpe. Predicting elections with twitter: What 140 characters reveal about political sentiment. In *Proceedings of the Fourth International Conference on Weblogs and Social Media*, ICWSM '10, 2010.

[23] S. Vieweg, A. L. Hughes, K. Starbird, and L. Palen. Microblogging during two natural hazards events: what twitter may contribute to situational awareness. In *Proceedings of the 28th International Conference on Human Factors in Computing Systems*, CHI '10, 2010.

[24] I. H. Witten and E. Frank. *Data Mining: Practical Machine Learning Tools and Techniques, Second Edition*. Morgan Kaufmann, 2005.

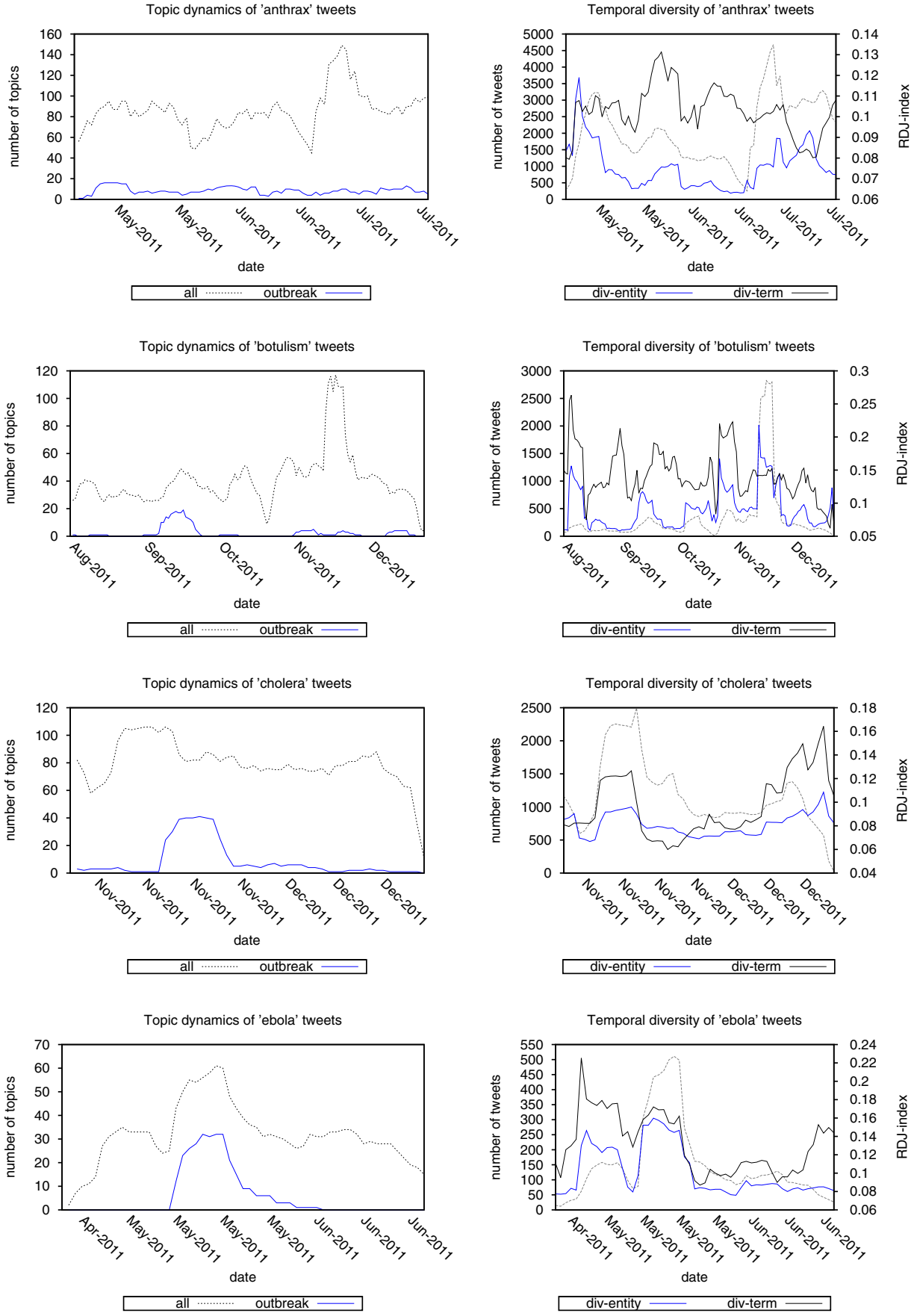

Figure 1: Illustration of topic dynamics vs. temporal diversity for different infectious diseases.

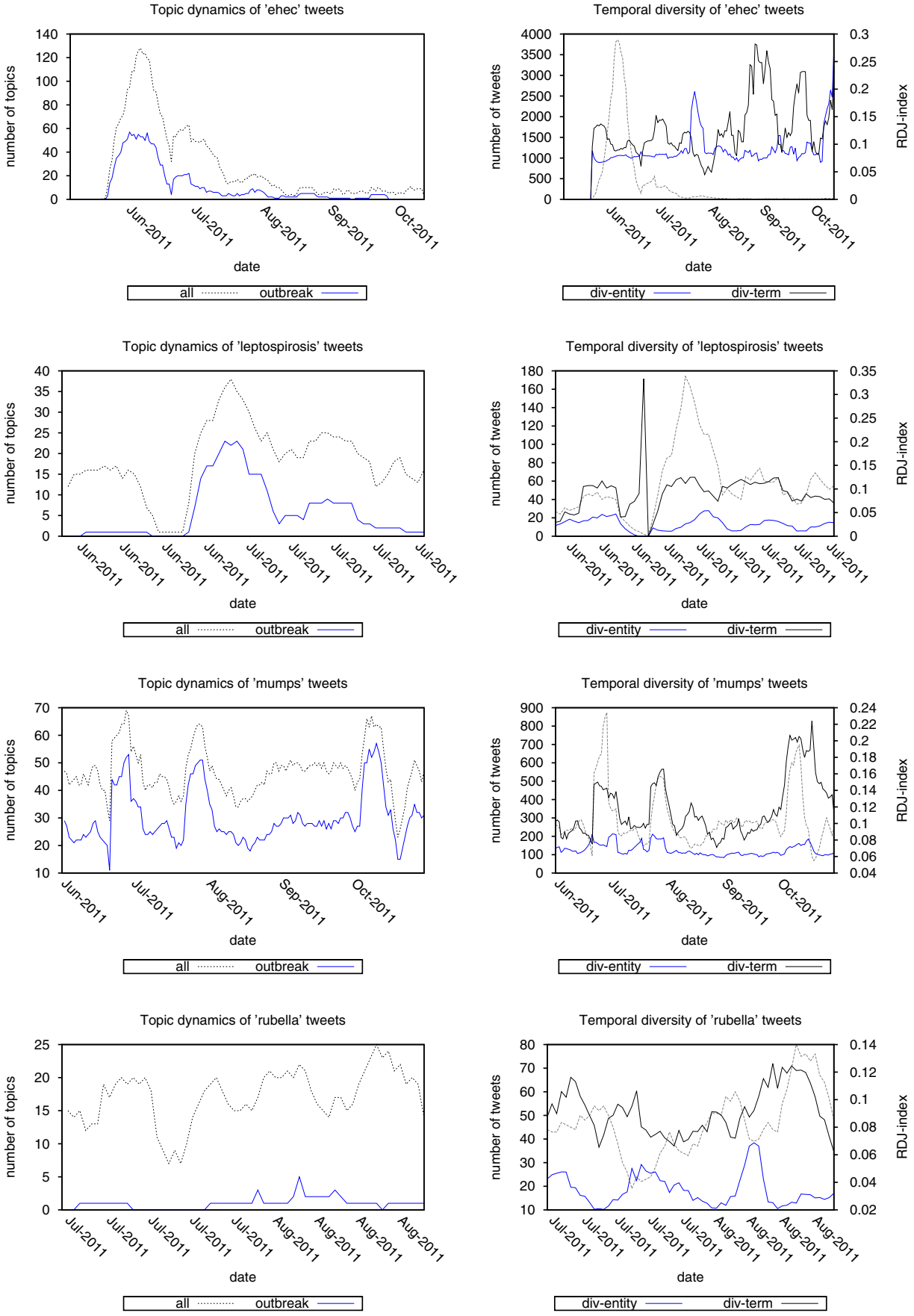

Figure 2: Illustration of topic dynamics vs. temporal diversity for different infectious diseases.