Tracking Topics of Influential Tweets on Fukushima Disaster over Long Periods of Time

Hiroshi Nagaya
School of Engineering
The University of Tokyo
Tokyo, Japan
hiroshi-nagaya@g.ecc.u-tokyo.ac.jp

Kazuko Uno
Louis Pasteur Center
for Medical Research
Kyoto, Japan
kazukouno@louis-pasteur.or.jp

Hiroyuki A. Torii
School of Science
The University of Tokyo
Tokyo, Japan
torii@chem.s.u-tokyo.ac.jp

***Abstract*—Social media has been extensively and effectively deployed to share information and communicate during emergencies, such as the 2011 Fukushima Daiichi nuclear power plant accident in Japan. It is important to provide information during crises on social media and find the most effective way to transmit information in such situations. It is necessary to carefully preprocess Twitter data because it includes a considerable amount of noise. However, compared to other resources, such as government statistics and newspapers, Twitter provides varied information and is distinguished by its immediacy. We can also regard Twitter data as data that reflect human behaviors, thoughts, and intentions across different domains by characteristics of the platform. We propose an expansion model of Topic Dynamics for tracking the trend and detecting the moments of the occurrences of influential tweets on the Fukushima disaster. Using this method, we obtained the list of bursting words at different periods over a long duration following the Fukushima disaster.**

***Index Terms*—Twitter, Social media, Disaster situation, Text mining, Burst detection**

I. Introduction

TWITTER is a social media platform on which users can post small texts called "tweets," which can be shared in a process known as "Retweet". Users can also communicate directly with other users on the platform. The distinctive features of Twitter are the immediacy and diffusibility of information transmission. Because of these advantages, Twitter is an effective tool for sharing information and communicating. Furthermore, compared to other resources, such as government statistics and newspapers, the variety of information on Twitter and its immediacy makes Twitter data preferable as a subject of analysis. We can also regard Twitter data as data that reflect human behaviors, thoughts, and intentions across different domains by characteristics of the platform.

On the 11th of March, 2011, the Great East Japan Earthquake occurred. It led to the Fukushima Daiichi nuclear power plant accident. It was difficult to collect detailed information, and the lack of information was a particular source of concern. Because of the resultant traffic congestion and severe damage to the network infrastructure, media reporters were unable to assess the disaster-struck area. Under the situation, social media was extensively used and attracted a considerable amount of attention for its information sharing capabilities [1]. In this and similar situations, it is necessary to clarify occurrences on social media and find an efficient method to propagate the required information. If used correctly, social media can be a most effective tool for transmitting information.

II. Literature Review

According to Tsubokura et al. [2], the majority of retweets on the Fukushima disaster were based on original posts sent out by a few hundred accounts held by people described as influencers. Influencers play an important in information propagation on social media. In the context of the dataset of interest in this study, the influencers communicated important information relating to the emergency efficiently.

Although some studies explore the behavior of Twitter users under emergency situations [3], [4], scant studies have analyzed the topics that have emerged under these situations especially over long periods following disasters. In this study, based on the influential tweets over time, we analyze tweet data relating to the Fukushima disaster.

III. Dataset and Method

A. About Data

We used the Twitter data related to the Fukushima disaster. It comprised of tweets from the 1st of January, 2011 to the 30th of June, 2019. Only 8 percent of the entire tweets were sampled, because the tweets were surplus. Each tweet included at least one of the keywords listed in table I. These keywords are the same as those used in a previous research [2]. The data was purchased from the NTT DATA Corporation, a major telecommunications company in Japan.

B. Preprocessing

The dataset includes a lot of unrelated tweets that included the keywords. For instance, "Monitoring", the title of a popular TV program in Japan, is mentioned in a lot of tweets. To circumvent this problem, we re-extracted tweets from the original dataset with an additional constraint that the tweets included a Fukushima ("福島", "ふくしま" or "フクシマ") in addition to the keywords; we also removed tweets that included no Japanese. Following sampling, we obtained a new dataset comprising of the 21,898,729 tweets and retweets of 2,809,329 users.

2375-9259/19/$31.00 ©2019 IEEE
DOI 10.1109/ICDMW.2019.00010

For the subsequent analysis, it was necessary to conduct some preprocessing. First, we removed the URLs and some Twitter-specific signifiers: "RT @[userID]:", "#". Next, we conducted morphological analysis and represented the data as Bag-of-Words (BOW) using MeCab [5], an open-source library developed by the Nara Institute of Science and Technology. We also used NEologd [6] to supplement Mecab. In our analysis, we focused solely on nouns. Furthermore, we eliminated some stop words that were unrelated to the contents (e.g. the Japanese equivalents of the English "they", "I" or "we")

TABLE I
KEYWORDS; SAME AS THE ONES USED IN [2]

Keywords in Japanese	English transition
放射	radio- or radia-
被ばく, 被曝, 被爆	exposure
除染	decontamination
線量	dose
ヨウ素	iodine
セシウム	cesium
Sv, mSV, μ SV, uSV , msv, μ sv, usv, シーベルト	Sv, sievert
ベクレル	becquerel
Bq	Bq
ガンマ線, γ 線	gamma ray, γ -ray
核種	isotope
甲状腺, 甲状線	thyroid
チェルノブイリ	Chernobyl
規制値	regulation value
基準値	standard value
学会	academic society
警戒区域	no-entry zone
避難区域	evacuation zone
産科婦人科	obstetrics and gynecology
周産期・新生児医	perinatal and neonatal care
日本疫	nuclear medicine
核医	nuclear medicine
電力中央	central electric
学術会議	science council
環境疫	environmental epidemiology
物理学会	Physical Society
プルトニウム	plutonium
ストロンチウム	strontium
暫定基準	provisional standard
暫定規制	provisional regulation
屋内退避	sheltering
金町浄水場	Kanamachi Water Purification Plant
出荷制限	shipment restriction
管理区域	control area
避難地域	evacuation area
モニタリング	monitoring
スクリーニング	screening
ホットスポット	hot spot
汚染	contamination
検査 AND (食品 OR 水 OR 土)	inspection AND (food OR water OR soil)
リスク AND (がん OR ガン OR 癌)	risk AND cancer
影響 AND (妊婦 OR 妊娠 OR 出産 OR 子ども OR 子供 OR こども OR 児)	effect AND (pregnant woman OR pregnancy OR childbirth OR child)
母子避難	mother and child evacuation
避難弱者	people having difficulty in evacuation
自主避難	voluntary evacuation
避難関連死, 避難死	death associated with evacuation
(福島 OR ふくしま OR フクシマ) AND (避難 OR 米 OR 野菜 OR 牛肉 OR 食品 OR 産 OR 安全 OR 安心 OR 不安 OR 検査)	Fukushima AND (evacuation OR rice OR vegetable OR beef OR food OR product OR safety OR relief OR anxiety OR inspection)

C. Topic Jerk Detector: Detection of Bursting Words

We expanded the topic dynamics model [7], and propose the topic jerk detector, to track trends on twitter, and detect the moments of occurrences related to the Fukushima disaster. Topic dynamics is based on the move and change delete (MACD) histogram that is mainly used to analyze the trends of stock prices in the technical analysis of financial markets.

First, we defined the score $s_{w,t}$ for each word w at time t in Eq1. The scores correspond to the stock prices in financial markets.

$$s_{w,t} = \sum_{d \in D} \log(rt_{d,t}) * tf_{w,d,t} * \log(\frac{N_t}{df_{w,t}}) \quad (1)$$

Here, D is a set of tweets that include the word w of which tweet d is an element. $rt_{d,t}$ is the number of retweets of the tweet d at time t. We express the impacts of the tweets in $\log(rt_{d,t})$. $tf_{w,d,t}$ is the frequency at which word w appears in tweet d at time t. N_t is the total number of tweets before time t, and $df_{w,t}$ is the number of tweets containing the keyword before time t. Next, we calculate the exponential moving average (EMA) as follows:

$$\begin{aligned} \mathrm{EMA}^n[s_{w,t}] &= \alpha * s_{w,t} + (1-\alpha) * \mathrm{EMA}_w^{n-1}[s_{w,t}] \\ &= \sum_{k=0}^{n} \alpha(1-\alpha)^k s_{w,t-k} \end{aligned} \quad (2)$$

$$\alpha = \frac{2}{n+1} \quad (3)$$

n refers to the span of date. The MACD is defined as the difference between the moving averages ($n_2 > n_1$) of the n_1-day and n_2-day:

$$\mathrm{MACD}_w^{(n_1,n_2)} = \mathrm{EMA}^{n_1}[s_{w,t}] - \mathrm{EMA}^{n_2}[s_{w,t}] \quad (4)$$

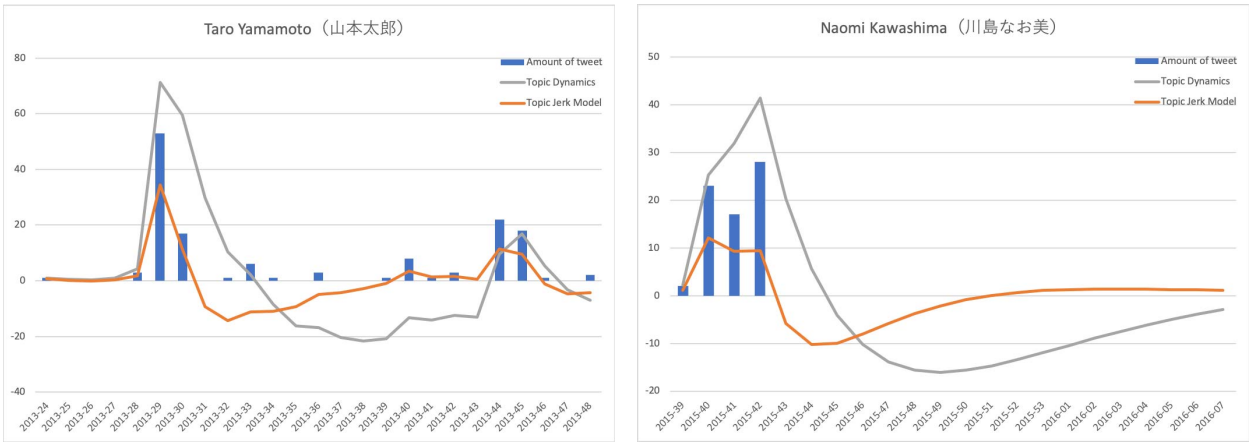
Furthermore, the MACD histogram is defined as the difference between the MACD and its moving average:

$$\begin{aligned} &\mathrm{histogram}_w^{(n_1,n_2,n_3)} \\ &\quad = \mathrm{MACD}_w^{(n_1,n_2)} - \mathrm{EMA}^{n_3}[\mathrm{MACD}_w^{(n_1,n_2)}] \end{aligned} \quad (5)$$

From the perspective of signal processing, the MACD can be interpreted as the derivative of $s_{w,t}$; topic dynamics(MACD histogram) is the second derivative [8]. Using the analogy of physical quality, the MACD corresponds to velocity, and topic dynamics corresponds to acceleration. Topic dynamics detects the bursting of words when the value of the MACD histogram is positive. We propose the topic jerk detector as an expansion of the topic dynamics model:

$$\begin{aligned} &\mathrm{TopicJerkDetector}_w^{(n_1,n_2,n_3,n_4)} \\ &= \mathrm{histogram}_w^{(n_1,n_2,n_3)} - \mathrm{EMA}^{n_4}[\mathrm{histogram}_w^{(n_1,n_2,n_3)}] \end{aligned} \quad (6)$$

Once again, we make reference to the analogy of physical quality; the acceleration derivative can be regarded as a jerk. The jerk is a physical quantity that indicates a large value corresponding to a sudden increase in acceleration (e.g. if a car peel outs, a jerk shoots a moon.) Because of this feature, it is anticipated that the topic jerk detector accurately captures the timing of words bursting. Here, we summarize the correspondence between the methods and physical qualities in Table II.



(a) Yamamoto (b) Kawashima

Fig. 1. Model comparison. The orange and gray lines are the values of the topic jerk detector and topic dynamics, and blue bar is the amount of tweet.

TABLE II
METHODS AND CORRESPONDING PHYSICAL QUALITIES

Indicator	Corresponding physical quantity
MACD	Velocity$[\frac{1}{s}]$
Topic Dynamics (Based on MACD histogram)	Acceleration$[\frac{1}{s^2}]$
Topic Jerk Detector (Our proposed method)	Jerk$[\frac{1}{s^3}]$

IV. Result and Discussion

We proposed and applied the topic jerk detector to our dataset to reveal the variation in the topics occurring over time. First, we extracted the top hundred retweeted tweets (defined as "influential tweets") each week and calculated their values in the topic jerk detector. We listed the most frequently occurring words for each week. In Eqs. 4-6, it was necessary to determine four parameters (n_1, n_2, n_3, n_4). In the technical analysis of financial markets, the values of (n_1, n_2, n_3) are customarily set as $(9, 12, 26)$. We diverted these values and qualitatively detected n_4 as 3. The result is shown in Appendix (Figs. 2-3). We succeeded in listing the burst words for each week. There were some key persons whose remarks were taken up on Twitter (people whose names are in red). It may be attributable to Twitter' s sensitivity to the remark of specific individuals.

To compare the topic jerk detector and topic dynamics, we extracted some words from conversations on selected famous people on Twitter and plotted the value of each model around the burst timings in Fig. 1. The orange and gray lines are the values of the topic jerk detector and topic dynamics, respectively; the blue bars correspond to the number of tweets comprising of the influential tweets and retweets each week. As can be seen from these graphs, these models simultaneously capture the timing of word bursts. Subsequently, the value of the topic jerk detector drops sharply and converges to zero faster than the value of the topic dynamics. Thus, it may be concluded that the topic jerk detector detects the timing of the burst more accurately.

V. Conclusion

A. Summary

Occurrences of disasters need to be communicated on social media, and finding the best way to transmit information in such situations is of great importance. Thus, we proposed the topic jerk detector, based on the expansion of topic dynamics, to track topics of influential tweets, and cache the timing of the word bursts. Because of the application of the model to our dataset, we succeeded at revealing the words that received the most attention on Twitter following the Fukushima Daiichi nuclear power plant accident. We observed that there were some important people whose remarks were presented on Twitter. We also compared the outputs of the models and demonstrated that our proposed topic outperformed the existing method at detecting the timing of bursts.

B. Future Works

In our future works, we will analyze the network of retweets to answer the questions of "what" and "how" by exploring the kind of users drawn by each topic and how the network changed. Furthermore, we will also explore the best way to transmit information during and in the wake of emergencies.

Acknowledgment

This work was supported by the Research on the Health Effects of Radiation initiative organized by the Ministry of the Environment, Japan.

REFERENCES

[1] F.Toriumi, T.Sakaki, K.Shinoda, K.Kazama, S.Kurihara, I.Noda. "Information sharing on Twitter during the 2011 catastrophic earthquake," in Proceedings of the 22nd International World Wide Web conference, 2013, pp. 1025-1028.

[2] M. Tsubokura, Y. Onoue, H. A. Torii, S. Suda, K. Mori, Y. Nishikawa, et al. "Twitter use in scientific communication revealed by visualization of information spreading by influencers within half a year after the Fukushima Daiichi nuclear power plant accident," in PLoS One, 2018; 13:e0203594.

[3] T. Sakaki, F. Toriumi, Y. Matsuo. "Tweet trend analysis in an emergency situation," in Proceedings of the Special Workshop on Internet and Disasters, SWID'11, 2011, pp. 3:1-3:8.

[4] M. Mendoza, B. Poblete, C. Castillo. "Twitter under crisis: can we trust what we RT?" in Proceedings of the First Workshop on Social Media Analytics - SOMA'10, 2010, pp. 71 - 79.

[5] T. Kudo, K. Yamamoto, Y. Matsumoto. "Applying Conditional Random Fields to Japanese Morphological Analysis," in Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing, 2004, pp.230-237.

[6] T. Sato. "Neologism dictionary based on the language resources on the Web for unidic-mecab, " https://github.com/neologd/mecab-unidic-neologd, 2015.

[7] D. He and D. S. P. Jr., "Topic Dynamics: An alternative model of bursts in streams of topics," in Proc. 16th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining, 2010, pp. 443-452.

[8] G. Stanley. "The MACD approach to derivative (rate of change) estimation, " https://gregstanleyandassociates.com/whitepapers/FaultDiagnosis/Filtering/MACD-approach/macd-approach.htm, 2010-2013. accessed on 25 July 2019.

APPENDIX

2011-10	2011-11	2011-12	2011-13
-	discrimination	TEPCO	plutonium
2011-14	2011-15	2011-16	2011-17
Masayoshi Son	Chernobyl	school	msv
2011-18	2011-19	2011-20	2011-21
damage	broadcast	radioactive contamination	sv
2011-22	2011-23	2011-24	2011-25
Follow-up report	detection	radioactivity	Fukushima nuclear power plant
2011-26	2011-27	2011-28	2011-29
cesium	Thyroid gland	beef	Straw
2011-30	2011-31	2011-32	2011-33
word	measurement	Radioactive iodine	msv
2011-34	2011-35	2011-36	2011-37
milk	pain	The Ministry	**Hachiryo**
2011-38	2011-39	2011-40	2011-41
fireworks	plutonium	Thyroid gland	Setagaya
2011-42	2011-43	2011-44	2011-45
Radiation dose	Decontamination	Nuclear fission reaction	ekiden
2011-46	2011-47	2011-48	2011-49
rice	rice	TEPCO	Meiji
2011-50	2011-51	2011-52	2012-01
Noda	cesium	ceder	cesium
2012-02	2012-03	2012-04	2012-05
cesium	Nuclear warfare	lump	Kumamoto
2012-06	2012-07	2012-08	2012-09
earthworm	temperature	cancer	son
2012-10	2012-11	2012-12	2012-13
shit	Japan	death	store
2012-14	2012-15	2012-16	2012-17
chip of wood	swimming pool	Fukushima Watari	Carnation
2012-18	2012-19	2012-20	2012-21
powdered milk	cost	evacuation	rubble
2012-22	2012-23	2012-24	2012-25
fertilizer	**Noda**	Hirosaki Univ.	Fukushima
2012-26	2012-27	2012-28	2012-29
plutonium	prayer	guardian	subcontract
2012-30	2012-31	2012-32	2012-33
strontium	cyst	Hiroshima	cc
2012-34	2012-35	2012-36	2012-37
greenling	other than	milk	thyroid cancer
2012-38	2012-39	2012-40	2012-41
cyst	childhood thyroid cancer	baby girl	Fukushima Prefectural Office
2012-42	2012-43	2012-44	2012-45
people	kg	drying	monitoring post
2012-46	2012-47	2012-48	2012-49
glimpse	specimen	**Ichiro Ozawa**	people
2012-50	2012-51	2012-52	2013-01
resident of Fukushima	declaration	TEPCO	decontamination
2013-02	2013-03	2013-04	2013-05
element	way	sad news	Oishinbo
2013-06	2013-07	2013-08	2013-09
Oishinbo	thyroid cancer	thyroid cancer	operator
2013-10	2013-11	2013-12	2013-13
erasing	**Yuzuru Nishida**	power cuts	Sazae-san
2013-14	2013-15	2013-16	2013-17
underground	contaminated water	leakage	evacuation

Fig. 2. Words with the largest values according to topic jerk detector for each week (from 2011-11 to 2013-17). Each period is expressed using the pair of years and ISO week. People's names are in red.

2013-18	2013-19	2013-20	2013-21
radioactivity	photographing	immigration	health
2013-22	2013-23	2013-24	2013-25
United Nations	Doubt	Oita Prefecture	strontium
2013-26	2013-27	2013-28	2013-29
Tritium	well	**Hiroshi Suzuki**	**Taro Yamamoto**
2013-30	2013-31	2013-32	2013-33
Taro Yamamoto	flowing out	flowing out	ice
2013-34	2013-35	2013-36	2013-37
thyroid cancer	contaminated water problem	Tokyo	harbor
2013-38	2013-39	2013-40	2013-41
block	Becquerel	Yoshinoya	operator
2013-42	2013-43	2013-44	2013-45
typhoon	last night	**Taro Yamamoto**	**Taro Yamamoto**
2013-46	2013-47	2013-48	2013-49
thyroid cancer	interim storage facility	hearing	well
2013-50	2013-51	2013-52	2014-01
education	TEPCO	contaminated water	homeless
2014-02	2014-03	2014-04	2014-05
contaminated water	contaminated water	contaminated water	Tokyo
2014-06	2014-07	2014-08	2014-09
Tamogami	thyroid cancer	tank	medical school
2014-10	2014-11	2014-12	2014-13
doctor	news station	purification	exposure dose
2014-14	2014-15	2014-16	2014-17
United Nations	**Obokata**	nuclear power plant	dead bodies
2014-18	2014-19	2014-20	2014-21
nosebleed	nosebleed	nosebleed	doubt
2014-22	2014-23	2014-24	2014-25
associate professor	tank	Shukan Gendai	monetary value
2014-26	2014-27	2014-28	2014-29
permeation	return	Tokiwa	withdrawal
2014-30	2014-31	2014-32	2014-33
sv	monitoring post	bomb victim	dry ice
2014-34	2014-35	2014-36	2014-37
exposure	thyroid cancer	dengue fever	**Obuchi**
2014-38	2014-39	2014-40	2014-41
passage	west	Tokyo	groundwater
2014-42	2014-43	2014-44	2014-45
groundwater	dismantling	agricultural land	cameraman
2014-46	2014-47	2014-48	2014-49
bag	cement	contaminated water	analysis
2014-50	2014-51	2014-52	2015-01
MSv	tritium	thyroid cancer	scattering
2015-02	2015-03	2015-04	2015-05
pharmaceuticals	GPS	purification	dosimeter
2015-06	2015-07	2015-08	2015-09
newspaper	newspaper	drainage	contaminated water
2015-10	2015-11	2015-12	2015-13
Fukushima	decontamination	**Otsuka Ai**	impact
2015-14	2015-15	2015-16	2015-17
discrimination	abnormality	Taiwan	drawn
2015-18	2015-19	2015-20	2015-21
drawn	radioactive contamination	housing	thyroid cancer
2015-22	2015-23	2015-24	2015-25
harbor	vegetable plants	voluntary evacuation	voluntary evacuation
2015-26	2015-27	2015-28	2015-29
Fukushima Prefecture	through	private	rain water
2015-30	2015-31	2015-32	2015-33
decontamination	assumption	Paris	Sendai nuclear power plant
2015-34	2015-35	2015-36	2015-37
decontamination	fir	thyroid cancer	flowing out
2015-38	2015-39	2015-40	2015-41
flowing out	surgical operation	**Naomi Kawashima**	**Toshihide Tsuda**
2015-42	2015-43	2015-44	2015-45
Naomi Kawashima	leukemia	events	inspection
2015-46	2015-47	2015-48	2015-49
refuge	wiping	paper	inspection
2015-50	2015-51	2015-52	2015-53
underground	report	forest	zone
2016-01	2016-02	2016-03	2016-04
North Korea	operator	Small fish	construction
2016-05	2016-06	2016-07	2016-08
neighborhood	high school students	thyroid cancer	cause
2016-09	2016-10	2016-11	2016-12
record	thyroid cancer	work	wheat
2016-13	2016-14	2016-15	2016-16
forest fire	salted plum	Sendai nuclear power plant	cost
2016-17	2016-18	2016-19	2016-20
Chernobyl	reflux	high official	exposure
2016-21	2016-22	2016-23	2016-24
Hiroshima	frozen soil wall	thyroid cancer	Green Co-op
2016-25	2016-26	2016-27	2016-28
nationwide	apparatus	non-prosecution equivalent	foreigner
2016-29	2016-30	2016-31	2016-32
agricultural products	marine	sediment disaster	reduction
2016-33	2016-34	2016-35	2016-36
leukemia	reduction	burden	prevalence
2016-37	2016-38	2016-39	2016-40
donation	dam	dam	**Taro Yamamoto**
2016-41	2016-42	2016-43	2016-44
Yoneyama	group	group	car wash
2016-45	2016-46	2016-47	2016-48
refuge	bullying	high school students	homeroom teacher
2016-49	2016-50	2016-51	2016-52
voluntary evacuation	Workers' compensation certification	nosebleed	thyroid cancer
2017-01	2017-02	2017-03	2017-04
inspection	dose	Toyosu	voluntary evacuation
2017-05	2017-06	2017-07	2017-08
store	store	train crew	lecturer
2017-09	2017-10	2017-11	2017-12
Fukushima Airport	msv	professor	account book
2017-13	2017-14	2017-15	2017-16
Moritomo	voluntary evacuation	voluntary evacuation	rapid
2017-17	2017-18	2017-19	2017-20
fire	fire	fire	defoliation
2017-21	2017-22	2017-23	2017-24
United Nations	West	Impact	Niigata
2017-25	2017-26	2017-27	2017-28
death	voluntary evacuation	-	-

Fig. 3. Words with the largest values according to topic jerk detector each week. (from 2013-18 to 2017-26)