『学習院大学　経済論集』第53巻　第３号（2016年10月）

Topic Extraction Analysis for Sidoardjo Mudflow Disaster Impacts

Yussanti Nur Fajrina[1)], Yukari Shirota[2)], Riri Fitri Sari[3)]

ABSTRACT

In this paper, we present our work on analyzing the impact of the Mudflow Disaster in Sidoardjo, Indonesia, based on text mining technologies. We conducted a topic extraction using the Latent Dirichlet Allocation model. To handle the difficult expressions and grasp the points, we use various techniques such as bigram segmentation for documents related to the Mudflow in English. The TreeTagger is the morphological analysis tool used. The extracted topics clearly showed the impact of the Sidoardjo Mudflow. The most widely discussed topic found was the resettlement conditions and the compensation for the victim corresponding to the presidential regulation. We also found other frequently mentioned topics, such as the payment of resettlement, water pollution, and the verification process for the households.

Keywords: *Topic extraction, Dirichlet Allocation Model, Sidoardjo Mudflow, Compensation, Resettlement, Presidential Regulation*.

1 Introduction

On May 29th, 2006, mud and gases began erupting unexpectedly from a hydrocarbon exploration well near Sidoardjo, East Java, Indonesia. The eruption, called the LUSI (Lumpur Sidoardjo [Lumpur means mud in Indonesian]) of mud volcano, has continually flow out from the well since then at rates as high as 180,000 m^3 per day [1]. The Sidoardjo Mudflow spread widely and devastated many villages. The mudflow is still spreading, and is predicted to continue flowing for many decades to come. Responsibility for it was credited to the blowout of a natural gas well drilled by a company called Lapindo Brantas Inc. On the other hand, some scientists and company officials contend that it was caused by a distant earthquake [2].

1）　Department of Electrical Engineering, Faculty of Engineering, University of Indonesia, yussanti.nur@ui.ac.id
2）　Department of Management, Faculty of Economics, Gakushuin University, Tokyo, Japan, yukari.shirota@gakushuin.ac.jp
3）　Department of Electrical Engineering, Faculty of Engineering, University of Indonesia, riri@ui.ac.id

The Sidoardjo mudflow is a new type of disaster. The duration of this disaster is estimated to be 23-35 years, much longer than other types of disasters such as earthquakes, tornadoes, tsunamis, and floods [3]. Japan is an earthquake-prone country with many volcanoes, which caused some earthquake damage. However, there were no such case as Mudflow disasters found in Sidoardjo.

This paper presents the impacts of the large-scale mudflow and clarify the feature difference between the mudflow and an earthquake. The mudflow damage affect the population around the location in various ways. To investigate that, we need extensive reading of the documents and reports of the Sidoardjo mudflow. Then text mining techniques can help us. In this paper, we analyzed documents and reports using text mining technologies, focusing on the impacts and effects of the Sidoardjo mudflow. Our research aim is to support readers of the documents, so that many people including foreign people can instantly understand the contents. We implemented the topic extraction using the Latent Dirichlet Allocation (LDA) model. To handle the difficult expressions and grasp the points, we use various techniques such as bi-gram segmentation. The target of our analysis is English documents, using a morphological analysis tool called TreeTagger [4].

This paper is organized as follows. In Section 2, we explain about the Sidoardjo Mudflow briefly. In Section 3, we explain the topic extraction method used, which is based on the LDA model and the Gibbs sampling algorithm for implementing the LDA model. Subsequently, in Section 4, we explain the topic extraction results by the LDA model. In section 5, we discuss the differences between the mudflow and an earthquake disaster impact. Finally, we conclude the paper in the Section 6.

2 Sidoardjo Mudflow

In this section, we explain the disaster area and the reason why we analysed the Sidoardjo Mudflow disaster and its economic impacts [2, 3, 5-7].

(a) Disaster Area

The Sidoardjo mudflow area is located in Renokenongo village in the Porong subdistrict in Sidoardjo regency. There are 12 villages from 3 districts affecfed. The total area covered by the mud is

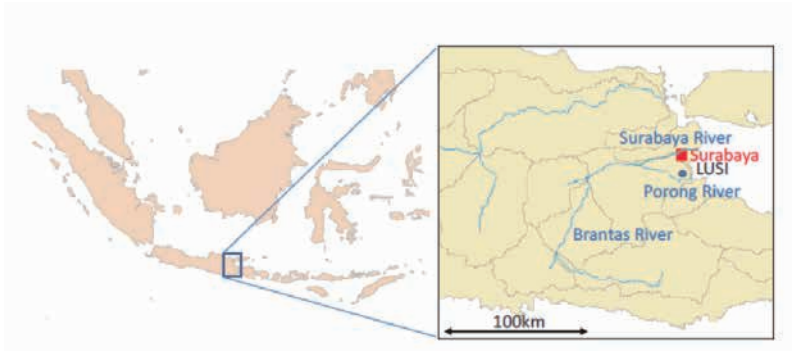

Figure 1. Location of the Brantas River basin, Surabaya city and the LUSI mud volcano(cited from [1])

approximately 640 hectares, or equal to 1,600 football fields wide. The area affected determined by the Indonesian Presidential Regulation. The mudflow spread area and the affected area determined (as per March 22nd 2007), so far still remains the same. Figure 1 shows the location of the LUSI mudflow in East Java

(b) Effects on Economics

The economic impact of the mudflow affected all facet of life, and damaged the economic to business sector in this large area and its surroundings. The region suffering the biggest lost is the central corridor from South Surabaya to Malang. Leather processing, food, hotels and restaurants industries were the most affected sectors. There were also hundreds of farms, rice fields, small businesses and 10 (ten) large factories directly affected by the mudflow [8]. Table 1 shows the Direct Economic Costs from 2006 to 2015.

Table 1: Direct Economic Costs - 2006 - 2015 (US$) (cited from [8])

No.	**Cost Component**	**2006**	**2007-2015**	**Total**
1	Lost Assets	131,467,000	1,729,972,000	$1,861,439,000
2	Lost Income	16,736,000	215,547,000	232,283,000
	Total	**148,203,000**	**1,945,519,000**	**$2,093,722,000**
Source: Brawijaya University Report on Economy Impacts Assessment of the Mud Flow 2006[9]				

Nowadays, the mudflow area becomes a tourist attraction. The people in the affected area created some statues and monuments to represent their sadness and madness to Lapindo Brantas Inc. Both International and local Indonesian tourists are eager to visit the area and to witness the peculiarity of the disaster area. Many ex-factory workers have become tour-guides on motorcycle. Tourism increased the income of those mudflow victims and has significant effect on the economic growth in Sidoardjo area.

(c) Relocation and Compensation

For compensation and relocation, the purchase of land and building for former residents of the disaster area come from two sources of financing. Land and buildings which have been submerged by the mudflow in the affected area map were solely financed by Lapindo Brantas Inc. The area outside the affected area were fully funded by the government through the state budget. The basic scheme of payment was by an advance payment (20%) and a further payment of redemption (80%). Table 2 depicted the amount agreed to be disbursed to the victims and the actual number of claim for that purpose.

Table 2: Compensation for Resettlement (cited from [8])

	Amount Agreed	Number of Claimants
Land And Building Compensation	$ 15,000 per Household on average	25,000
Evacuation Cost/Moving Cost	$ 50 per family	25,000
House Lease Assistance/House Rental Contract	2-year of $ 500 per family,	25,000
Monthly Living Assistance	$ 30 per month per person for 9 months,	50,000
Provide Food (3 Times/Day) at Shelter Locations	$ 2 per person per day	50,000
Provide Amenities and Facilities at Shelter Locations	No Agreement	50,000
Source: Brawijaya University Report on Economy Impacts Assessment of the Mud Flow 2006[9]		

From the document review, it can be summarized that there is another process needed to evaluate the economic impact by evaluating each topic probability from the documents. We can evaluate the economic impact deeper using text mining. First we need to collect and identify a set of textual materials, then we use text analytics methods and analysis.

3 Latent Dirichlet Allocation Model

In this section, we shall explain the LDA model and the Gibbs sampling process that we used for the topic extraction.

The LDA is a widely-used multi-topic document model based on Bayesian inteference method [10, 11]. The following is a simple explanation of the framework. In the LDA model, each topic is supposed to have a set of related words and one document is supposed to have several topics. To express the possible various distributions, we use the Dirichlet distribution by using a hyper parameter $\boldsymbol{\alpha}$. On the same way, we define per-topic word distribution based on the Dirichlet distribution by using another hyper parameter $\boldsymbol{\beta}$. The used symbols are as follows:

$\boldsymbol{\alpha}$ is the parameter of the Dirichlet prior on the per-document topic distributions,

$\boldsymbol{\beta}$ is the parameter of the Dirichlet prior on the per-topic word distribution,

θ_i is the topic distribution for document i,

ϕ_k is the word distribution for topic k,

z_{ij} is the topic for the j^{th} word in document i, and

w_{ij} is the specific word.

The w_{ij} are the only observable variables. The other variables are latent variables. The ϕ is a Markov matrix of which size is $K \times V$ (V is the dimension of the vocabulary). Each row denotes the word distribution of a topic. The LDA generative process for a corpus $\boldsymbol{D}$ consist of M documents each of length N_i, where K denotes the number of topics: as follows:

1. Choose $\theta_i \sim \mathrm{Dir}(\boldsymbol{\alpha})$, where $i \in \{1,..., M\}$ and $\mathrm{Dir}(\boldsymbol{\alpha})$ is the Dirichlet distribution for parameter $\boldsymbol{\alpha}$
2. Choose $\phi_k \sim \mathrm{Dir}(\boldsymbol{\beta})$, where $k \in \{1,..., K\}$
3. For each of the word positions i, j, where $j \in \{1,..., N_i\}$, and $i \in \{1,..., M\}$
   (a) Choose a topic $z_{ij} \sim$ Multinominal(θ_i).
   (b) Choose a word $w_{ij} \sim$ Multinominal(ϕ_{zij}).

The multinominal and Dirichlet distributions are defined in machine learning textbooks. We want to obtain an estimate of $\boldsymbol{Z}$ that gives high probability to the words that appear in the corpus. z_{ij} represents the topic for the j^{th} word in document i. This problems becomes a maximum posteriori estimation of $P(\boldsymbol{W}, \boldsymbol{Z}, \boldsymbol{\Theta}, \boldsymbol{\Phi}|\boldsymbol{\alpha}, \boldsymbol{\beta})$. By an integration concerning $\boldsymbol{\theta}$ and $\boldsymbol{\phi}$, the expression becomes a simple one, $\mathrm{P}(\boldsymbol{W}, \boldsymbol{Z}|\boldsymbol{\alpha}, \boldsymbol{\beta})$. Therefore, we want to obtain $\boldsymbol{Z}$ so that $\mathrm{P}(\boldsymbol{Z}|\boldsymbol{W}, \boldsymbol{\alpha}, \boldsymbol{\beta})$ is maximum. The $\boldsymbol{W}$ is given data. The cost of the calculation is too high because the estimation space size is the number of topics (K) to the power of the dimension of the vocabulary (V), K^V. Each word has K options independently.

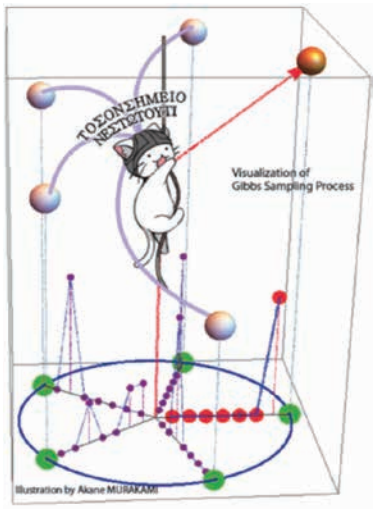

Figure 2. The image of Gibbs sampler concept with the background image developed using Mathematica tools (cited from[12]).

For the LDA program, we used R. The R packaged used is based on "the Comprehensive R Archive Network (CRAN) entitled "lda: Collapsed Gibbs sampling methods for topic models" developed by Jonathan Chang (https://cran.r-project.org/web/packages/lda/index.html).

So instead of that, a random walk search method by Gibbs sampling is widely used[13]. The Gibbs sampling is one method out of Markov chain Monte Carlo methods[10, 14]. The concept image of Gibbs sampling is illustrated in Figure 2 [12, 15]. In this case, the number of documents is five, and the number of topics is seven. In Figure 2, there are five balls on the cylinder edge. Each of the ball corresponds to a document. The height of the ball indicates the topic identification number. On the bottom plane, there is a circle and the five radius lines. On the radius line, the topic distribution probability of each document is illustrated.

The system collects other (n – 1) document data to calculate the topic probability distribution. The Figure 2 illustrates that as a cat wearing a helmet with (n-1) connections to documents. The feature of Gibbs sampling is that the other (n – 1) document data are used to calculate the topic probability density of the target document. From the result, the high probability topic ID is selected. In Figure 2, topic ID 7 is selected for the document. When the topic ID of the target document has been determined, the target document then moved to the topic ID group. That is the classification process. Then, the process is repeated. The next turn will begin on the next target document.

Figure 2 shows the result of our program in Mathematica by Wolfram. We transformed the Mathematica programs to the Wolfram CDF[1] version and published them on the web to be freely access by users (http://www-cc.gakushuin.ac.jp/~20010570/mathABC/SELECTED/). The Wolfram CDF player is a free software. By installing the player, everyone can conduct interactive operations using web browsers. With the teaching materials, the user can interactively operate the sampling by using the slider location of the top page. The reverse motion is also available.

4 Topic Extraction Results

In this section, the topic extraction results by LDA model is presented. We analyzed the LUSI Mudflow reports, paper, and news articles [2, 3, 5-7]. To create the source input file, first we have to remove figures and reference list parts from the documents. The volume of each documents is shown in Table 3:

Table 3: Input volume of each document

Title of the Document	words	Characters
Social and Economic Impacts of the Sidoardjo Mudflow Community Resettlement After Disaster[3]	**7,099**	38,942
The Lapindo mudflow disaster: environmental, infrastructure and economic impact[5]	**4,458**	23,952
Sidoardjo mud flow[2]	**5,001**	26,239
Lapindo Brantas Social Impact Report[6]	**3,841**	20,391
Report into the past, present and future social impacts of lumpur Sidoardjo[7]	**37,737**	197,888

If a document is too long, we divide that file to several text files. Input the source text, and conducted topic extraction by the Latent Dirichlet Allocation (LDA) model and Gibbs samplings. The number of the topics selected was four, because that offers more clear classification than five or six.

We used both unigram and noun-noun bigram segmentations. This is because noun-noun bigrams analysis can prevent lack of connections meaning between words. To make the LDA model, at first, only noun-noun bigrams are extracted from the input files and we count the number of the appearance (See

1) https://www.wolfram.com/cdf-player/

Table 4). However, we could not interpret the latent semantics from the result clearly. Therefore, we took a different approach as follows:

1. We remove the only one time appearance from the result in Table 4.
2. Using the noun-noun bigrams and the appearance, we make a unigram based LDA model. In other words, the word distribution is made of a set of unigram nouns.

Table 5 shows the result of the term distribution of each topic. From the term frequencies, we interpreted the topic. The topic ID in Table 4 and Table 5 has no correspondence. If we cannot interpret the meaning of a term in Table 5, we can refer to a corresponding bigram that includes the noun, so that we can obtain the meaning. For example, for a noun "payment", we found the following bigrams: "assistance payment", "compensation payment", "payment compensation", and "payment claim". Then we can guess that the "payment" might be a payment of compensation to devastated people or districts.

From Table 5, we make the topic titles as follows, i.e. The Lapindo compensation based on the Presidential Regulation. The following topic i to topic u explanations shows our process of analysis to the topic extraction as shown in Table 5.

Topic 1: Lapindo compensation based on the Presidential Regulation

We interpreted the implications of the topic as Lapindo compensation based on Presidential Regulation. The terms of the topic are "lapindo compensation", "community", "regulation", and "Renokenongo Pejarakan". Lapindo is the company name that is used to reter to the mud flow case. These topics are related to the process of compensation from the Lapindo disaster to the declared victims based on Presidential Regulation. Thereby Presidential Regulation refers to the "Presidential Regulation No. 14/2007 on the Sidoardjo Mudflow Settlement Board."[16]. In the official hierarchy of Indonesia legislation, a Presidential Regulation is higher than a Regional Regulation. This Presidential Regulation is a regulation by the president of the Republic of Indonesia to declare the victim area, and the amount of compensation from Lapindo Brantas Inc.

The terms "Renokenongo" and "Pejarakan" refers to the location of the Sidoardjo Mudflow area which is located in Renokenongo Pejarakan village. Therefore, we conclude the topic is related to compensation for the affected village. To explore more about that, we found related compensation words from the bigram noun-noun result, such as "compensation payment", "compensation scheme", "land compensation", "building compensation", "demand compensation", "regulation compensation", and "compensation package". From these words, we think that it is related to the compensation scheme for land and building victims, based on the Presidential Regulation.

Topic 2: Payment of resettlement and relocation

The most frequently appeared word is "Rp" which stands from Rupiah, the currency of Indonesia, IDR. We found in Table 5 that words "payment", "resettlement", "regulation", "relocation" and "Kedungcangkring" appeared many times. Kedungcangkring is one of the affected area in Sidoardjo regency. In Indonesia, both regency and city are at the same administration level. A regency is immediately below a province, and consist of some districts. For the resettlement word, we found "type resettlement", "resettlement scheme", and "cash resettlement" at the noun-noun bigram result. We think

Table 4: Noun-noun bigram distribution of each topic

Topic 1	Freq	Topic 2	Freq	Topic 3	Freq	Topic 4	Freq
payment	66	Porong-river	32	mud-volcano	76	Sidoarjo-mudflow	35
land-building	32	Presidential-Regulation	30	East-Java	65	Indonesia-Year	23
Executing-Agency	25	Gazette-Republic	21	Lapindo-Brantas	62	State-Gazette	22
compensation	25	mud-Porong	19	Republic-Indonesia	53	fs	19
Social-Assistance	21	Number-year	17	Porong-River	34	Place-A	18
Indonesia-Number	21	mudflow-Sidoarjo	14	PT-Lapindo	32	Place-B	17
BPLS-Website	21	purchase-land	13	map-area	27	job-type	17
Besuki-Kedungcangkring	19	Lapindo-Brantas	12	toll-road	20	resettlement-area	14
Kedungcangkring-Pejarakan	19	cost-Rp	12	mudflow-management	19	house-holdincome	14
source-BPLS	19	Area-Map	11	Year-Number	18	mudflow-disaster	12
Rp-month	18	Affected-Area	11	eruption-site	18	source-Author	12
compensation-package	14	volume-mud	11	Lapindo-compensation	17	income-level	12
Rp-family	13	sale-purchase	11	verification-team	15	Author	11
effort-mudflow	13	compensation-property	10	LUSI-mud	14	fs-survey	11
Rp-m2	12	compensation-scheme	10	Rp-metre	12	oil-gas	11
Rp-Rp	12	resident-village	10	village-Siring	12	Renokenongo-village	10
Siring-Jatirejo	11	eruption-zone	10	mud-eruption	11	income-change	10
assistance-Rp	10	value-property	9	Sidoarjo-Mud	11	Banjar-Panji1	10
Assistance-payment	10	housing-estate	9	Head-Executing	11	Sidoarjo-Mudflow	9
president-Republic	10	land-area	9	water-quality	11	et-al	9
payment-Rp	10	compensation-payment	9	LUSI-eruption	10	resettlement-behavior	8
village-Besuki	10	Lapindo-BPLS	9	Regulation-Number	10	significance	8
mud-water	10	Siring	9	Sub-Total	10	mudflow-area	7
refugee-camp	9	resettlement-home	8	compensation-process	10	impact-Sidoarjo	7
instalment-Rp	9	Management-Agency	8	claim	10	income-household	7
Land-Buildings	9	methane-gas	8	impact-mudflow	9	Bother	7
proof-ownership	9	property-value	8	March-map	9	schoolage-child	7
Claims-Rp	9	table-status	8	Brantas-Inc	8	household-head	7
Mitigation-Agency	8	Housing-Estate	8	Bakrie-Group	8	drilling-mud	7
Mud-Mitigation	8	disaster-area	7	Yogyakarta-earthquake	8	Lusi-mud	7
courtesy-BPLS	8	area-village	7	Lapindo-Rp	8	Sidoarjo-regency	6
Total-source	8	National-Team	7	cost-mudflow	7	resettlement-dummy	6
life-insurance	8	Presidential-Decree	7	fault-reactivation	7	survey-estimation	6
month-person	7	compensation-village	7	Siring-Renokenongo	7	number-schoolage	6
percent	7	Jatirejo-Mindi	7	fault-system	6	type-resettlement	6
Government-Regulation	7	Website-table	7	Java-Indonesia	6	resettlement-preference	6
PBP-refugee	7	Mil-Claims	7	flow-mud	6	refugee-area	6
Target	7	Land-Building	7	problem-mudflow	6	mudflow-impact	6
land-ownership	6	rice-field	6	area-December	6	Aburizal-Bakrie	6
Lumpur-Sidoarjo	6	September-resident	6	Agency-article	6	paragraph-paragraph	6
Toll-road	6	Oil-Gas	6	form-assistance	6	Map-March	6
Article-paragraph	6	Perumtas-resident	6	Rail-line	6	Act-No	6
Village-Village	6	Besuki-Pejarakan	6	Regulation-No	6	Jati-rejo	6
H2S-gas	6	RT-RT	6	compensation-claim	6	Brantas-Inc	5
verification-process	6	state-budget	6	West-Siring	6	household-Renokenongo	5
Sosial-December	6	New-Market	6	I	6	business-activity	5
Kegiatan-Deputi	6	Sidoarjo-East	6	Bidang-Sosial	6	villager-relative	5
bubblea-rea	6	crop-failure	6	Kedung-cangkring	6	change-household	5
Jatirejo-Siring	6	Deputi-Bidang	6	payment-claim	6	significance-level	5
Wunut	6	PowerPoint-Presentation	6	mud-sample	6	number-relative	5

Table 5: Noun unigram distribution of each topic

Topic 1	Frequency	Topic 2	Frequency	Topic 3	Frequency	Topic 4	Frequency
Lapindo	550	Rp	401	mud	696	Mil	741
compensation	465	**payment**	354	**BPLS**	397	area	576
eruption	268	**resettlement**	211	village	376	mudflow	481
cost	258	Indonesia	196	resident	308	**Sidoarjo**	449
volcano	188	claim	187	land	296	government	269
number	176	December	143	disaster	285	Porong	246
income	172	Agency	142	**water**	202	**time**	168
year	172	month	133	**gas**	194	Number	150
impact	171	Social	127	assistance	176	**household**	148
property	170	problem	124	**LUSI**	170	**infrastructure**	148
community	169	Republic	112	people	168	management	128
loss	152	building	108	Brantas	166	issue	118
victim	148	March	106	Java	159	Surabaya	112
Total	133	agreement	102	process	158	Besuki	112
Regulation	128	Village	100	East	150	PT	100
company	124	map	94	family	145	day	96
scheme	121	**regulation**	94	table	132	earthquake	96
result	116	refugee	90	**drilling**	118	group	86
level	112	**relocation**	88	**river**	114	River	84
August	112	ownership	78	November	106	change	80
road	108	article	75	report	99	**verification**	80
No	108	**Kedungcangkring**	70	Presidential	94	effect	72
September	108	Year	70	school	88	worker	68
location	106	responsibility	68	metre	86	material	66
house	106	Assistance	68	October	80	effort	62
value	104	Table	66	factory	78	life	60
Jatirejo	102	Mindi	64	paragraph	78	dike	60
business	100	IDR	63	flow	76	concern	54
Renokenongo	96	member	62	district	68	Executing	52
Pejarakan	95	m2	60	bubble	64	asset	48
activity	94	Article	60	source	64	Bakrie	47
well	94	Area	58	President	64	National	46
home	92	provision	58	quality	60	oil	45
Mud	90	program	56	field	58	mitigation	42
July	89	source	56	figure	56	protest	42
team	86	instalment	54	package	56	pond	42
Land	83	purchase	54	Lusi	54	Website	42
job	82	sale	52	health	52	system	40
study	80	person	50	housing	50	facility	40
Place	78	Government	47	form	50	budget	38
site	78	information	46	Gazette	50	date	38
datum	76	Management	45	event	48	al	36
villager	76	Kedungbendo	44	volume	48	status	36
region	74	dispute	42	price	46	action	34
order	72	Affected	42	toll	46	methane	34
type	72	Claims	42	fault	46	Team	34
May	72	Humanitus	40	January	44	Deputy	34
April	72	survey	38	sea	42	distribution	32
Minister	67	progress	38	example	40	service	32
card	66	child	37	line	40	MLJ	32

that it is related to the cash payment scheme based on the type of the resettlement and relocation.

From these words, we think that the topic is related to a resettlement and relocation payment from Lapindo Brantas Inc. for Kedungcangkring village. We conduct the evaluation of Topic 2 as a part of Topic 1. Topic 1 focused just on Presidential Regulation from the standpoint of public legislation. On the other hand, Topic 2 describes the way of the implementation of the law concerning the amount of compensation.

Topic 3: BPLS prevent water pollution from mud spreading

In the word list, as the second mostly found words, there is the word "BPLS." BPLS is the abbreviation of the "Badan Penanggulangan Lumpur Sidoardjo". The English version of BPLS is SMMA (Sidoardjo Mud Mitigation Agency). BPLS/SMMA performs tasks such as handling, controlling, monitoring the mud eruption and its sediments, rescuing people, handling social issues, and handling the relocation of infrastructure.

The most frequently appeared words include "BPLS", "water", "gas", "LUSI", "drilling" and "river". From the noun-noun bigram result in Table 4, we found the terms "BPLS compensation", "water quality", and "water village". From these words, we assume that it is related to the quality of water. Therefore, it may be related to how BPLS could prevent water pollution by the mud spreading. In the residential area surrounding the disaster area, many clean water sources were polluted or damaged by the eruption and mudflow. Subsequently we classified Topic 3 as a topic concerning water pollution by the mud spreading.

Topic 4: Verification of household Infrastructure.

This topic is considered to be related to verification of household infrastructure. The most frequently appearing words are "Sidoardjo", "time", "household", "infrastucture" and "verification". We also found the "process verification" and the "claim verification" from the noun-noun bigram result on the term "verification". The claim means that the verification process was slow because the verification was difficult.

From these words, we think that the topic is related to a verification concerning household and infrastructure of the Mudflow Sidoardjo victims. In order to arrange the compensation, a verification process of the household infrastructure is needed. However, many victims could not show the documents of proof of their former household. As a result, the compensation process had been extremely slow and there is a large discrepancy between the victims' expectation of the compensation and PT Lapindo's willingness to pay[17].

We classify Topic 4 as the implementation related topic on the compensation similar to Topic 2. Topic 1 describes the compensation based on the Presidential Regulation, Topic 2 describes the amount of the payment and Topic 4 is related to the verification process to implement the payment. We can spot the differences between the three compensation related topics.

Among the four topics, there are three compensation related topics and one topic is on the water pollution by the mudflow (Topic 3).

5 Discussion

In this section, we shall discuss the extracted topics and compared them with the results from the East Japan Great Earthquake. Our team had conducted topic extraction to investigate the transition of people' s needs after the East Japan Great Earthquake [18, 19]. Hashimoto et al. found the following 12 kind of their needs:

- T1 : Request for supply of goods
- T2 : Need of Job
- T3 : Request for moving to temporary houses
- T4 : Complaints about governmental responses
- T5 : Need of money support
- T6 : Complaints about transportation
- T7 : Need of new houses (not temporary houses)
- T8 : Complaints about temporary houses
- T9 : Needs of cars
- T10 : Needs of mental care
- T11 : Feel fear about the future
- T12 : Hope to live with families

Concerning the requirements of the victims, Hashimoto et al. recognized the time series changes as follows[18]:

- T1 (Request for supply of goods) and T2 (Need of Job) are basic topics (needs) for afflicted people. They appear for long periods of time.
- T3 (Request for moving to temporary houses), T4 (Complaints about governmental responses) and T5 (Need of money support) appeared in the early period, because these requirements were directly related to support the afflicted people.
- As time passed, people's needs gradually changed to T7 (Need of new houses which are not temporary houses), T8 (Complaints about temporary houses), T9 (Needs of cars), T10 (Needs of mental care), T11 (Feel fear about the future) and T12 (Hope to live with families).

When we spot the difference between the two disasters, the most noted finding is the lack of information of the Sidoardjo afflicted people's needs. The East Japan Great Earthquake happened in 2011 and the Sidoardjo disaster happened in 2006. The date difference must have made differences of Social Networking Service (SNS) functions. In [18], Hashimoto et al. used as a source text, a blog about the afflicted people's needs provided by a non-profit organization in Tohoku, Japan. SNS development enabled us to collect their needs. On the other hand, because there was no SNS facilities in 2006, we think that it was so difficult for the Sidoardjo afflicted people to communicate their needs to others.

Another reason of the difficulties would be that Sidoardjo disaster's enormous scale. When a disaster

scale is too large, it becomes difficult to respond to individuals' different needs and interests. For example, the report[6] says "Lapindo Brantas also provided ancillary social assistance payments to the affected families and individuals.," which included the following payments:

- Monthly salary to unemployed laborers of Rp 700,000 per month per person;
- 2-year house lease assistance of Rp 5,000,000 per family;
- Provision of food (3 times/day) at shelter location at a cost of Rp 15,000- 20,000 per person."

They are a part of the original sentences in[6]. The exchange rate is approximately 100 Rp equals to 1 JPY.

We found the fact that the compensation payment is not sufficient for persons or family expenses. The standard living expenses for Sidoardjo area by Indonesia's government is Rp 3050,000 per month the disaster scale was so large and the verification of the former life became so difficult to acquire. In addition, the land ownership records were in many cases inadequate, incorrect or lost [6]. The confusion level was larger in the Sidoardjo disaster than that of the East Japan Great Earthquake. In such a turbulence, we would not be able to collect individual different needs in a right way to comply with their needs.

6 Conclusion

We analyzed the Sidoardjo Mudflow to find the impact of the mudflow. We have achieved the goal to analyze the impact of the Sidoardjo Mudflow from a neutral viewpoint using the topic extraction method. We found that the text mining is very effective when we find an unexpected fact, such as Presidential Regulation and water pollution from the results. In analyzing the impacts of the Sidoardjo Mudflow using text mining technology we particularly employed the Latent Dirichlet Allocation model. To handle the difficult expressions and grasp the points, we use various techniques such as bigram segmentation. Bigram segmentation is used for finding connection between each words from unigram result. We found that we gained a more rich meaning than using only unigrams for the current target of English documents found on literature on the Sidoardjo Mudflow. As the morphological analysis tool, we used TreeTagger. The extracted topics clearly showed the impact of the Sidoardjo Mudflow. We could find the following topics, i.e. the Lapindo compensation based on the Presidential Regulation, payment of the resettlement and relocation, BPLS prevents water pollution from mud spreading, and the verification of household infrastructure.

Subsequently we compared the results with previous text analysis on other disaster in Japan. We compared the results with our team's analysis concerning the afflicted people's individual requirements. We noticed that the Sidoardjo Mudflow victim must have had different needs and requirement to survive from the disaster and lost they experienced during the difficult situation. However, detailed investigation on their needs could not be conducted because the disaster scale was so large and that the turbulence level was very high.

From the humanitarian standpoint, we think that we have to continue to think on how information technology can contribute to the afflicted people. One answer would be to collect their real needs to survive and continue their lives via information from Social Networking Service (SNS). The future work

could involve recording the current status of the land ownership based on location based services such as maps with different layers and services.

Acknowledgement

This research was partly supported by the grant of 2015-2016 research project at the Telecommunications Advancement Foundation and by the Global Exchange Organization for Research and Education (GEORE) of Gakushuin University. In addition, we thank Ms Akane Murakami for her beautiful scientific illustration about Gibbs sampling that helps our research.

References

[1] Kure, S., et al., *Effects of Mud Flows from The LUSI Mud Volcano from The Porong River Estuary, Indonesia*. Journal of Coastal Research, 2013. 70 (Special Issue): p. 568-573.

[2] Wikipedia. *Sidoardjo mud flow*. Available from: https://en.wikipedia.o-rg/wiki/Sidoardjo_mud_flow.

[3] Putro, P.B.S., *Social and Economic Impacts Of The Sidoardjo Mudflow: Community Resettlement After Disaster*, 2012, Master Thesis. Graduate School of Agricultural Science. Tohoku University. Japan.

[4] Schmid, H. *TreeTagger - a part-of-speech tagger for many languages*. Available from: http://www.cis.uni-muenchen.de/~schmid/tools/TreeTagger/er/.

[5] McMichael, H., *The Lapindo mudflow disaster: environmental, infrastructure and economic impact*. Bulletin of Indonesian Economic Studies, 2009. 45(1): p. 73-83.

[6] Brantas, L. *Lapindo Brantas Social Impact Report - Revised April 2014*. April 2014; Available from: http://www.hsf.humanitus.org/library/social-impact-research/reports/.

[7] Richards, J.R. *Report into the past, present and future social impacts of lumpur Sidoardjo (Humanitus Sidoardjo Fund)*. March 2011; Available from: http://www.hsf.humanitus.org/library/social-impact-research/reports/.

[8] Ratnatunga, J. and A. Sopanah, *Disaster Financing: A Contingent Valuation Approach*. Journal of Applied Management Accounting Research, 2015. 13(2).

[9] Badan Pemeriksa Keuangan Republik Indonesia (The Audit Board of The Republic Indonesia), *Laporan Pemeriksaan Atas Penanganan Semburan Lumpur Panas Sidoardjo (The title : Auditing Report on Hot Mud Sidoardjo Handling)*. 2007; Available from: http://www.environmental-auditing.org/portals/0/auditfiles/indone-sia_s_hotmud.pdf (written in bahasa which is the national language of the Republic of Indonesia).

[10] Bishop, C.M., *Pattern Recognition and Machine Learning*. 2006: Springer.

[11].Koller, D. and N. Friedman, *Probabilistic Graphical Models: Principles and Techniques*. 2009: The MIT Press.

[12] Shirota, Y., T. Hashimoto, and B. Chakraborty, *Visual Materials to Teach Gibbs Sampler.* International Journal of Knowledge Engineering, 2016. 2(2 & 3): p. 92-95.

[13] Griffiths, T.L. and M. Steyvers, *"Finding scientific topics*. Proceedings of the National Academy of Sciences, 2004. 101 (Suppl. 1): p. 5228–5235.

[14] Rubinstein, R.Y. and D.P. Kroese, *Simulation and the Monte Carlo Method*. second edition ed. 2009: John Wiley & Sons.

[15] Shirota, Y., T. Hashimoto, and B. Chakraborty, *Deductive Reasoning for Joint Distribution Probability in*

Simple Topic Model. Proc. of IIAI International Congress on Advanced Applied Informatics 2016, 10-14 July, 2016, Kumamoto, Japan, 2016,. (in printing).

[16] The President Of The Republic Of Indonesia. The Sidoardjo Mudflow Settlement Board (Presidential Regulation No. 14/2007 Dated April 8, 2007) (*Presidential Regulation No. 14/2007 Dated April 8, 2007*). Available from: faolex.fao.org/docs/pdf/ins74997.pdf.

[17] Indra Harsaputra (The Jakarta Post), *Lapindo Mudflow Victims Still Waiting for the Payment (May 30, 2013)*. Available from: http://www.thejakart-apost.com/news/2013/05/30/lapindo-mudflow-victims-still-waiting-paymen-t.html.

[18] Hashimoto, T., et al. *Temporal Awareness of Needs after East Japan Great Earthquake using Latent Semantic Analysis*. in *EJC*, 200-212, 2013.

[19] Hashimoto, T., et al. *Affected people's needs detection after the East Japan Great Earthquake—Time series analysis using LDA*. in *Signal and Information Processing Association Annual Summit and Conference (APSIPA), 2014 Asia-Pacific*, IEEE, 2014.