Crisis Communication in Major Disaster Using Natural Language Processing

白　承志

SeungJi Baek

2016

Crisis Communication in Major Disaster Using Natural Language Processing

by

SeungJi Baek

B.Sc. in Engineering, 2011
M.Sc. in Engineering, 2013

A dissertation submitted in partial satisfaction of the
requirements for the degree of
Doctor of Engineering

in the

GRADUATE SCHOOL OF ENGINEERING
DEPARTMENT OF URBAN MANAGEMENT
of
KYOTO UNIVERSITY

Crisis Communication in Major Disaster Using Natural Language Processing

Abstract

Preparing for catastrophes that may happen in the future is an important issue in risk management. In this research we investigates crisis communication in disaster to clarify the nature of crisis communication between people who are in various position in the society.

With advance of information and communications technology, new type of communication was emerged which is process of exchanging information and opinions regarding the crisis in disaster situation. The online communication using social media make a wide range of communications possible globally and regionally. Considerable attention has been given to the social media's role in reconstructing and strengthening cooperation at various levels and in reinforcing the real-world networks, which deteriorated following the disaster.

The crisis communication contains objective information based on the facts regarding disasters and accidents as well as subjective assessment and perception of the public toward disasters. Since communication in crisis is an significant element in management of crisis, communication during major disasters have to be deliver appropriately, otherwise it would create societal instability eventually.

From a perspective not limited only to victims but inclusive of society as a whole, it is important to investigate peoples reactions to crises and the risk perception they have developed through their experience of the crisis.

In this research, we attempt to examine the contents of actual communication. We are aiming for clarifying an underlying sentiment of the public in extreme situations, since it is essential factor to leading people to organize a reasonable process for disseminating and receiving information. We examine tweets comprising information related to risks caused by disasters and how they transmitted in the Twitter. And based on the discussion that it is essential to investigate sentimental elements of the communication, we focus on latent concerns of the public changing along with state of disaster. Lastly, we suggest the application to measure anxiety as one of the index to measure sentiment of the public in disaster. In this research, we apply methodologies in the field of Natural language processing.

This dissertation is dedicated to my family.

Contents

List of Figures

List of Tables

Acknowledgments

During my 3 years tenure in doctoral courses at Kyoto University, this acknowledgement is the chance to express my appreciation to all those who contributed to make this dissertation possible.

I would like to express the deepest appreciation to my supervisor, Professor Kiyoshi Kobayashi who provided me the opportunity to research in his laboratory. He showed the sincere attitude toward research by giving me advice and comments during discussions in all the seminars. He will always be my inspiration in which he showed endless passion for quest for knowledge and conduct research.

I also want to express my gratitude to Associate Professor Kakuya Matsushima, who guided me from second year of my doctoral course. All of his advice and comments during research meeting and seminar made me accomplish my doctor degree, and his enthusiasm for educating widen my intellectual horizon.

I wish to thank, Assistant Professor Masamitsu Onishi who helped me in research work as well as the life in the laboratory. Without his careful concern and understanding, especially regarding scholarship and all the financial things, this dissertation would not have been possible.

I also wish to thank, Dr. Hayeong Jeong who guide me during first year of my doctor course and encourage me to enter the course. She helped me to determine the topic of my dissertation.

I would like to thank my dissertation defense committee member, Prof. Fujii for his comments and insights.

I also would like to thank to Professor Shun Shiramatsu who helped me a lot especially for research methods with kind advice and comments.

I thank secretary Ms. Inoue and Ms. Fujimoto providing many kinds of help with much kindness and consideration.

I am also very indebted to my many colleagues who supported me in lab. I consider it a fortune and priceless experience in my life to study and discuss with all of them. I thank all of my senior, Dr. Abe, Dr. Segi, Dr. Li, Dr. Hiep, Dr. Nakano, Dr. Gunawan, Dr. Parisa, Dr. Shin, Dr. Zhang, and my lab member who share time with me in the lab, Mr. Fernando, Mr. Kobayashi, Ms. Doyle, Ms. Phuong, Mr. Mizuta, Ms. Zhou, Mr. Yamato, Mr. Han, Ms. Kakihara, Mr. Waki, Mr. Nagaya, Mr. Ito, Ms. Pan, Mr. Wu, Mr. Oga, Mr. Bian, Mr. Park, Mr. Yamashita. Last but not the least, Mr. Ruampongpattana my friend in the same journey of doctoral course.

I also thank Epson International Scholarship Foundation for providing me a scholarship that support my study and my life for last two years in doctoral course.

Finally, I thank my family for their sincere support and patience during my 9 and half year studying abroad in Japan and Jesus Christ, my savior, who always make a way for me.

Chapter 1

Introduction

1.1 Backgrounds of Research

In a disaster, many people participate in crisis communication, including those in various positions in public and private organizations, as well as victims and non-victims of the disaster. Mostly, they interchange ideas about potential risks and at the same time estimate their own capability to deal with them. With the rapid advancement of mobile communication technologies, it has become easy to share ideas and feelings with anyone, anywhere, and anytime by means of portable electronic devices. This has led people to communicate globally and regionally. In fact, in the aftermath of Great East Japan Earthquake, communication via social media and the Internet immediately emerged through the use of portable devices. The Great East Japan Earthquake of March 11, 2011, had a serious impact not only on the region directly affected by the earthquake but also on the entire eastern part of Japan. Since social media such as Twitter played a role during the disaster as a means of communication, considerable attention has been devoted to the effect of social media in reconstructing cooperation at various levels of communication and in reinforcing real-world networks. Because of these new media, information sharing during the Great East Japan Earthquake differed significantly from what it had been in previous disasters. In the 2011 disaster, information about the accident at the regional or personal level became available on a global scale immediately after the earthquake.

However, the rapid spread of information does not always have a helpful effect on communication, since fabricated information can also be circulated (i.e., lies and hoaxes).

While mobile technology can provide tools for intensifying the communication of useful information, unfortunately it can also intensify communication that may disrupt social cooperation in a disaster situation and lead to societal instability.

To deliver crisis communication properly, it is necessary to take proper measures for disaster control. From a perspective not limited only to victims but inclusive of society as a whole, it is important to investigate peoples reactions to crises and the risk perception they have developed through their experience of the crisis. In this study, anxiety is noted as an important emotional sensation in crisis communication, because anxiety, an underlying sentiment of the public in extreme situations, is a factor essential to leading people to organize a reasonable process for disseminating and receiving information. In investigating the contents of crisis communication, it is important to clarify how anxiety was diffused during the crisis. To clarify, it is necessary that anxiety be measured as a figure (numerically). This study proposes a methodology and an application for measuring anxiety, in order to analyze crisis communication containing contents related to risk perception posted on Twitter during the Great East Japan Earthquake.

1.2 Objectives of the Research

In this research we are dealing with crisis communication in disaster situation. To accomplished appropriate communication between the public and organization such as government or media, it is necessary to conduct investigation with the actual crisis communication that have actively taken place on the internet by the introduction of social media and advanced electronic communication device. Our main objectives are:

1. investigating how the tweets containing disaster risks were spread and the contents of information distribution by organization such as government agencies and how the information transmitted throughout Twitter.

2. investigating concern of the public and how the concern changed by time and situation transition.

3. developing an anxiety index to measure anxiety which is achieved by estimating the risk perception of Twitter users. And examining the feature of 'anxiety' of the

public which emerge with societal catastrophe.

1.3 Contribution of the Research

Crisis communication which is using social media as a new communication platform and utilizing advanced techniques of portable communication device was newly appeared, though there had been crisis communication before which is done locally in each area. The new type of communication is not only reinforce existing communication, but play a important role that had been a place to exchange and transmit risk information regardless of sites where the people are. Consequentially, this leads crisis communication to macro level, which people in non-disaster area could easily get into the communication being discussed in disaster areas. This research attempt to investigate crisis communication in macro level which necessary to deal with this age of advancements in information technology.

The research attempts to be one contribution of application corpus linguistics and natural language processing into field of crisis and risk management. by investigating the contents of an actual communication. The research suggests some considerable implications for crisis and risk management, especially policy-making when government agency who need to conduct effective communication by providing appropriate information in proper time to the public.

1.4 Structure of the Research

The structure of the research is organized as follows (Figure 1.1). Chapter 2 explains crisis communication in Twitter which is newly emerged with advanced communication technologies and its corpus based on corpus linguistics. For investigating sentiment of the public using crisis communication corpus, we carry out literature reviews on sentiment analysis and its methodologies.

Chapter 3 shows examination by means of actual communication data of the Great East Japan Earthquake. By estimating the term frequency and retweets information of Twitter corpus, firstly, it describes information transmission regarding risks in disaster

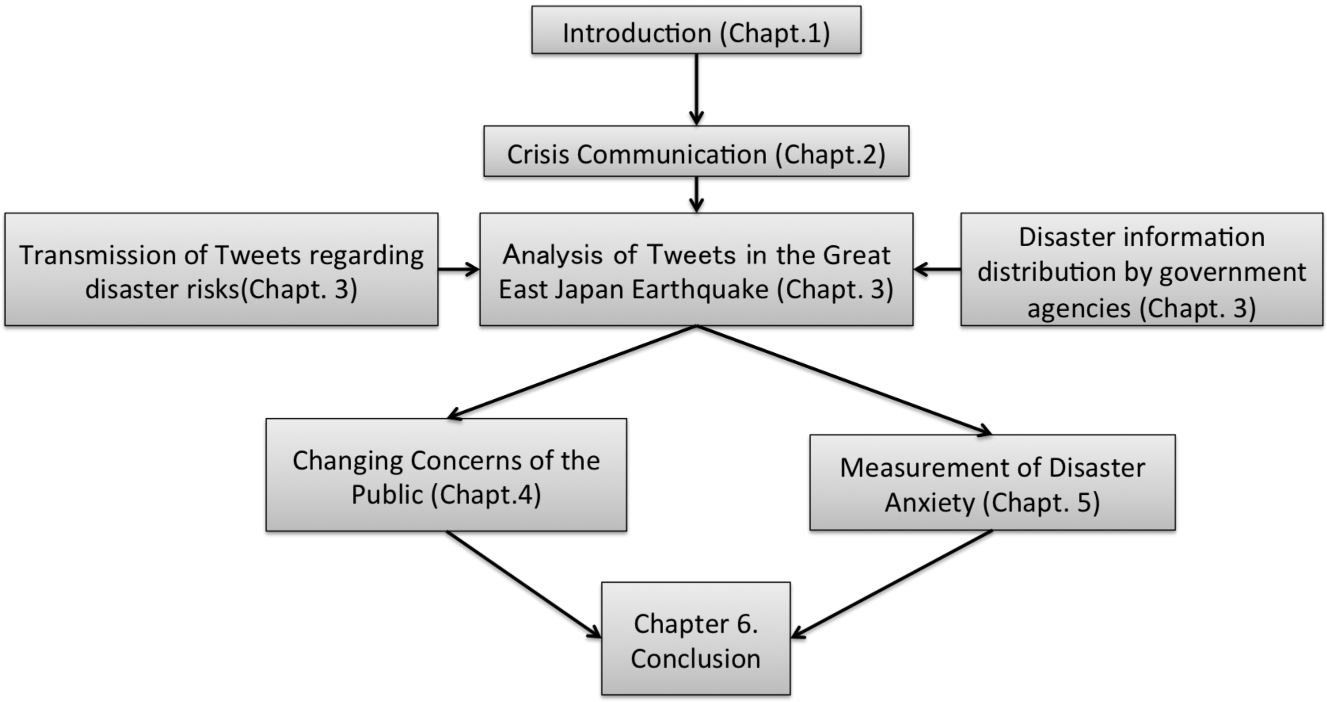

Figure 1.1: Research Process and Framework

to identify elements of risks that the public encounter, caused by series of disasters and accidents. Secondly, it explains disaster information distribution by government agencies through Twitter to examine the effectiveness of using Twitter as a information distribution tool.

After we understand how and why the information transmission occur in Twitter, Chapter 4 analyzes how the the concerns of the public changes along the disaster situation to identify latent risk perception of the public. The aim of this analysis is to clarify what the public really concern about along time changing. Topic model LDA (Latent Dirichlet Allocation)[Blei et al., 2003] is applied to extract the concerns of the public.

After we see the latent concern of the public, In Chapter 5, we focus on one of crucial underlying emotion 'anxiety' for understanding crisis communication. we presents measurements of anxiety caused by the disaster to investigate the latent emotional sensation of the public in extreme situations. As anxiety, an underlying sentiment of the public in extreme situation, it is a essential factor to lead people to organize a reasonable process for disseminating and receiving information [Stieglitz et al., 2013][Oh et al., 2010]. In this Chapter, we propose Anxiety Index to clarify the risk perception based on computational linguistics.

Chapter 6 concludes this research and proposes some potential future research topics.

Bibliography

[1] David M Blei, Andreq Y Ng, and Michael I Jordan. Latent Dirichlet Allocation. *Journal of Machine Learning Research*, pp. 993-1022, 2003.

[2] Onook Oh, Kyounghee Hazel Kwon, and H. Raghav Rao. An Exploration of Social Media in Extreme Events : Rumors Theory and Twitter During the Haiti Earthquake 2010. *Thirsty First International Conference on Information Systems, St. Louis*, pp. 231, 2010.

[3] Stefan Stieglitz and Linh Dang-Xuan. Emotions and Information Diffusion in Social Me- diaSentiment of Microblogs and Sharing Behavior. *Journal of Management Information Systems*, 29(4), pp. 217-248, 2013.

Chapter 2

Crisis communication and Methodologies for Measuring Sentiment of the Public

2.1 Crisis Communication in Disasters

Communication in crisis is an significant element in management of crisis. In a case that organizations such as government agency communicate with individuals in poor way during crises, it often make bad situations worse [Marra, 1998].

In discussions of crises such as natural disasters, topics usually include risk communication, which is similar to the concept of crisis communication. While risk communication is a sort of thought movement reflecting the democratic values of the time, crisis communication does not encompass values but focuses solely on issues regarding the strategic skills necessary for bringing about appropriate communication [Kikkawa, 2000]. By carrying out risk communication on a daily basis, society can function and organization can be upheld, but once a crisis occurs, these tend to fall apart. In a crisis, therefore, crisis communication is activated and communication must be accomplished strategically to reduce the damage as much as possible [Kikkawa, 2000]. According to Coombs [Coombs, 2014] and Kikkawa [Kikkawa, 2000], crisis communication is defined as series of communications that take place before (pre-crisis), in the middle of (crisis event), and after (post-crisis) a crisis, which could seriously threaten the security of a whole society, including organizations or social activities, while risk communication mainly fulfills only the pre-crisis function. Kikkawa [Kikkawa, 2000] pointed out that especially in the middle of a crisis, strategic communication, which minimizes the scale of damage directly caused by the disaster, is necessary. Specifically, it is desirable for communication to be prompt and for all information to be disclosed. Furthermore, information needs to be transferred through several different channels and explained redundantly. If communication is not carried out properly and in a timely manner, the crisis may cause cross-societal panic. For instance, a lack of information could lead to the proliferation of rumors or fabricated information.

Crisis communication research has been carried out mainly from the perspective of the organization, focusing on damage to organizations reputations or legitimacy face to face stakeholders reactions in a structuralist context [Schultz et al., 2011]. However, in a disaster, communication takes place mainly between organizations such as the government and the public through such traditional media as newspapers and radio and television broadcasts, or, more recently, through blogs or social networking services. Since these new types of news media have changed the traditional structure of communication, Schultz et al.[Schultz et al., 2011] pointed out that more multiple perspectives are needed to overcome the organization-centered communication model. They analyze the effects on reputation of different crisis communication strategies via different traditional and social media. During a disaster, it is necessary for the process of exchanging information between the government and public to encourage society to restore order out of chaos.

In this study, crisis communication mainly refers to communication whereby the government and the public mutually send information about their respective statuses, mainly in the middle of a disaster.

2.2 Corpus and Corpus Linguistics for Analyzing Contents of Crisis Communication

2.2.1 Corpus and Corpus Linguistics

To enhance the effectiveness of crisis communication, methodology to investigate public sentiment is proposed in this study. Public sentiment was analyzed by means of Corpus based on Corpus linguistics.

Corpus linguistics, like all linguistics, is concerned primarily with the description and explanation of the nature, structure and use of language and languages and with particular matters such as language acquisition, variation and change. But corpus linguistics nevertheless has a tendency to focus on lexis and lexical grammar of languages in use through corpora[kennedy, 1998]. A corpus is a large sample of how people have used language. And it is a reconstructive method for analysis of language data using a computer [Jeong et al.,2008]. Bennet [Bennett, 2010] pointed out that corpus linguistics serves to answer two fundamental research questions:

- What particular patterns are associated with lexical or grammatical features?
- How do these patterns differ within varieties and registers?

More specifically, investigating corpora (singular:corpus) provides answers to questions like below

- What are the most frequent words and phrases in English?
- What tenses do people use most frequently?
- Which words are used in more formal situations and which are used in more informal ones?

According to Stubbs [Stubbs, 2002], the meaning of words depends on how they are combined into phrases, and on how they are used in social situations. In other words, their meaning depends on both linguistic conventions and inferences from real-world knowledge. And the main evidence for these constraints, linguistic and social expectations, comes from observations of what is frequently said, and this can be observed, with computational help, in large text collections.

In this study, contents of crisis communication are investigated by means of corpus which was created during a disaster.

In urban management research fields, there are few studies that utilize a corpus related to investigate the contents of public debate. For examining both structural analaysis and contents analysis of public debate, Jeong et al. [Jeong et al., 2008][Jeong et al., 2007] developed a computational method to analyze and visualize the semantic similarities of utterances between participants involved in public debates for accurate understanding of the contents and structure.

2.2.2 Crisis Communication using Twitter

As communication technologies advanced a great deal, especially internet enable public to connect each other without media, the way of how people communicate and obtain source to assess a situation during disasters has changed in recent years. Coombs [Coombs2014] argues that these advances make the transmission of communication easier and faster. And also they make the world more visible and crises are easy to be revealed. So it is difficult that crises are isolated from rest of the world. In recent years, it has been possible to communicate easily with anyone, anywhere, and anytime by means of portable electronic devices and the rapid advancement of mobile communication technologies and social media. Social media is one of communication tools that attract great attention these days. It is a broad term that overs a variety of different online communication tools and dominated by user-created content [Coombs2014]. Crises all over the world are continuously showing that citizens, traditional media (newspapers or TV), and organization(government agency) use social media such as Twitter, Facebook and YouTube extensively to express their feelings and to share opinions and information [Terpstra et al., 2012].

In this study, the corpus of Twitter data has been highlighted as a representative social media among all of them. Since Twitter played a major role in the aftermath of the Great East Japan Earthquake as a new type of communication tool between people who are in disaster area and the rest of the society. Twitter is responsible for a large stream exchanging subjective information between its users regarding disaster risks in crisis. And the most distinguishable feature from previous crisis communication is that, it contains the individual context of communication.

For clarification, Twitter is a form of social media that allows its users to send short messages (140 characters or less) to others [Kireyev et al., 2009]. Its user can decide that the contents are visible to only a limited group of users or in general. It is a new type of chat service based on real-time platforms, and strict sorting of the vast amount of

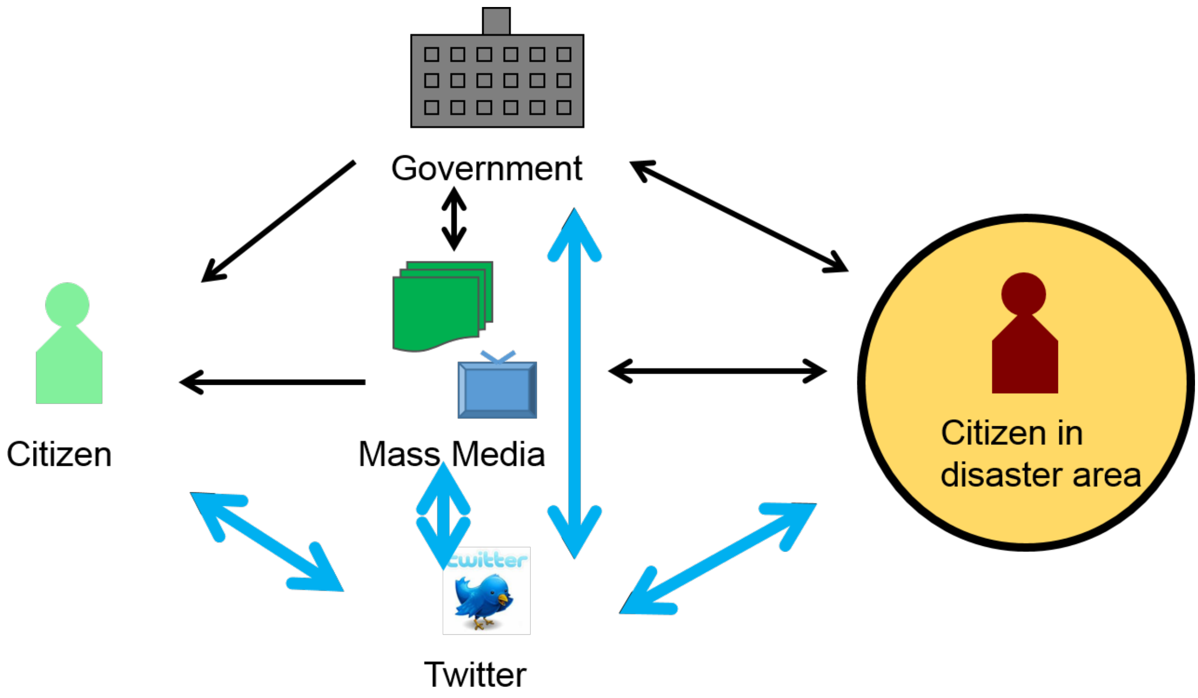

Figure 2.1: Crisis Communication in Disaster using Twitter

information produced based on the social relationships of users is not required. Starbird et al.[Starbird et al., 2010] pointed out that unlike Wikipedia, content passed through Twitter is short-lived; therefore, it cannot be discussed, verified, or edited. While most social media have places for interaction, interaction on Twitter occurs in and on the data itself, and through its distribution, manipulation, and redistribution. Information is part of a life cycle of generation, derivation, synthesis, and innovation that combines skills with information production to shape the information space. Because of the unique characteristics of information interaction, information diffusion is determined by its users who decide what is valuable and what is not.

There are researches that examine the communication which is newly appeared especially in natural disasters (Figure 2.1). Several case studies have been conducted that highlight unique characteristics of interaction such as stream of information or user's behavior in social media. Starbird et al.[Starbird et al., 2010] examined Twitter activity over a concentrated period, where stable elements of geography and features of the hazards threat may be connected to Twitter communications. They examined computer-mediated communication that took place during the flooding of the Red River Valley in the US and

Canada in March and April 2009. They show that Twitter user's who are in disaster areas, tend to tweet more about flood-related issues. However, once the river level begins to subside, they return their interest to everyday lives. They emphasized that Twitter users have evolved their own curation mechanisms, a form of bottom-up self-organizing.

For investigating user's behavior focusing on aspects of the derivative information propagation function, Retweet behavior was observed which is reposting same contents that another twitter user published [Starbird and Palan 2010]. This analysis showed that during crisis, for tweets authored by local users and tweets that contain emergency-related search terms, retweets are more likely than non-retweets to be about the event. Focusing on the contents of tweets, it is more likely to be retweeted when it contains information generated by traditional media (organization) especially local media.

Bruns et al.[Bruns et al., 2012] pointed out that crisis information posted on Twitter by organization such as government agencies were retweeted many times, that is messages sent by organization was able to cut through effectively massive stream of communication. The messages that organization sent contained timely and important information and advice for flood victims and other information-seekers. It could be said that if messages contain emergency-related information and sent by organization such as traditional news media, it is easily spread all over the Twitter. So, these analysis suggest that there is significant scope for official agencies to play an great role in providing up-to-date information and coordinating relief and volunteer efforts through social media, alongside their more established emergency management procedures.

On the other hand, it is hard to say that all communication was carried out ideally in crisis communication process because certain information was fabricated (i.e., lies and hoaxes). Oh et al.[Oh et al., 2010] pointed out that despite many advantages, warnings have been raised about the information quality of Twitter. As mentioned, transmission of communication is also faster than before. So, messages containing unconfirmed information also can spread easily and rapidly, which make people feel insecurity, anxiety and ultimately society unstable. From the point of view of considering Twitter as a tool for information distribution in emergencies, it is necessary to understand particular way of transmission of information as well as motivations that people interact through social media.

For investigating crisis communication using Twitter, we can obtain an understanding about the collective "wisdom of crowds"[Surowiecki, 2005] and leverage its data in policymaking, decision support, economic analysis, epidemic behavior (the "tipping points" theorem [Gladwell, 2006]) and various other applications [Cheong and Lee, 2009].

In this research, we are focusing on public response and sentiment that change over time pass during disasters in Twitter.

2.3 Sentiment Analysis for Crisis Communication using Corpus

2.3.1 Reasons for Sentiment Analysis

Emotions are crucial elements to explain actions that occur in society. Barbalet [Barbalet, 2002] argued that "a well-developed appreciaion of emotions is absolutely essential for sociology because no action can occur in a society without emotional involvement". When natural disaster happen which give a fatal impact on society stability, it delivers social division and societal panic, accompanied by collective actions such as stocking up heavily with food or even riots.

According to a study [Baker, 2012]investigating riots recently broke out, They discussed that the new social media facilitates the riots extension using social networking in diverse temporal and spatial boundaries, but, more to the point, social media is not reason of riots. They emphasized that it is a facilitator rather than the underlying cause of riots. Instead, attempts to understand the causes of riots must recognize that emotions play a crucial role in motivating thought and action with the "mediated crowd" which is a nascent social phenomenon that emerges before the evident crowd gathering. The emergence of "mediated crowd" in recent riots, relay on newly emerging social media by means of advanced technology of internet and mobile communication, and this feature make it distinguish from a standard crowd. And it is necessarily emerging from either a common "emotional atmosphere" or a shared "emotional climate"[Baker, 2012][Rivera, 1992]. The "emotional atmosphere" refers to collective and temporary moods or behavior towards a common event that group of society may show, so it is event generated moods

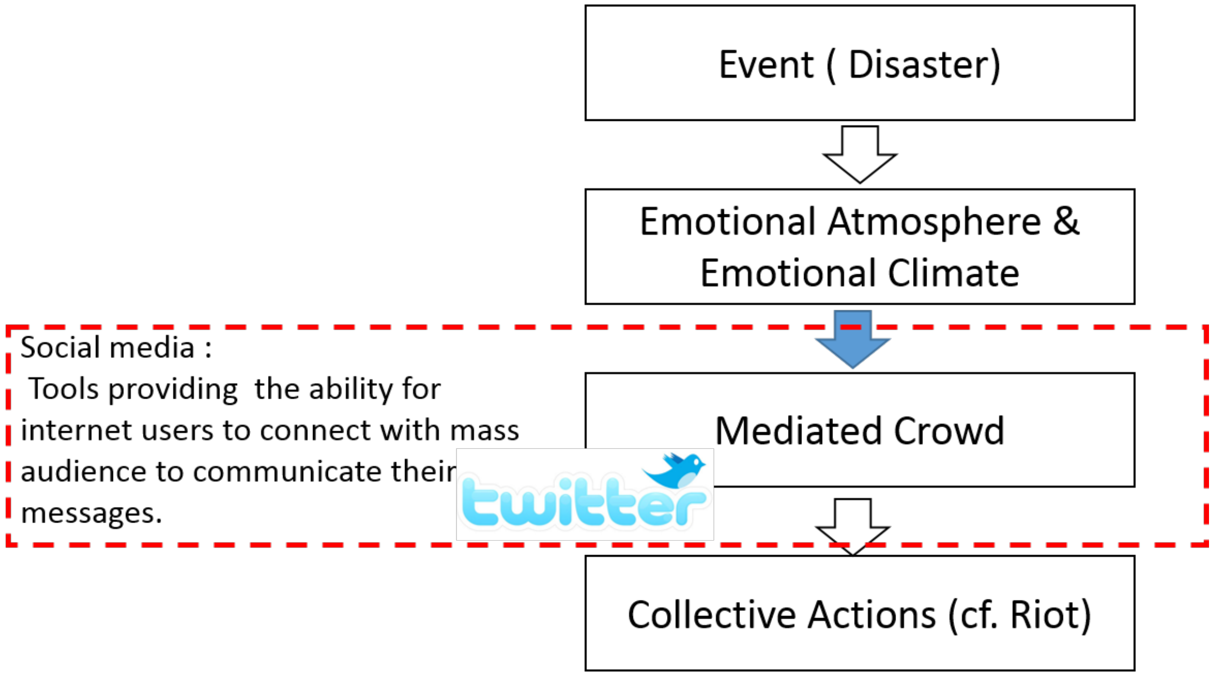

Figure 2.2: The Concept of Formation of Collective Actions

such as the grief arise from sudden death of celebrity or the collect joy ensues from victory of sports events, while "emotional climate" is more lasting than emotional atmosphere and "more pervasive emotional phenomena that are related to underlying social structures and political programs", often be said as names of emotions, such as joy and fear[Baker, 2012][Rivera, 1992]. During a formation of collective behavior such as riots and after it broke out, "mediate crowd" appear which interact each other disregarding the time and place in new social media. In this sense "emotion operate as the intermediary between social structure and agency, with individual and collective action reflecting the agent's evaluation of a given circumstance"[Baker, 2012][Barbalet, 2002] (Figure 2.2). Therefore, investigating the emotion in disaster may give us to understand the cause of social division by unexpected catastrophe.

2.3.2 Utility of Twitter Corpus as a Data for Crisis Communication Analysis

The societal context of risk perception measures during the disaster was examined based on Twitter data from public who addressed their conditions through social media. It contains both cross-societal and local context. With regard to local context, Twitter corpus includes information regarding locations of users and times when users tweet with their mobile communication devices. Twitter provides information valuable in understanding the areas affected by a disaster. Risk information disseminated by the government and news media tends to ignore the local context, as evidenced by the fact that crisis communication geared to the local level had not been observed in prior disasters. Communication by means of social media can overcome this problem.

Another feature of Twitter corpus is that it contains very little lexical redundancy in a single tweet and the distribution of information is rapid [Kireyev et al., 2009]. By investigating the contents of Twitter, it enables researchers to capture the early phases of crisis communication. This feature also encourages the government to utilize for distribution of information via Twitter (i.e., real-time risk information).

Twitter is a social media containing subjective assessment created by individuals, while newspapers and Wikipedia are inter-subjective social media that publish information through cross-validation. Therefore, this study considers Twitter corpus as an important means of information exchange within crisis communication.

2.4 Methodologies for Sentiment Analysis in Disasters

2.4.1 Review of Related Research

In the research field of crisis communication, there have been studies examining public response or reaction during a disaster, in order to demonstrate which factors are correlated to the publics perception of the emergency situation.

Utz et al.[Utz et al., 2013] examined three factors: medium, crisis type, and emotions. They demonstrated that anger, which is chosen as an emotion factor, is related to reputation, public communication, and reaction. In the analysis, they used a questionnaire survey to obtain the reactions of participants about a crisis scenario, in this case the Fukushima nuclear incident. Answers were given on seven-point Likert scales, and a co-relation was seen among factors and reactions. Since it acquired participants response in a direct way, the questionnaire was different from the application method for crisis communication via social media.

Stefan et al.[Stieglitz et al., 2013] focused on the relationship between emotions and information diffusion in social media. They demonstrated that emotionally charged Twitter messages tend to be re-tweeted (information diffusion) more often and more quickly than neutral ones. To determine the level of sentiment strength in a short message, SentiStrength [Thelwall et al., 2010] was applied. This is an algorithm that uses a dictionary of sentiment words with associated strength measures [Thelwall et al., 2010]. Since it was developed through comments from social media (Myspace), the algorithm includes a correcting process for non-standard spellings.

There is little research that focuses on anxiety as a significant factor to explain a pattern of information-distribution in extreme events. Oh et al.[Oh et al., 2010] investigated social media (Twitter) in a disaster scenario, the Haiti Earthquake of 2010. They applied rumor as an important factor for analysis, focusing on emotional statements as a proxy of the anxiety variable. Twitter messages were categorized mainly as emotional statement or authenticating statement. This classification of Twitter messages was done manually. The authors observed that the quantity of emotional statements and authenticating statements changes according to time series, with emotional statements (anxiety) arising rapidly at the early stage of events and authenticating statements, such as new reports, increasing gradually toward the later stage. However, this research did not adopt an automatic classification algorithm. It is difficult to examine the trend of fluctuating emotions in the extremely rapidly expanding volume of social media data in a natural disaster situation.

Baek et al.[Baek et al., 2013] investigated the co-relation between information distributed by organizations and citizens risk perception in a catastrophe. They focused on comparing the anxiety of citizens and organizations appearing in Twitter post.

This study focused on an anxiety index proposed as a sentiment analysis application to accomplish proper crisis communication. Anxiety is utilized as an indicator to estimate risk perception. To evaluate anxiety, the authors proposed an Anxiety Index containing polarity of terms and frequency of terms. The polarity (positive or negative) of each term in the corpus is categorized automatically by a dictionary of term-polarity (semantic orientation). Though the methodology of identifying the topics applied in this research is general in the field of sentiment analysis, output derived from the analysis would give useful information for effective crisis communication, which could open a new window in the public policy strategy dimension of sentiment analysis.

2.4.2 Topic models as a Tools for analyzing Twitter Data

In this research, we propose applications to investigate the Twitter data during disaster. To clarify the contents of the communication, Topic models are applied. Since Topic models are appropriate to examine corpus such as Twitter.

Kiyeyev et al.[Kireyev et al., 2009] examine the use of Topics models for processing Twitter data. The topic models are not originally developed for analyzing Twitter data but probabilistic models for analyzing the semantic contents of large document corpora. They argue that topic models is a particularly promising methods for analysis of disaster-related Twitter data. They raised 4 specific reasons that are Bag-of words, Latent variables, Representation and Adaptability.

Ramage et al.[Ramage et al., 2010] also argue that contents analysis on Twitter are relatively short compared to the standard written language on which many supervised models in machine learning and Natural Language Processing are trained and evaluated. For effective modeling to analyze contents on Twitter data, it is required to adapt method with little supervision.

Text mining techniques that we applied in this study to investigate contents of Twitter data are unsupervised models such as LDA (Latent Dirichlet Allocation) and $TFIDF$.

2.4.3 Topic Model (LDA)

In this study, we apply LDA(Latent Dirichlet Allocation) model to clarify the concerns of the public in disasters. LDA is one of the latent variable topic models that require no manually constructed training data[Ramage et al., 2010]. It is a generative probabilistic model for collections of discrete data such as text corpora developed by Blei et al.[Blei et al., 2003]. With this methodology, we can find topics which is latent variable of data collection. To identify topics in documents, it is assumed that documents is mixtures of probabilistic topics, which the main problem to solve is for discovering the set of topics that are used in a collection of documents [Griffiths and Steyvers, 2004]. For parameter estimation, we apply variational inference algorithm proposed by Blei et al.[Blei et al., 2003].

2.4.4 Topic Model (TFIDF)

To measure Anxiety of the public, we use natural language proceedings techniques with combined methods including Term Frequency Inverse of Document Frequency Implementation ($TFIDF$) and Co-Occurrence Frequency (COF). $TFIDF$ scheme was a topic-extracting model proposed by Salton et al.[Salton et al., 1988]. Terms are weighted with $TFIDF$ scores and it is used to determine the significant keywords in a document. The term frequency of word appearance is used to define the score which could be calculated as below.

$$TFIDF_{w,a} = TF_{w,a} \times IDF_w \tag{2.1}$$

$$IDF_w = log\left(\frac{N}{DF_w}\right) + 1 \tag{2.2}$$

$TF_{w,a}$ = Number of occurrences of term w in a document a

DF_w = Number of documents containing term w

N = Total number of documents

TFIDF score, the total weight of significance $TFIDF_{w,a}$, is defined with term w frequency in document a ($TF_{w,a}$) and inverse of document frequency containing term w (IDF). Term w with high TFIDF represents a significant term in a document a. However, high frequency terms are not necessarily important. So, the IDF is applied to express whether the term appears frequently in other documents as well. So, IDF considers the number of documents that the term occurs (DF_w) and is defines as the logarithm of rate of the total number of documents and DF. A term with high $TFIDF$ score is the term that occurs in a document frequently but not in other documents. Likewise, A term with low TFIDF score is the term that occurs in a document infrequently in a document but appears in many other documents. In our research, we utilized this topic model to extract the keywords to identify the concerns of the public. Among terms with high $TFIDF$ score, HOUSYA (radiation) was chosen as a topic related closely to risks originated from the Fukushima Daiichi nuclear accident.

2.5 Conclusion

In this chapter, we examined the necessity of the sentiment analysis to investigate the crisis communication and availability of Twitter data as a representative social media in the recent disaster. since the new type of news media contains the local context and individuals level of communication which never observed in pre-existing communication, it have changed the traditional structure of communication. According to the several researches related to crisis communication with Twitter, it could have dual-aspect. One is reinforcing the communication between the stakeholders in our society most importantly the public and the government. On the contrary, it also have a aspect of disturbing their interaction. From the point of view of considering Twitter as a tool for information distribution for the public, especially, in emergencies, the 'emotion' turns out that the most crucial factor motivating the public to make the emotional atmosphere which is the stream of gathering and interacting in society such that was shocked by catastrophe.

For investigating the emotion of the public using Twitter data, we explained some unique characteristic of information interaction in Twitter and its adaptability to analysis of crisis communication. We reviewed other researches that developed the methodologies and application tools to demonstrate the co-relation of emotional statements and information diffusion using data collected from social media. As we deal with the text data, we explained topic models such as TFIDF and LDA which is promising method for analysis of disaster-related Twitter data.

Bibliography

[1] Seung Ji Baek, Hayeong Jeong, and Kiyoshi Kobayashi. Disaster anxiety measurement and corpus-based content analysis of crisis communication. *In Proceedings - 2013 IEEE International Conference on Systems, Man, and Cybernetics, SMC 2013*, pp.1789-1794, 2013.

[2] Stephanie Alice Baker. From the criminal crowd to the mediated crowd: The impact of social media on the 2011 English riots. *Safer Communities*, 11(1), pp.40-49, 2012.

[3] Jack Barbalet. Introduction: why emotions are crucial. *The Sociological Review*, 50(S2) pp.1-9, 2002.

[4] Gena R. Bennett. Using Corpora in the Language Learning Classroom: Corpus Linguistics for Teachers Part 1. pp. 4, 2010.

[5] David M Blei, Andreq Y Ng, and Michael I Jordan. Latent Dirichlet Allocation. *Journal of Machine Learning Research*, pp. 993-1022, 2003.

[6] Axel Bruns, Jean Burgess, Kate Crawford, and Frances Shaw. #qldfloods and @QPSMedia: Crisis Communication on Twitter in the 2011 South East Queensland Floods. *Methodology*, (Cci) pp.1-57, 2012.

[7] Marc Cheong and Vincent Lee. Integrating web-based intelligence retrieval and decision- making from the twitter trends knowledge base. *ACM Workshop on Social Web Search and Mining (SWSM)*, pp.18, 2009.

[8] W Timothy Coombs. *Ongoing crisis communication: Planning, managing, and responding*. Sage Publications, 2014.

[9] Malcolm Gladwell. *The tipping point: How little things can make a big difference.* Little, Brown and Company, 2006.

[10] Thomas L Griffiths and Mark Steyvers. Finding scientific topics. *Proceedings of the National Academy of Sciences of the United States of America*, (101) Suppl, pp.5228-5235, 2004.

[11] Hayeong Jeong, Tsuyoshi Hatori, and Kiyoshi Kobayashi. Discourse analysis of public de- bates: A corpus-based approach. *Conference Proceedings - IEEE International Conference on Systems, Man and Cybernetics*, pp. 1782-1793, 2007.

[12] Hayeong Jeong, Shun Shiramatsu, Kiyoshi Kobayashi, and Tsuyoshi Hatori. Discourse analysis of public debates using corpus linguistic methodologies. *Journal of Computers*, 3(8), pp.58-68, 2008.

[13] Graeme Kennedy. An Introduction to Corpus Linguistics. *In Studies in Language and Linguistics*. Citeseer, 1998.

[14] Toshiko Kikkawa. Risk Communication. *Japanese Journal of Health Physics*, 35(4), pp.485-490, 2000.

[15] Kirill Kireyev, Leysia Palen, and Kenneth Anderson. Applications of topics models to analysis of disaster-related twitter data. *NIPS Workshop on Applications for Topic Models: Text and Beyond*, 2009.

[16] Francis J. Marra. Crisis communication plans: Poor predictors of excellent crisis public relations. *Public Relations Review*, 24(4), pp.461-474, 1998.

[17] Onook Oh, Kyounghee Hazel Kwon, and H. Raghav Rao. An Exploration of Social Media in Extreme Events : Rumors Theory and Twitter During the Haiti Earthquake 2010.*Thirsty First International Conference on Information Systems, St. Louis*, pp. 231, 2010.

[18] Joseph De Rivera. Emotional climate: Social structure and emotional dynamics. *In KT Strongman (Eds.), International review of studies on emotions*. Citeseer, 1992.

[19] Term-weighting approaches in automatic text retrieval. *Information processing & management*, 24(5), pp.513-523, 1988.

[20] Friederike Schultz, Sonja Utz, and Anja G oritz. Is the medium the message? Perceptions of and reactions to crisis communication via twitter, blogs and traditional media. *Public Relations Review*, 37(1), pp. 20-27, 2011.

[21] Stefan Stieglitz and Linh Dang-Xuan. Emotions and Information Diffusion in Social Me- diaSentiment of Microblogs and Sharing Behavior. *Journal of Management Information Systems*, 29(4), pp. 217-248, 2013.

[22] Kate Starbird, Leysia Palen, Amanda L. Hughes, and Sarah Vieweg. Chatter on the red: what hazards threat reveals about the social life of microblogged information. *CSCW 10 Proceedings of the 2010 ACM conference on Computer supported cooperative work*, pp. 241-250, 2010.

[23] Kate Starbird and L Palen. Pass it on?: Retweeting in mass emergency. *Proceedings of the 7th International ISCRAM Conference*, pp. 1-10, 2010.

[24] Michael Stubbs. *Words and phrases: Corpus studies of lexical semantics.* Blackwell Publishers Oxford, 2002.

[25] James Surowiecki. The Wisdom of Crowds. *American Journal of Physics*, 75(0908) pp. 336, 2005.

[26] Daniel Ramage, Susan Dumais, and Dan Liebling. Characterizing Microblogs with Topic Models. *Icwsm*, pp.1-8, 2010.

[27] Teun Terpstra, R. Stronkman, a De Vries, and G.L. Paradies. Towards a realtime Twitter analysis during crises for operational crisis management. *Proceedings of ISCRAM 2012*, (April) pp. 1-9, 2012.

[28] Mike Thelwall, Kevan Buckley, Georgios Paltoglou, Di Cai, and Arvid Kappas. Sentiment in short strength detection informal text. *Journal of the American Society for Information Science and Technology*, 6(12), pp.2544-2558, 2010.

[29] Sonja Utz, Friederike Schultz, and Sandra Glocka. Crisis communication online: How medium, crisis type and emotions affected public reactions in the Fukushima Daiichi nuclear disaster. *Public Relations Review*, 39(1), pp.40-46, 2013.

Chapter 3

Analysis of Tweets in Disaster

3.1 Introduction

In the Great East Japan earthquake, disaster information regarding disaster areas, damage condition by earthquake or tsunami, the confirmation of someone's safety or radioactivity quantity was transmitted quickly through Twitter. Social media such as Twitter was utilized as a means of information delivery. Disaster information that produced by government agencies and media such as newspapers, new broadcast was transmitted in Twitter as Twitter users who were sent the messages containing disaster information resent the messages by retweeting (resenting the message to their account followers). Especially, in Twitter, the information was also transferred without going through the traditional media, but government agencies or experts group distributed the disaster information directly to the public. So, information flow was more direct passing through few channel to reach the public, comparing with previous disasters. Both of central and local government released the disaster information by means of Twitter to notify the public who are in local areas and also all over the Japan.

In this chapter, we are focusing on an actual communication occurred in The Great East Japan Earthquake and The Fukushima Daiichi Nuclear Disaster between the public, the government agencies and media. We investigated how the information containing disaster risks were spread and the contents of information distribution by organization such as government agencies and how the information transmitted throughout Twitter.

3.2 The Great East Japan Earthquake and The Fukushima Daiichi Nuclear Disaster

The Great East Japan Earthquake occurred on 11 March, 2011. It was the strongest earthquake ever recorded in Japan. The magnitude that shows a scale of the earthquake was 9.0 and it also generated tsunamis which were over 10 meters high and caused great damage to a wide range of coastal areas of Japan. By the earthquake and tsunami, 18,460 were reported dead and missing and almost 4 hundred thousand building were overly destroyed. After about an hour later the earthquake, the tsunami caused The Fukushima Daiichi nuclear disaster. It produced power supply equipment failure followed with three nuclear meltdowns and releases of radioactive materials beginning on 12 March [Strickland, 2011]. Since the amount of electric supply by The Fukushima Daiichi nuclear power plant occupied great deal for electric provision in east Japan, it affected not only the region struck by the large earthquake but the whole eastern part of Japan and led planned power outage and unstable electricity supply. Additionally, the transportation network suffered severe damage, as many roads were cut off and more than 16,000 people were isolated in the disaster area (Iwate, Miyagi, and Fukushima prefectures). In this analysis, we decided that Iwata, Miyagi and Fukushima prefectures as disaster areas, of which the number of dead and missing is over a thousand. Especially, these three prefectures suffer a severe damages by the earthquake, tsunami and Nuclear accident.

3.3 Tweets as an utterance of the public and its Transmission in Twitter

3.3.1 Data

Twitter data was provided by Twitter Japan via Project 311 (The Great East Japan Earthquake Big Data Workshop Project 311, 2015). It was collected over the seven days from March 11 (about an hour before the earthquake occurred) to March 17, 2011. Tweets posted during those 7 days and also those written in Japanese were included. The data comprises tweet IDs, user IDs, time and tweet contents (Figure 3.1). Every tweet has its own ID, so there is no tweet which has same tweet ID. And each user account have its own ID, so a tweet that a same user tweeted (posted a message) has the same user ID. Figure 3.2 shows the content of the data arranged randomly regardless of time. Table 3.1 shows the number of tweets for each days and the total number of tweets in dataset. The number of tweets was biggest on March 12 and showed decreasing trends after the day.

Figure 3.1: Twitter Data comprising tweet IDs, user IDs, time and tweet contents

Table 3.1: The Number of Tweets

11-Mar	12-Mar	13-Mar	14-Mar	15-Mar	16-Mar	17-Mar	Total
26,099,910	27,315,744	23,065,384	25,404,539	25,338,601	24,368,393	22,465,493	174,058,064

```
46293849374597120SOH66720011SOH2011-03-12 04:38:08SOH新聞屋さんが来るタイミングが神すぎて笑えてきたっていうか笑いが止まらないって言うか
47162379284385792SOH52560393SOH2011-03-14 14:09:22SOH仕事は少しずつだけど始めてます。
47451783513116672SOH96744666SOH2011-03-15 09:19:21SOHていうか、おじいちゃんありがとう。RT @yagi1207 @nanahoshi10 エレベーターも電気だもんね。お
47175889942888448SOH105517314SOH2011-03-14 15:03:03SOHRT @yoko_19900820: @SHUYA_Mizuno 【拡散希望】茨城県北大子町は、すべてのライフラ
47155747670327296SOH168670956SOH2011-03-14 13:43:01SOH今日バイトあるか確認してこなくては！
46421659325964288SOH225819728SOH2011-03-12 13:06:00SOH野球には興味がないです。WBCは見ますが RT @Research_Panel 3月25日に開幕する日本
46172308682969088SOH15413828SOH2011-03-11 20:35:10SOH途中まではバスでいけそう。
46947966577147904SOH188747597SOH2011-03-13 23:57:22SOHRT @hagetoruyanaika: 日本は八つの島からなる国です。日本中が協力して一つになるという意
46000325236228096SOH182428004SOH2011-03-11 09:11:46SOH@yukinkoNoyuki 飯食いにいこうず！
46978929650696192SOH218426802SOH2011-03-14 02:00:24SOHそろそろ寝るの。おやすみー
46947949070135296SOH106405366SOH2011-03-13 23:57:17SOHRT @hiraerika: 停電時の冷蔵庫の扱い@東芝 だそうです http://bit.ly/fgWHUG
47889982509162496SOH173827312SOH2011-03-16 14:20:36SOHRT @fbous1: @kopipedoujou 歌手の松山千春氏がラジオで，今回の地震に関して「知恵が
48299246134362112SOH133675993SOH2011-03-17 17:26:52SOHちょっと揺れたな
47934012043431936SOH102442831SOH2011-03-16 17:15:33SOH「私は命が3つあるんだ」男は高らかに告げた。「多少の無茶もできるし、色んな世界を見r」俺は隠し持って
47720326800478208SOH90947246SOH2011-03-16 03:06:27SOH@ad_reader 確かにそれは感じてました。しかしありがとうございます！
47653902274920448SOH99251369SOH2011-03-15 22:42:30SOH@hukupo 把握しないでください><
47617123857600512SOH110397696SOH2011-03-15 20:16:21SOHとまっちゃんがあまりにもナチュラルに「あきちゃんお肉ちょっとちょうだい」と言ってて吹いた笑 #sphere
47297835800068097SOH163474317SOH2011-03-14 23:07:37SOHたったいま、カナル型ヘッドフォンのLチャンネルがお亡くなりになりました。数ヶ月の間でしたが、お疲れ様でした
46778609133821952SOH132179599SOH2011-03-13 12:44:24SOHRT @iwakicity: RT @riko62: 福島県いわき市小名浜在住のご家族等と連絡とれない方達
46013161555492864SOH83518840SOH2011-03-11 10:02:47SOHRT @TokikoKato: 昔ある某レコード会社のディレクターに，いくら曲を作っていただいても困ります。会
46829926250315777SOH125988896SOH2011-03-13 16:08:19SOHRT @nananoki: 以前阪神大震災の時、電話ボックスに入ろうとしたら突然「走れ！」って聞こえて、
47885106882428928SOH137995902SOH2011-03-16 14:01:13SOH【りよこさん通信】規制されることがあるので@rykp06フォローお願いします。
46120067804561409SOH27812779SOH2011-03-11 17:07:35SOH日本で大地震があったみたいですけど、津波、火事、倒壊の被害が最小限になることを祈ってます。Huge
46853929937551360SOH195320153SOH2011-03-13 17:43:42SOHRT @toyokatsu1985: @takapon_jp 堀江さん。ご協力お願いします。宮城県石巻市宇田
47642416232144896SOH175572340SOH2011-03-15 21:56:51SOHプレッシャーをかけられた仕事が無事ひとくぎり。
48402772495843328SOH115369614SOH2011-03-18 00:18:14SOH@_LuKa__おやすみなさい♪
47029323156299776SOH99208078SOH2011-03-14 05:20:39SOH会社に行けない・・
47366869782503424SOH183550968SOH2011-03-15 03:41:56SOHTokio 03.41am Good night y Arigato!!おやすみ(ｕσω-)｡*ﾟzzzzzzzzzz
47058275178524672SOH111886715SOH2011-03-14 07:15:41SOH悲しんでいる人が、苦しんでいる人がいっぱいいっぱいいるのに、あいりはいつもと変わらない月曜日。連絡取
48514173646282752SOH71818158SOH2011-03-18 07:40:55SOH挨拶するたびトモダチ増えるポポポポポーン RT @mtk0929: ふっくんもう完全に洗脳されたな
47227197043376128SOH126537596SOH2011-03-14 18:26:55SOH@i_squirrel_bell なでなでなで
46186222195965952SOH241952077SOH2011-03-11 21:30:28SOH@cooozooo 私も絶対出ると思ってました。政府の発表の仕方怪しすぎます。
47238275945074688SOH85592570SOH2011-03-14 19:10:57SOHて、ちょwww 黙祷の後のデモ騒乱の切り替えがパネぇwwwww ギャップワロタw
47290837721288704SOH121074554SOH2011-03-14 22:39:48SOHRT @Asahi_Shakai: 【被災地以外の方へ】「自衛隊が支援物資を募集している」というチェーンメ
48386668650168320SOH94495670SOH2011-03-17 23:14:15SOH自分の放送のwww第二回www黒歴史wwwwww http://live.nicovideo.jp/watch/
47913353259139072SOH44841381SOH2011-03-16 15:53:28SOH今日も停電開始。昨日と違って昼だから引き続きお掃除するね。
47611300989255680SOH210813615SOH2011-03-15 19:53:13SOH@OshIno_o しんどかったら遠慮せず休みな。体がいちばんやしな。こっちはまぁ何とかなるYo！今日は
47185878438780928SOH97318991SOH2011-03-14 15:42:44SOHただいま
47077733158756352SOH72236463SOH2011-03-14 08:33:00SOH@kassy1069 うむ。来た。全く影響なしorz...
46581482713657345SOH230690794SOH2011-03-12 23:41:05SOH緊急地震速報のエリアメールが携帯に来た時の着信音が怖い 怖い夢みてる時に来て軽いトラウマになって
47206959241379840SOH182965160SOH2011-03-14 17:06:30SOH@Ayaaa_18 全然いいよ＼(^o^)／言ってくれればつぶやきやす!!!!!!気をつけてね(T_T)
47945259900076032SOH241794757SOH2011-03-16 18:00:15SOHただいま！ためてたプリキュアみる～セイレーンかわいい
```

Figure 3.2: Data provided by The Great East Japan Earthquake Big Data Workshop Project 311

3.3.2 Tweeting by the public Related to the Disaster

After the disaster occurred, People exchanged the information regarding the disaster. As disasters such as radiation(HOUSYA), earthquake(JISHIN) and Tsunami that gave serious impact on the Japanese Society, people recognized those disaster as 'risk' may or may have given damage to their life. So, those terms were included in the tweets a lot when people communicate in disasters. Figure3.3 and table 3.3 shows the time series of the quantity of tweets containing the words. The number of tweets which contain the words earthquake was highest on 11 March and it declined after 12 March. Those of tsunami shows the similar pattern. On the other hands, those of radiation was biggest on 15 March which was second time increasing after 12 March. We could observe that people were more like to mention the risks when it is actually happening or immediately after, since the severe damages caused by earthquake and tsunami were recognized on 11 march right after the first violent shaking of earthquakes, while Fukushima Nuclear power plant accident was facing serious difficulties on 15 March with series of explosion accidents.

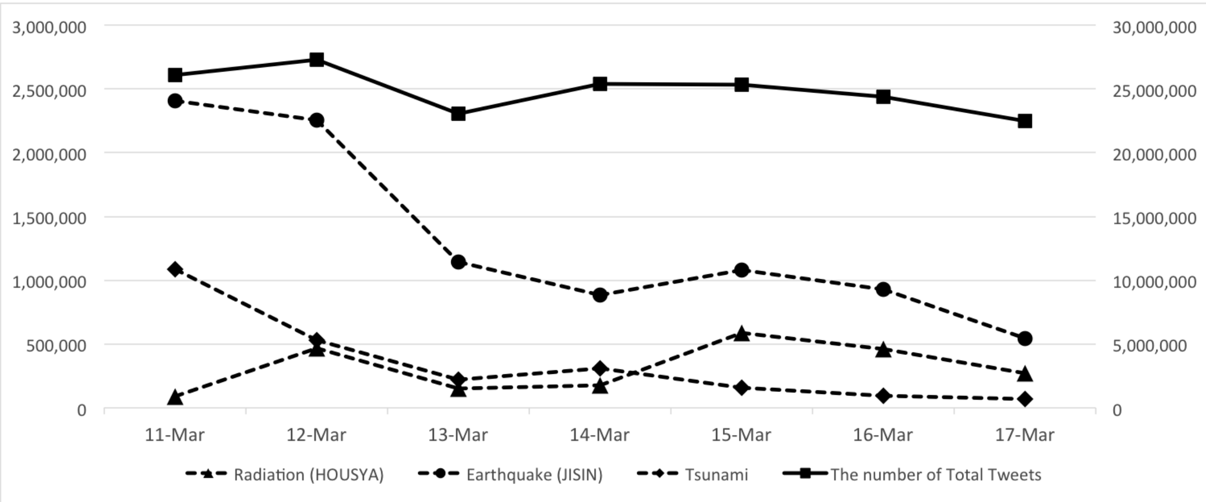

Figure 3.3: The Time Series of the Quantity of Tweets Containing Words Radiation(HOUSYA), Earthquake(JISHIN) and Tsunami

Table 3.2: The Total Number of Tweets containing words Radiation(HOUSYA), Earthquake(JISHIN) and Tsunami

	11-Mar	12-Mar	13-Mar	14-Mar	15-Mar	16-Mar	17-Mar	Total
Radiation (HOUSYA)	90,344	467,414	153,138	175,560	590,087	458,443	274,153	2,209,139
Earthquake (JISIN)	2,405,752	2,253,085	1,143,233	885,780	1,080,522	928,951	545,696	9,243,019
Tsunami	1,086,893	532,505	224,088	309,546	160,491	95,519	68,812	2,477,854
The number of Total Tweets	26,099,910	27,315,744	23,065,384	25,404,539	25,338,601	24,368,393	22,465,493	174,058,064

3.3.3 Transmission of Tweets by Retweet Related to the Disaster

In Twitter, tweets containing disaster information transmitted rapidly as twitter user retweeted, that resent the messages they provided by their following users. It is a behavior that transmit the information which they were given from their following users to the users who follow them. There are two ways to retweet which is called an official retweet and unofficial retweet [Yamamoto et al., 2012]. First one is created by pressing the "retweet button" and it automatically transmitted the messages to the follower of the user. Second one is created by user themselves that copy and paste the messages conforming to forms as below.

- RT @name of Twitter account: The contents of tweet

In the data we have, official retweet is represented as same as unofficial retweet, which is, both way of RT represented as same as above. So, in this analysis, we consider both of an official and an unofficial retweet as a retweet. Therefore, tweets can be sorted as original tweets and retweets. As mentioned, retweet is defined as information transmission behavior. In the disaster, Twitter was utilized as a platform to share disaster information which is originally created by organization such as government agencies, broadcast and experts. Especially, a series of explosion accidents in nuclear power plant made people express alarm at the danger of radiation. Since, the radiation is one of the risk that people perceived it as poorly understood and unknown risk but also dread risk that latent cancer fatalities are expected, it produced huge societal impacts compare to other risks [Slovic, 1987]. Table3.3 shows the proportion of original tweets and retweets in tweets mentioning 'radiation(HOUSYA)'. After the messages regarding radiation is created, they were transmitted almost twice as many as originally created tweets. Table3.4 shows which

Table 3.3: The Proportion in Tweets including Word 'Radiation (HOUSYA)'

Tweets (Original Tweets)	Retweets
34%	66%

Twitter account were retweeted the most in those days. Mainly, Experts in radiation-related research field and broadcating organization such as NHK (Japan Broadcating Corporation) were played great role in initial distributor of information about danger of the radiation. Particularly, tweets that created by the expert in physics who has Twitter account name 'hayano', was widely shared. Because the feature of the risk radiation is complicated for public to understand the meaning of information without background knowledges such as technical terms, people were seeking views and opinion about the unfamiliar risk and sharing it with others in Twitter.

Table 3.4: The 10 most Retweeted Twitter Account and the The number of Times their Tweets were Retweeted in Tweets Containing Word Radiation (HOUSYA)

	Name of Twitter Account	Organization/Profession	The number of Times their Tweets were Retweeted	The Proportion in Retweeted tweets including Word 'Radiation (HOUSYA)'
1	hayano	Expert in Physics	79,882	5%
2	NHK_PR	Broadcasting Organization	58,755	4%
3	fukanju	Radiotherapist	50,467	3%
4	nhk_HORIJUN	Broadcasting Organization	38,225	3%
5	team_nakagawa	Group of Expert in Radiation Therapy	36,761	2%
6	nhk_kabun	Broadcasting Organization	35,543	2%
7	funky_konbu	Unknown	25,087	2%
8	itokenstein	Composer	23,998	2%
9	CAjapan	Flight Crew Association of Japan	23,077	2%
10	ikedanob	Economist	26,071	2%

3.4 Disaster Information Distribution by Government Agencies

3.4.1 The Contents of Disaster Information Distribution

Government agencies is one of the key information distributor in the disasters. Some of them were actively communicate through Twitter which is useful platform to inform the emergency instructions of evacuation when communication through telephone call or short messages was disable but connecting to the internet was possible with their smart phone. Not only the disaster area, but central government agencies were also provided disaster information to citizen in all parts of Japan. Since not all of government agencies send messages in those period of time, Table 3.5 shows the information of government agencies that distributed the disaster-related information through Twitter on the 7 days. For selecting the agencies, we investigated every prefecture if there are official twitter account and they utilized Twitter on the days. Especially, among government agencies of disaster areas which are Fukushima, Miyagi, Iwate Prefecture and Hachinose City(Aomori Prefecture), only 5 agencies tweeted on the 7 days. Osyu city (place in Iwate prefeture) tweeted the most, over 700 times in the days.

The contents of disaster information sent to Twitter were different from each government agency and areas. Table 3.6 shows the 30 most spoken terms (Term Frequency) of agencies in each areas. As the central government agencies who had made announcement aiming at people in all over the Japan, their contents ranged over all major risks such as ‘GENPATSU (nuclear power plant)’, ‘JISHIN(earthquake)’ and ‘TSUNAMI(tsunami)’ and also ‘TEIDEN(planned power cut)’. On the other hand, the agencies in the disaster areas mentioned many times regarding areas suffering from disaster such as ‘OSHU(oshu)’, ‘HACHINOSE (hachinose)’ and ‘iwate’. And also the information(‘JYOUHOU (information)’) of evacuation(‘HINAN (evacuation)’) and restoration(‘HUKKYUU (restoration)’) had become a much-talked-about issue of those agencies. About agencies of areas close to the disaster area, ‘TEIDEN(planned power cut)’ was the most spoken risk, because the planned power cut was emergently carried out in east-japan area and they talked a lot about ‘HOUSYASEN (radiation)’ as well which is influential on a broad area compared with ‘TSUNAMI (tsunami)’ and ‘JISIN (earth- quake)’. In non disaster- stricken area,

Table 3.5: The number of Tweets provided by Government Agency

Government Agency	User ID	The number of Tweets
Central Government Agency		
Prime Minister of Japan and His Cabinet	Kantei_Saigai	198
Fire and Disaster Management Agency	FDMA_JAPAN	98
Ministry of health, Labour and Welfare	MHLWitter	17
Ministry of Economy, Trad and Industry	meti_NIPPON	18
(Tokyo Electric Power Company)	(OfficialTEPCO)	(2)
The Disaster Area		
Iwate Prefecture Oshu City	oshu_city	731
Miyagi Prefectrue Kesennuma City	bosai_kesennuma	67
Fukushima Prefecture Aizuwakamatus City	aizuwakamatsuct	74
Fukushima Prefecture Minamiaizu Town	minamiaizu_town	53
Aomori Prefecture HachinoheCity	HachinoheCity	178
Area Close to The Disaster Area		
Aomori Prefecture	AomoriPref	300
Aomori Prefecture Aomori City	AomoriShi	165
Aomori Prefecture Mutsu City	mutsukoho	79
Akita Prefecture	pref_akita	64
Ibaraki Prefecture	Ibaraki_Kouhou	54
Saitama Prefecture Tokigawa Town	tokigawamachi	147
Saitama Prefecture	SaitamaCityPR	24
Chiba Prefecture Matsudo City	matsudo_city	163
Chiba Prefecture Urayasu City	urayasu_koho	136
Kanagawa Prefecture	KanagawaPref_PR	38
Niigata Prefecture	Niigata_Press	179
Ishikawa Prefecture Nomisi City	nomicity	4
Sizuoka Prefecture	rc_shizuokaken	118
Non Disaster-Stricken Area		
Mie Prefecture Kuwana City	kuwana_city	20
Tottori Prefecture	tottori_kouhou	24
Kochi Prefecture	pref_kochi	69
Saga Prefecture	saga_kouhou	332
Nagasaki Prefecture Hirado City	HIRADOcity	32
Kumamoto Prefecture	KumamotoPre_koh	3
Oita Prefecture Oita City	OitaCity_PR	74
Miyazaki Prefecture	miyazakipref	28

even risks such as ‘TSUNAMI (tsunami)’ and ‘JISIN (earth- quake)’ were still spoken a lot, words ‘SHIEN (sup- port)’ and ‘BUSHI (commodities)’ was ranked in first and third most spoken terms. It seems that they mentioned things about supporting commodities or contribution (GIENKIN) for disaster areas. As shown above, government agencies provided information considering the concerns of the people in corresponding area.

Table 3.6: The Contents of Information provided by Government Agencies

	Central Government Agency		The Disaster Area		Area Close to The Disaster Area		Non Disaster-Stricken Area	
	Word	TF	Word	TF	Word	TF	Word	TF
1	KAIKEN (interview)	140	oshu	480	aomori	336	SHIEN (support)	219
2	CHOUKAN (secretary)	125	jishin	364	JYOUHOU (information)	332	JYOUHOU (information)	186
3	HIGAI (damage)	116	JYOUHOU (information)	306	AOMORI (aomori)	267	BUSHI (commodities)	132
4	FUKISHIMA (fukushima)	102	OSHU (oshu)	221	TEIDEN (power cut)	260	HISAI (suffering from)	130
5	GENPATSU (nuclear power plant)	97	HINAN (evacuation)	205	TSUJYOU (ordinary)	212	TSUNAMI (tsunami)	127
6	JYOKYOU (state of things)	71	HACHINOSE (hachinose)	165	JISIN (earthquake)	191	SAGA (saga)	118
7	SYOUBOUCHOU (fire Defense Agency)	56	JYOKYOU (state of things)	123	JYOKYOU (state of things)	187	JISHIN (earthquake)	110
8	PDF	55	TAISAKU (countermeasure)	98	KEIKAKU (plan)	181	HINAN (shelter)	109
9	KANBOU (secretariat)	53	KAIGI (meeting)	98	UNKOU (operation)	178	HIGAI (damage)	93
10	TEIDEN (power cut)	50	iwate	98	HINAN (evacuation)	165	CHIHOU (region)	83

Table 3.6: The Contents of Information provided by Government Agencies

	Central Government Agency		The Disaster Area		Area Close to The Disaster Area		Non Disaster-Stricken Area	
	Word	TF	Word	TF	Word	TF	Word	TF
11	SOURI (prime Minister)	46	HUKKYUU (restoration)	93	TSUKI (month)	162	ONEGAI (please)	80
12	JISHIN (earthquake)	40	save	85	KENNAI (within the prefecture)	147	TAIHEIYOU (pacific)	80
13	JYOUHOU (information)	39	HONBU (headquarter)	83	NIIGATA (niigata)	135	RT	77
14	KOKUMIN (citizen)	37	SAIGAI (disaster)	82	aomorist	128	OKI (offshore)	77
15	TSUKI (month)	35	SHINAI (in the city)	81	MINAMIUONUMA (minami-unuma)	128	TOUHOKU (tohoku region)	74
16	KISYA (journalist)	34	koho	79	OSHIRASE (notification)	127	oita	69
17	KEIKAKU (plan)	33	ANZEN (safety)	79	YOTEI (schedule)	123	pref	68
18	HISAI (suffering from)	33	JISIN (earthquake)	76	SOUDAN (consultation)	111	iwate	63
19	EDANO (edano)	31	hachinohe	75	GURUUPU (group)	110	UKETSUKE (acceptance)	60
20	APPU (Up)	29	ONEGAI (please)	74	HOUSYASEN (radiation)	110	KEIHOU (warning)	57
21	BAKUHATSU (explosion)	27	TSUKI (month)	71	AGA (aga)	102	KENNAI (within the prefecture)	54
22	MESSEIGI (message)	27	TSUNAMI (tsunami)	69	urayasu	99	SAIGAI (disaster)	52

Table 3.6: The Contents of Information provided by Government Agencies

	Central Government Agency		The Disaster Area		Area Close to The Disaster Area		Non Disaster-Stricken Area	
	Word	TF	Word	TF	Word	TF	Word	TF
23	HINAN (evacuation)	25	TAIOU (correspondence)	68	GENPATSU (nuclear power plant)	99	JYOKYOU (state of things)	51
24	ONEGAI (please)	24	ANSHIN (relief)	64	KOUTSUU (traffic)	97	KOUCHI (kochi)	48
25	SAIGAI (disaster)	24	OSHIRASE (notification)	62	KASHIWAZAKI (kashiwazaki)	93	GIENKIN (contribution)	46
26	km	22	KOUSHIN (renew)	59	BASU (bus)	92	KYOURYOKU (cooperation)	44
27	DOUGA (video)	22	ENGAN (coast)	59	TAIHEIYOU (pacific)	92	OSHIRASE (notification)	43
28	TSUNAMI (tsunami)	21	HOMUPEEJI (homepage)	58	NISHI (nishi)	91	RT	41
29	TEKISUTO (text)	21	ESASHI (esashi)	57	NAGAOKA (nagaoka)	90	KEN (prefecture)	39
30	CHIHOU (region)	21	MINAMIAIZU (minamiaizu)	54	SENTA (center)	90	TSUKI (month)	39

Table 3.7 shows more clear difference between agencies. It compares agencies in disaster areas. Their tweets was containing terms of serious risks which they faced with. Kesennuma City (Twitter account name: bosai_kesennuma) which had huge damage from the tsunami emphasized the ‘HINAN(evacuation)’ from tsunami by going up ‘TAKADAI(hill)’ in their tweets. On the other hand, Minamiaizu town (Twitter account name: minamiaizu_town) concerned about ‘HOUSYASEN (radiation)’ issue, mentioning the information related to radiation using scientific words such as ‘microsievert’, since the town is placed near Fukushima nuclear power plant.

Table 3.7: The Contents of Information provided by Government Agencies of The Disaster Areas

	Disaster Area									
	oshu_city		bosai_kesennuma		aizuwakamatsuct		minamiaizu_town		HachinoheCity	
	Word	TF	Word	TF	Word	TF	Word	TF	Word	TF
1	oshu	480	HINAN (shelter)	43	SAIGAI (disaster)	31	MINAMIAIZU (minamia	54	HACHINOHE	102
2	jishin	318	TSUNAMI (tsunami)	43	aizu	30	minamiaizu	31	JYOUHOU (information)	84
3	OSHU (oshu)	221	KEIHOU (warning)	20	TSUKI (month)	30	HOUSYASEN (radiation)	29	koho	79
4	JYOUHOU (information)	191	TAKADAI (a hill)	12	SOKUTEI (measurement)	19	FUKUSHIMA (fukushima)	15	ANZEN (safety)	76
5	JYOKYOU (state of things)	100	KASAI (fire)	9	JYOUHOU (information)	18	KANKYOU (environment)	14	hachinohe	75
6	KAIGI (meeting)	98	HASSEI (an outbreak)	7	HINAN (shelter)	18	REBERU (level)	14	HATHINOHE (hachinohe)	63
7	iwate	98	JYOUHOU (information)	6	SIZEN (nature)	14	TAIKI (the atmosphere)	14	ANSHIN (relief)	63
8	TAISAKU (countermeasur e)	95	HATSUREI (an official announcement)	6	AIZUWAKAMATSU	14	OHABA (big gap)	14	JISIN (earthquake)	54
9	HINAN (shelter)	87	SHIGAICHI (urban area)	6	BORANTIA (volunteer)	13	JINTAI (human body)	14	hMail	53
10	save	84	KESENNUMA (kesennuma)	5	SENTAA (center)	13	CHOUSA (investigation)	14	HINAN (shelter)	47
11	HUKKYUU (restoration)	81	YOSHIN (an aftershock)	5	AIZU (aizu)	13	EIKYO (influence)	14	jishin	46
12	HONBU (headquarter)	81	SHINAI (in the city)	5	MATSUNAGA(mat sunaga)	13	GOUDOUCHOSYA (common building for government offices)	13	SHINDO (a seismic intensity)	40
13	SHINAI (in the city)	58	SHINSUI (flooded)	5	SUKURININGU (screening)	12	MIKUROSIIBERUTO (microsievert)	12	HAPPYOU (an announcement)	32
14	TAIOU (correspondenc	58	MIYAGI (miyagi)	4	CHIHOU (region)	12	HINAN (shelter)	10	OKI (sea)	30
15	ESASHI (esashi)	57	FUKIN (nearby)	4	HOUSYASEN (radiation)	11	YAKYOKU (farmacy)	9	MAGUNICYUDO (magnitude)	28
16	HOUMUPEIJI (homepage)	52	kesennuma	3	KANKYOU (environment)	10	TAIOU (correspondence)	9	SHINGEN (origin of earthqaueke)	28

3.4.2 Disaster Information Transmission by Retweets

After each agency release disaster information, people who were sent the information passed it on to another user by retweeting. Table 3.8 shows the number of retweeted of immediately after their tweets. Tweets of two central government agencies (Prime Minister of Japan and His Cabinet and Fire and Disaster Management Agency) were retweeted the most, indicating that their information got a lot of attentions as they released the information targeting audience who were not only disaster areas but whole nation. Meanwhile, there was no a clear correlation between the number of tweets and retweeted. There could be several reasons, but one is that retweeting is depends on network which is already formed. The information in Twitter is transmitted through preformed network. This indicates that if government agency make use of Twitter for disaster information announcement, it is necessary to investigate process of delivery of information in Twitter.

3.5 Conclusion

In this chapter, we investigate the actual crisis communication in The Great East Japan Earthquake and The Fukushima Daiichi Nuclear Disaster. People who are in disasters, use Twitter to describe the events or accidents that is happening or happened just before. In addition, it it observed that they apply twitter as a tool for delivery of information that originally created by group of experts or official organization. It indicates that as they retweeted, they distributed the information by themselves without the media. As for government organization who play a role as a information provider in communication, the contents of information they provide usually depend on concern risks that they are encounter with. So, the contents of central government covered disaster risks such 'radiation' which affected broadly across wide areas, while local government focusing on specific risks such as 'tsunami' that they are faced with. As there was no a clear correlation between the number of tweets and retweeted, it is necessary to consider other elements such as a preformed network which is constructed before the disaster occur to investigate process of delivery of information in Twitter.

Table 3.8: The Number of Retweets of Government Agencies's Tweets

Government Agency	The number of Tweets	The number of Retweeted	The Proportion in Total number of Retweets
Central Government Agency			
Prime Minister of Japan and His Cabinet	198	96,099	0.21%
Fire and Disaster Management Agency	98	134,945	0.30%
Ministry of health, Labour and Welfare	17	24,957	0.05%
Ministry of Economy, Trad and Industry	18	997	0.00%
Tokyo Electric Power Company	2	23,162	0.05%
The Disaster Area			
Iwate Prefecture Oshu City	731	8,032	0.02%
Miyagi Prefectrue Kesennuma City	67	13,981	0.03%
Fukushima Prefecture Aizuwakamatus City	74	15,990	0.04%
Fukushima Prefecture Minamiaizu Town	53	3,109	0.01%
Aomori Prefecture HachinoheCity	178	273	0.00%
Area Close to The Disaster Area			
Aomori Prefecture	300	25,256	0.06%
Aomori Prefecture Aomori City	165	1,374	0.00%
Aomori Prefecture Mutsu City	79	441	0.00%
Akita Prefecture	64	1,723	0.00%
Ibaraki Prefecture	54	1,000	0.00%
Saitama Prefecture Tokigawa Town	147	167	0.00%
Saitama Prefecture	24	407	0.00%
Chiba Prefecture Matsudo City	163	3,039	0.01%
Chiba Prefecture Urayasu City	136	42,352	0.09%
Kanagawa Prefecture	38	194	0.00%
Niigata Prefecture	179	5,211	0.01%
Ishikawa Prefecture Nomisi City	4	11	0.00%
Sizuoka Prefecture	118	508	0.00%
Non Disaster-Stricken Area			
Mie Prefecture Kuwana City	20	72	0.00%
Tottori Prefecture	24	30	0.00%
Kochi Prefecture	69	2,478	0.01%
Saga Prefecture	332	3,423	0.01%
Nagasaki Prefecture Hirado City	32	81	0.00%
Kumamoto Prefecture	3	57	0.00%
Oita Prefecture Oita City	74	202	0.00%
Miyazaki Prefecture	28	45	0.00%

Bibliography

[1] Eliza Strickland. Explainer: What Went Wrong in Japans Nuclear Reactors, IEEE spectrum 2011 (Jan.15,2016) http://spectrum.ieee.org/tech-talk/energy/nuclear/explainer-what-went-wrong-in-japans-nuclear-reactors.

[2] P.E. Slovic. Perception of risk. *Science*, 236(4799), pp. 280-285, 1987.

[3] Yamamoto Masahito, Ogasawa Hiroya, Suzuki Ikuo, and Fukukawa Masashi. Information Propagation Network for 2012 Tohoku Earthquake and Tsunami on Twitter. *Information Processing Society of Japan*, 53(11), pp.1184-1191, 2012.

Chapter 4

Estimating Concerns of the Public with Latent Dirichlet Allocation

4.1 Introduction

In this chapter, we examine how concerns of the public changes along with the state of situation changes in disaster. In modern society, risk has become extremely complex [Koabayashi, 2013]. And it leads people to make difficult to predict and reason the situation. In a disaster, situation regarding damages or accidents changes from moment to moment, the public are surrounded by various kind of information.

To deliver communication and make appropriate response, it is important to identify the concerns of the public toward emergency situation and risks that newly emerge.

So, the aim of this analysis is to clarify the concerns which people possess along time changing in disaster and examine risk perception in disaster. Since there has been great advancement in mobile communication technology, tools for individuals to communicate have been highly diversified. It leads the public to actively communicate with others regardless of time and place using mobile phone. Recently, social media (Twitter) has been used as a place to exchange the information related to disaster or to share their perception of the risks and sentiment.

In the analysis, topic model LDA (Latent Dirichlet Allocation) [Blei et al., 2003] is applied to extract the concerns of the public. For inferencing model, we use variational inference algorithm. We extract the significant terms for every 12 hours and examine the topics of each document from Twitter data collected right after The Great East Japan Earthquake. Given the topic of Twitter in disaster, results showing what people are concern about and how it changes may provide clues about risk perception of public in the early stage of the disaster.

This chapter is organized as follows. 4.2 explains the basic ideas about the concerns of the public in disaster. 4.3 explains the Twitter data we use. 4.4 describes the topic models (LDA) and application proposed in this study. 4.5 presents Twitter corpus based analysis using proposed application, investigating the concern of the public during the disaster. 4.6 explains implication. Finally, 4.7 describes conclusions and future work regarding this research.

4.2 Basic Idea

4.2.1 Concerns of the public in disasters

In the middle of the crisis, people carry out communication for minimizing the scale of the damage [Kikkawa, 2000]. They focused on not only the tasks that directly caused by disaster but also the threatening factors that they may widen the scale of the damage in the near future. So, the former is more related to, for instance, restoration or recovery from the damages and the latter is regarding the risks they newly encounter.

Concern of the public could be defined as public awareness of a problem. In other words, it is a thing that people pay attention to in disasters. Among numerous information, they focused on certain subjects. According to Lupia [Lupia, 2008], for making reasoned choices, people pay attention to the information which help them to avoid the risk of future pain or increase the opportunity for future pleasure. When people evaluated information for reasoned behavior, the way that they perceive the risk influences it.

In modern society, the profound development of science and technologies has been accomplished. Risk has become increasingly complex nowadays. As the dependency on nuclear technologies and biological and chemical substances generated by genetic engineering grows, it will be difficult for people to assess the hazard by statistical analysis [Slovic, 1987] [Kobayashi, 2013]. Since risks caused by a catastrophe are specialized, expertise has become fragment. There exist a large gap between experts and the public in background knowledges which require for understanding the impact or the harmfulness of the hazard in disaster [Kobayashi, 2013]. And this gap affect to both of them and it determines how differently they define the risk they encounter. According to Slovic [Slovic, 1987], people's risk perception often significantly deviate from objective risk and concept of risk means different from each person, while experts judge risk by correlating with technical estimates of annual fatalities. It seems that people judge risk by assessing catastrophic potentials or impact to future generations. This finding explains why the public's perception of risk sometimes subjective, hypothetical and even irrational in the disaster compared to risk that assess by experts. For people, the risk is not the thing that can be evaluated only with objective value, rather it is inherently subjective and reflect people's social value or societal situation that they are surround by [Slovic, 2001] [Kobayashi, 2013]. The disagreement regarding risk between main participants of crisis communication could cause misunderstanding or confusion throughout a whole society. In disasters, an announcement by government containing information regarding a disaster produced by experts of the each area of expertise sometimes does not reflect the public's concept of the risk, as a result, it have little effect on changing people's behavior and attitude. So, first, to understand the risk perception of the public, it is necessary to clarify that what actually people are concerning about in disasters.

4.2.2 LDA Topic Model for Identifying Concerns of the Public

In this subsection, we explain the LDA model which is a generative model that viewing documents as mixtures of probabilistic topics the latent variable topic model. Each document are represented by a multinomial distribution over latent variable topics and each topic is decided by a multinomial distribution over words. It is unsupervised model which is requiring feature for Twitter corpus. Contents analysis on Twitter corpus requires little supervision since it contains few words compared to the standard written English of which normally many natural language processing model are developed for analysis [Ramage et al., 2010]. Kireyev [Kireyev et al., 2009] explains several challenges on utilizing Twitter in Natural language processing. First, the corpus contains "Esoteric language and grammar", second, short "message length" as explained, third, "Locale-specific references" which explain that Twitter including lots of proper noun representing specific location, events or name of entities. However, Topic model have several promising features for coping with those points. It is argue that topic model such as LDA is available for analyzing the particular data like Twitter for following reasons. First, topic model does not consider the syntactic construction and order of words, but it only matters the words occurring. Second, topic model such as LDA are focusing on inferring latent relationships between words in corpus. So, misspellings is more easy to handle [Kireyev et al., 2009].

4.3 Data

Twitter data was provided by Twitter Japan via Project 311 (The Great East Japan Earthquake Big Data Workshop Project 311, 2015). It was collected over the seven days from March 11 (about an hour before the earthquake occurred) to March 17, 2011. Tweets posted during those 7 days and also those written in Japanese were included to the data. The data comprises tweet IDs, user IDs, time and tweet contents. In this analysis, a 2.9% sample extracted from the provided data is used. As the corpora in LDA model is composed of documents, we amalgamate tweets for every 12 hours (AM, PM) that 14 documents are made for those 7 days. Every document contains 420,000 tweets and 14 documents is prepared for a corpora (except 11 March, the provided dataset contains the tweets from 9 a.m. in 11, March, so tweets before 9 a.m. is omitted).

Table 4.1: The Outline of Sample Data

Document	Date	Sample Data	Total (Provided)
1	11-Mar A.M.	105,000	26,099,910
2	11-Mar P.M.	420,000	
3	12-Mar A.M.	420,000	27,315,744
4	12-Mar P.M.	420,000	
5	13-Mar A.M.	420,000	23,065,384
6	13-Mar P.M.	420,000	
7	14-Mar A.M.	420,000	25,404,539
8	14-Mar P.M.	420,000	
9	15-Mar A.M.	420,000	25,338,601
10	15-Mar P.M.	420,000	
11	16-Mar A.M.	420,000	24,368,393
12	16-Mar P.M.	420,000	
13	17-Mar A.M.	420,000	22,465,493
14	17-Mar P.M.	420,000	
Total		5,145,000 (2.95%)	174,058,064

4.4 The Methodology and The Application

4.4.1 Latent Variable Topic Model (Latent Dirichlet Allocation)

The LDA model makes it possible to formulate the problem of discovering the set of topics that are latent variable in a collection of documents. First, text collections that assume in the model is collection of "documents" which called "corpora" and a document is consist of "words". Formally, the terms are defined as below.

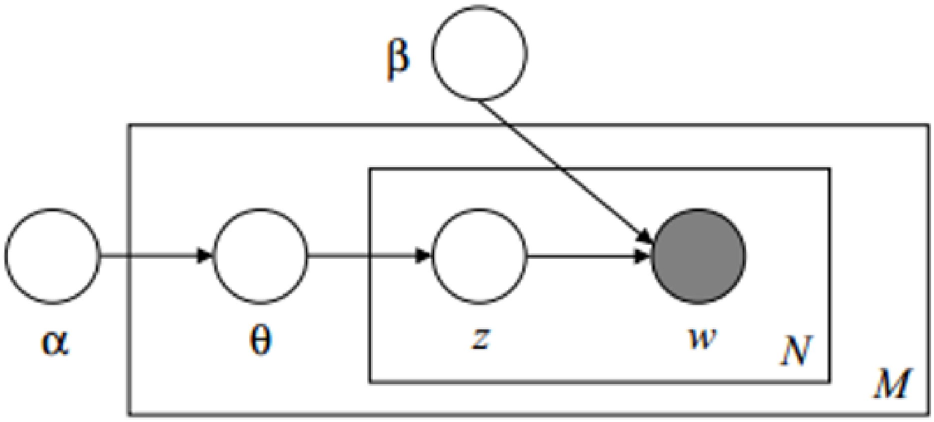

Figure 4.1: Graphical model representation of LDA

Source:Blei et. al (2003) Latent Dirichlet Allocation

- A word is the basic unit of discrete data, defined to be an item from a vocabulary indexed by $\{1, \ldots, V\}$. So, in corpora, it is assumed that V number of terms exit which is not duplicated.

- A document is a sequence of N words represented by $\boldsymbol{w} = (w_1, w_2, \ldots, w_N)$, where w_n is the nth word in the sequence.

- A corpus is a collection of M documents represented by $D = \{\boldsymbol{w_1}, \boldsymbol{w_2}, \ldots, \boldsymbol{w_M}\}$.

As explained, LDA is a generative probabilistic model of a corpus. The basic idea is that documents are represented as random mixtures over latent topics, where each topic is characterized by a distribution over words. Figure 4.1 shows graphical model representation of LDA. A generative process for each document $\boldsymbol{w}$ in a corpus D is:

1. Choose $N \sim Poisson(\xi)$

2. Choose $\boldsymbol{\theta} \sim Dir(\alpha)$

3. For each of the N words w_n:

   (a) Choose a topic $z_n \sim Multinomial(\theta)$

   (b) Choose a word w_n from $p(w_n|z_n, \beta)$, a multinomial probability conditioned on the topic z_n

Before explaining each variables, we briefly explained several simplifying assumptions that are made in this model. First, the number of topics in a corpus D is assumed known and fixed. Second, the word probabilities are parameterized by a $K \times V$ matrix $\boldsymbol{\beta}$ where $\beta_{kj} = p(w^j = 1|z^k = 1)$, which treated as a fixed quantity that is to be estimated. Finally, the Poisson distribution for choosing number of sequence in a document is not critical assumption and it is independent of all the other data generating variables($\boldsymbol{\theta}$ and $\boldsymbol{z}$).

k-dimensional Dirichlet random variable $\boldsymbol{\theta}$ has probability density as below. $\boldsymbol{\theta}$ is a K-vector and lies in the $K-1$ simplex if $\theta_k \geq 0$ and $\sum_{k=1}^{k} \theta_k = 1$. And its parameter α is a K-vector with $\alpha_k > 0$.

$$p(\boldsymbol{\theta}|\boldsymbol{\alpha}) = Dir(\boldsymbol{\theta}|\boldsymbol{\alpha}) = \frac{\Gamma(\sum_{k=1}^{K} \alpha_k)}{\prod_{i=k}^{K} \Gamma(\alpha_k)} \theta_1^{\alpha_1 - 1} \cdots \theta_K^{\alpha_K - 1} \tag{4.1}$$

Given parameters α and β, the joint distribution of a topic mixture $\boldsymbol{\theta}$, a set of N topics $\boldsymbol{z}$, and a set of N words $\boldsymbol{w}$ is given by:

$$p(\boldsymbol{\theta}, \boldsymbol{z}, \boldsymbol{w}|\boldsymbol{\alpha}, \boldsymbol{\beta}) = p(\boldsymbol{\theta}|\boldsymbol{\alpha}) \prod_{n=1}^{N} p(z_n|\boldsymbol{\theta}) p(w_n|z_n, \boldsymbol{\beta}) \tag{4.2}$$

where $p(z_n|\theta)$ is simply θ_i. Integrating over $\boldsymbol{\theta}$ and summing over z, the marginal distribution of a document is obtained as below.

$$p(\boldsymbol{w}|\boldsymbol{\alpha}, \boldsymbol{\beta}) = \int p(\boldsymbol{\theta}|\boldsymbol{\alpha}) \left(\prod_{n=1}^{N} \sum_{z_n} p(z_n|\boldsymbol{\theta}) p(w_n|z_n, \boldsymbol{\beta}) \right) d\theta \tag{4.3}$$

Since this is the marginal distribution of a single document, Finally, taking the product of the marginal probabilities of a document, probability of a corpus is obtained as below.

$$p(\boldsymbol{D}|\boldsymbol{\alpha}, \boldsymbol{\beta}) = \prod_{d=1}^{M} \int p(\boldsymbol{\theta}|\boldsymbol{\alpha}) \left(\prod_{n=1}^{N_d} \sum_{z_{dn}} p(z_{dn}|\boldsymbol{\theta}_d) p(w_{dn}|z_{dn}, \boldsymbol{\beta}) \right) d\theta_d \tag{4.4}$$

We have described the LDA model that how the corpora (a collection of documents) is generated using data generating variables ($\boldsymbol{\theta}_d$, z_{dn})and the parameters ($\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$).

4.4.2 Model Inference

For inferencing to use LDA model, it is necessary to estimate the latent variables in a document. The distribution of the hidden variables given a document is represented as below

$$p(\boldsymbol{\theta}, \boldsymbol{z}|\boldsymbol{w}, \boldsymbol{\alpha}, \boldsymbol{\beta}) = \frac{p(\boldsymbol{\theta}, \boldsymbol{z}, \boldsymbol{w}|\boldsymbol{\alpha}, \boldsymbol{\beta})}{p(\boldsymbol{w}|\boldsymbol{\alpha}, \boldsymbol{\beta})} \tag{4.5}$$

Unfortunately this distribution is intractable to compute in general. Also, the distribution of the model which we marginalized Eq (4.3) is intractable due to the coupling between θ and β [Dickey, 1983]:

$$p(\boldsymbol{w}|\boldsymbol{\alpha}, \boldsymbol{\beta}) = \frac{\Gamma(\sum_{k=1}^{K} \alpha_k)}{\prod_{i=k}^{K} \Gamma(\alpha_k)} \int \left(\prod_{i=k}^{K} \theta_i^{\alpha_k - 1} \right) \left(\prod_{n=1}^{N} \sum_{i=k}^{K} \prod_{j=1}^{V} (\theta_k \beta_{kj})^{w_n^j} \right) d\theta$$

Although the distribution of the hidden variables given a document is intractable to compute for exact inference, there are several approximate inference algorithms that can be considered. Griffiths and Steyvers [Griffiths and Steyvers, 2004]represent method solving estimating problem by using a Monte Carlo procedure, resulting in an algorithm that is easy to implement. However in the procedure, they fixed the parameter α and β, so estimation is not done explicitly. Blei et al. [Blei et al., 2003] proposed variational Inference model which is method that estimating the simplified model which is modified from the original model. In this study, we choose variational algorithm for inference in LDA for Twitter data (Figure 4.2).

Variational Inference

The variational inference is accomplished by using adjustable lower bound on the log likelihood ($p(\boldsymbol{w}|\boldsymbol{\alpha}, \boldsymbol{\beta})$). We consider the lower bound indexed by a set of variational parameters which is different from original LDA model parameters. So we set a new model which is a simple modification of original graphical model. For modification, we drop the edges between $\boldsymbol{\theta}$,$\boldsymbol{z}$ and $\boldsymbol{w}$, since these edges make the model to be explicitly inference. Variational distribution simplified with variational parameters are as below:

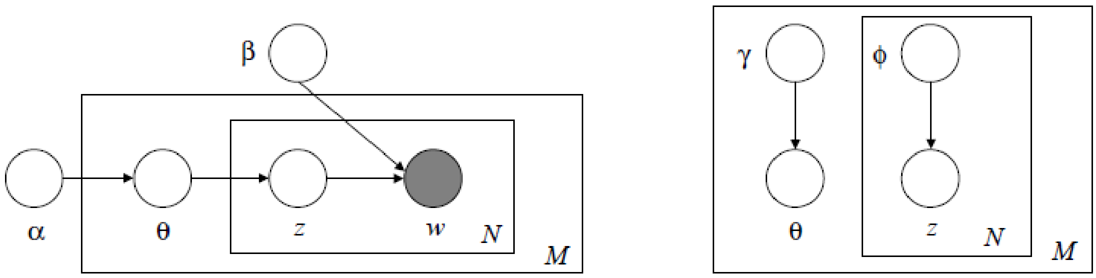

Figure 4.2: (Left) Graphical model representaion of LDA. (Right) Graphical model representation of the variational distribution used to approximate the posterior in LDA

Source:Blei et. al (2003) Latent Dirichlet Allocation

$$q(\boldsymbol{\theta}, \boldsymbol{z}|\boldsymbol{\gamma}, \boldsymbol{\phi}) = q(\boldsymbol{\theta}|\boldsymbol{\gamma}) \prod_{n=1}^{N} q(z_n|\phi_n) \tag{4.6}$$

where the Dirichlet parameter $\boldsymbol{\gamma}$ and the multinomial parameters $(\phi_1, \ldots, \phi_N)$ are the free variational parameters. Now, we have simplified new probability distribution, and then we determine the variational parameter $\boldsymbol{\gamma}$ and ϕ by optimization procedure. Basically, we measure the closeness of the two distributions with Kullback-Leibler(KL) divergence. It comes from information theory, our goal is to minimize the KL divergence. In the case that we have original distribution p and variational distribution q, the KL divergence for variational inference is (omitting the parameters $\boldsymbol{\gamma}$ and ϕ)

$$\begin{aligned} KL(q(\boldsymbol{\theta}, \boldsymbol{z})||p(\boldsymbol{\theta}, \boldsymbol{z}|\boldsymbol{w}, \boldsymbol{\alpha}, \boldsymbol{\beta})) &= E_q\left[\log \frac{q(\boldsymbol{\theta}, \boldsymbol{z})}{p(\boldsymbol{\theta}, \boldsymbol{z}|\boldsymbol{w}, \boldsymbol{\alpha}, \boldsymbol{\beta})}\right] \\ &= E_q\left[\log p(\boldsymbol{\theta}, \boldsymbol{z}|\boldsymbol{w}, \boldsymbol{\alpha}, \boldsymbol{\beta})\right] - E_q\left[\log q(\boldsymbol{\theta}, \boldsymbol{z})\right] \\ &= E_q\left[\log p(\boldsymbol{\theta}, \boldsymbol{z}|\boldsymbol{w}, \boldsymbol{\alpha}, \boldsymbol{\beta})\right] - E_q\left[\log q(\boldsymbol{\theta}, \boldsymbol{z})\right] + \log p(\boldsymbol{\theta}, \boldsymbol{z}|\boldsymbol{w}, \boldsymbol{\alpha}, \boldsymbol{\beta}) \end{aligned} \tag{4.7}$$

This could be also written as below

$$\log p(\boldsymbol{w}|\boldsymbol{\alpha}, \boldsymbol{\beta}) = E_q\left[\log p(\boldsymbol{\theta}, \boldsymbol{z}|\boldsymbol{w}, \boldsymbol{\alpha}, \boldsymbol{\beta})\right] - E_q\left[\log q(\boldsymbol{\theta}, \boldsymbol{z})\right] + KL(q(\boldsymbol{\theta}, \boldsymbol{z})||p(\boldsymbol{\theta}, \boldsymbol{z}|\boldsymbol{w}, \boldsymbol{\alpha}, \boldsymbol{\beta})) \tag{4.8}$$

Actually, minimizing the KL divergence exactly is impossible, but we can solve the problem by maximizing the lower bound on it. In other words, it coulbe be achieved by maximizing the first term of the right hand side of equation (4.8) which is a lower bound on the likelihood for an variation distribution $q(\boldsymbol{\theta}, \boldsymbol{z}|\boldsymbol{\gamma}, \boldsymbol{\phi})$. Let us denote this $L(\boldsymbol{\gamma}, \boldsymbol{\phi}; \boldsymbol{\alpha}, \boldsymbol{\beta})$. So, equation (4.8) could be also represented as below

$$\log p(\boldsymbol{w}|\boldsymbol{\alpha}, \boldsymbol{\beta}) = L(\boldsymbol{\gamma}, \boldsymbol{\phi}; \boldsymbol{\alpha}, \boldsymbol{\beta}) + KLD(q(\boldsymbol{\theta}, \boldsymbol{z}|\boldsymbol{\gamma}, \boldsymbol{\phi})||p(\boldsymbol{\theta}, \boldsymbol{z}|\boldsymbol{w}, \boldsymbol{\alpha}, \boldsymbol{\beta})) \tag{4.9}$$

So, maximizing the lower bound $L(\boldsymbol{\gamma}, \boldsymbol{\phi}; \boldsymbol{\alpha}, \boldsymbol{\beta})$ with respect to $\boldsymbol{\gamma}$ and ϕ maximizing is equivalent to the optimization problem:

$$(\gamma^*, \phi^*) = \arg\min_{(\gamma,\phi)} KLD(q(\boldsymbol{\theta}, \boldsymbol{z}|\boldsymbol{\gamma}, \boldsymbol{\phi})||p(\boldsymbol{\theta}, \boldsymbol{z}|\boldsymbol{w}, \boldsymbol{\alpha}, \boldsymbol{\beta})) \tag{4.10}$$

The lower bound could be expand by using the factorizations of p and q:

$$\begin{aligned} L(\boldsymbol{\gamma}, \boldsymbol{\phi}; \boldsymbol{\alpha}, \boldsymbol{\beta}) = \quad & E_q[\log p(\boldsymbol{\theta}|\boldsymbol{\alpha})] + E_q[\log p(\boldsymbol{z}|\boldsymbol{\theta})] + E_q[\log p(\boldsymbol{w}|\boldsymbol{z}, \boldsymbol{\beta})] \\ & - E_q[\log q(\boldsymbol{\theta})] - E_q[\log q(\boldsymbol{z})] \end{aligned} \tag{4.11}$$

And in terms of the model parameters ($\boldsymbol{\alpha}$, $\boldsymbol{\beta}$) and the variational parameters($\boldsymbol{\gamma}$,ϕ), the lower bound is derived as below:

$$\begin{aligned} L(\boldsymbol{\gamma}, \boldsymbol{\phi}; \boldsymbol{\alpha}, \boldsymbol{\beta}) \quad &= \log\Gamma(\textstyle\sum_{j=1}^k \alpha_j) - \sum_{i=1}^k \log\Gamma(\alpha_i) + \sum_{i=1}^k (\alpha_i - 1)(\Psi(\boldsymbol{\gamma}_i) - \Psi(\sum_{j=1}^k \boldsymbol{\gamma}_j)) \\ &+ \textstyle\sum_{n=i}^N \sum_{i=1}^k \boldsymbol{\phi}_{ni}(\Psi(\boldsymbol{\gamma}_i) - \Psi(\sum_{j=1}^k \boldsymbol{\gamma}_j)) \\ &+ \textstyle\sum_{n=i}^N \sum_{i=1}^k \sum_{j=1}^V \boldsymbol{\phi}_{ni} \boldsymbol{w}_n^j \log \boldsymbol{\beta}_{ij} \\ &- \log\Gamma(\textstyle\sum_{j=1}^k \gamma_j) - \sum_{i=1}^k \log\Gamma(\gamma_i) + \sum_{i=1}^k (\gamma_i - 1)(\Psi(\boldsymbol{\gamma}_i) - \Psi(\sum_{j=1}^k \boldsymbol{\gamma}_j)) \\ &- \textstyle\sum_{n=i}^N \sum_{i=1}^k \boldsymbol{\phi}_{ni} \log \boldsymbol{\phi}_{ni} \end{aligned} \tag{4.12}$$

Then, by using Lagrange multipliers we maximize the lower bound with respect to both of $\boldsymbol{\phi}_{ni}$, the probability that the nth word is generated by latent topic i, and $\boldsymbol{\gamma}_i$, the ith component of the posterior Dirichlet parameter and value of each variational parameter is

$$\boldsymbol{\phi}_{ni} \propto \boldsymbol{\beta}_{\boldsymbol{iw_n}} \exp\{E_q[\log(\boldsymbol{\theta}_i)|\gamma]\} \tag{4.13}$$

$$\boldsymbol{\gamma}_i = \boldsymbol{\alpha}_i + \sum_{n=1}^N \boldsymbol{\phi}_{ni} \tag{4.14}$$

Variational distribution with parameters γ^* and ϕ^* is conditional distribution varying as a function of $\boldsymbol{w}$. So, it could be written as $q(\boldsymbol{\theta}, \boldsymbol{z}|\gamma^*(\boldsymbol{w}), \phi^*(\boldsymbol{w}))$. Figure 4.3 shows the variational inference procedure.

initialize $\phi_{ni}^{0} := 1/k$ for all i and n
initialize $\gamma_i := \alpha_i + N/k$ for all i
repeat
  for $n = 1$ **to** N
    for $i = 1$ **to** k
      $\phi_{ni}^{t+1} := \beta_{iw_n} \exp(\Psi(\gamma_i^t))$
    normalize ϕ_n^{t+1} to sum to 1.
  $\gamma^{t+1} := \alpha + \sum_{n=1}^{N} \phi_n^{t+1}$
until convergence

Figure 4.3: A variational inference algorithm for LDA

Source:Blei et. al (2003) Latent Dirichlet Allocation

Parameter Estimation

Ultimate reason for inferencing the model is to estimate the parameter $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ that maximize the marginal log likelihood of the data:

$$\ell(\boldsymbol{\alpha}, \boldsymbol{\beta}) = \sum_{d=1}^{M} \log p(\boldsymbol{w}_d | \boldsymbol{\alpha}, \boldsymbol{\beta})$$

In this analysis, we use an empirical Bayes method for parameter estimation in the LDA model as Blei et. al [Blei et. al, (2003)] presented. As we explained $p(\boldsymbol{w}|\boldsymbol{\alpha}, \boldsymbol{\beta})$ is intractable to compute. So, we decided to use variational inference which provide us with a tractable lower bound on the log likelihood. So, first we maximize the lower bound with respect to the variational parameters $\boldsymbol{\gamma}$and ϕ, and then, with fixed value of the parameters $\boldsymbol{\gamma}$and ϕ, maximizes the lower bound with respect to the model parameters $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$. The derivation of Variational EM algorithm for LDA yields the following iterative algorithm:

- E-step For each document, find the optimizing values of the variational parameters $\{\boldsymbol{\gamma}_d^*, \boldsymbol{\phi}_d^* : d \in D\}$.

- M-step Maximize the resulting lower bound on the log likelihood with respect to the model parameters α and β.

EM step are repeated until the lower bound on the log likelihood converges. Since the maximizing the lower bound of the variational parameters (E-step) is estimated for each document, it is need to consider all documents when estimating model parameters α and β

$$L'(\boldsymbol{\gamma}, \boldsymbol{\phi}; \boldsymbol{\alpha}, \boldsymbol{\beta}) = \sum_{d=1}^{M} L(\boldsymbol{\gamma}, \boldsymbol{\phi}; \boldsymbol{\alpha}, \boldsymbol{\beta}) \tag{4.15}$$

Then we can update the M-step for the conditional multinomial parameter β and Dirichlet parameter α. First, we maximize L' with respect to β which is parameter of conditional multinomials by adding Lagrange multipliers:

$$L_{[\beta]} = \sum_{d=1}^{M}\sum_{n=1}^{N_d}\sum_{i=1}^{k}\sum_{j=1}^{V} \boldsymbol{\phi}_{dni}\boldsymbol{w}_{dn}^{j} \log \boldsymbol{\beta}_{ij} + \sum_{i=1}^{k} \boldsymbol{\lambda}_i (\sum_{j=1}^{V} \boldsymbol{\beta}_{ij} - 1) \tag{4.16}$$

then, take the derivative with respect to β_{ij}, set it to zero, then we have β_{ij}:

$$\boldsymbol{\beta_{ij}} \propto \sum_{d=1}^{M}\sum_{n=1}^{n_d} \phi_{dni} w_{dn}^{j} \tag{4.17}$$

Second, α is parameter of Dirichlet distribution and here we omit other terms and consider the terms which contains α in L':

$$L_{[\alpha]} = \sum_{d=1}^{M} \Big(\log\Gamma(\sum_{j=1}^{k} \boldsymbol{\alpha} + j) - \sum_{i=1}^{k} \log\Gamma(\boldsymbol{\alpha}_i) + \sum_{i=1}^{k} ((\boldsymbol{\alpha}_i - 1)(\Psi(\boldsymbol{\gamma}_{di}) - \Psi(\sum_{j=1}^{k} \boldsymbol{\gamma}_{dj}))) \Big) \tag{4.18}$$

Then, the derivative with respect to $\boldsymbol{\alpha}_i$ gives:

$$\frac{\partial L}{\partial \boldsymbol{\alpha}_i} = M(\Psi(\sum_{j=1}^{k} \boldsymbol{\alpha}_j) - \Psi(\boldsymbol{\alpha}_i)) + \sum_{d=1}^{M} (\Psi(\boldsymbol{\gamma}_{di}) - \Psi(\sum_{j=1}^{k} \boldsymbol{\gamma}_{dj})) \tag{4.19}$$

and this derivative depends on $\boldsymbol{\alpha}_j$, where $j \neq i$ and therefore we have to find the maximal $\boldsymbol{\alpha}$ using an iterative method. Here, the linear-time Newton-Raphson algorithm is applied to fine the $\boldsymbol{\alpha}$.

4.4.3 Application for clarifying the concerns of the publics

In this study, the application has been developed using the corpus of Twitter which collected every hour to clarify the concerns of the public in disaster (Figure 5.4). Since the disaster situation changes hour by hour, the response of the public followed the situation changes along

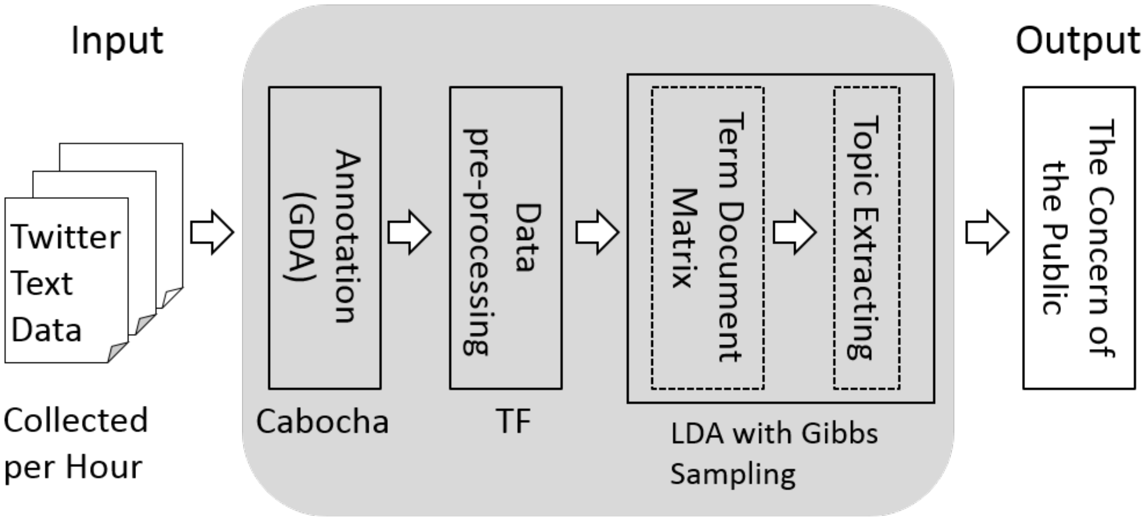

Figure 4.4: Outline of Application

with it. But to investigate what is the matter of concern to the public during disaster, it is important to clarify the latent response of the public.

Figure 4.4 shows the outline of the application.

First, we used Twitter data which is extracted every 12 hours from 11 March to 17 March (except . As we defined crisis communication as cross-societal communication, we consider entire tweets in an hour as response of the public in the time. So a document in out analysis is tweets which were tweeted in an hour. With those documents which consist of 14 documents (except 11 March), we examined the contents of Twitter using LDA topi model every day basis.

Second, it is necessary to tag every term in the text collection of natural language to apply computational linguistics or corpus linguistics. In this study, Cabocha (a Japanese dependency structure analyzer) was used to produce Global Document Annotation (GDA) tagging, which is processed language to be annotated and tagged by attributes.

Third, Data need to be pre-processed before analyzing using LDA, since row text data contains a lot of unnecessary words or characters. It is necessary to select suitable vocabulary to create a corresponding input data for topic model [Grun and Hornik, 2011]. In the pre-processing we removed punctuation characters, numbers, some english terms which used for disclosing web site address (such as http, www, com, or ly) and omitted low frequency terms occur less than 10 times.

Then, we made input data for topic model which is Term Document Matrix. The rows in this matrix correspond to the terms and the columns to the documents. The number of rows is equal to the size of the vocabulary and the number of columns to the size of the corpus.

Last, we fit an LDA model with 30 topics using variational inference. When we decided the number of topics, we conduct the 5 fold-cross validation to find out the optimal number of topics. Fitting LDA model, we utilized R Package (R Package 'topimodels') developed by Grun and Hornik [Grun and Hornik, 2011].

4.5 A Changing Concerns of the Public

4.5.1 Model Selection

The number of topics need to be decided a-priori for fitting the variational EM algorithm to our Twitter data set, since the optimal number of topics is different in every data set. So, for model selection with respect to the number of topics, we evaluate the likelihood for the data. In particular, the perplexity is used to evaluated the models on the data and is equivalent to the per-word likelihood. Model with lower perplexity indicates better probability distribution at predicting the sample.

$$Perplexity(w) = exp\left\{-\frac{\log(p(w))}{\sum_{d=1}^{D}\sum_{j=1}^{V} n^{(jd)}}\right\} \tag{4.20}$$

n^{jd} means how often the jth term occurred in the dth document. When we evaluate the model with perplexity, we uses 5-fold cross-validation. The Twitter data set is split into 5 data sets. And we conduct 5 testings, which we select one for test data with the remaining data as training data for every testing.

Figure 4.5 shows the perpelexities of the 5 test data for the models fitted using variational EM algorithm. The perplexity is lowest about 30 topics, which is the optimal number of topics of the Twitter data set.

And Figure 4.6 represents the α values estimated in cross-validation. And the mean value of α is estimated as 0.026 with 30 topics after 5-fold cross-validation.

4.5.2 Fitting the LDA Model to the Twitter Data Set Using 30 Topics

To identify the concerns of the public in disaster, we use LDA to identify topics in the disaster. Since topic is defined as latent variables in corpus, we apply this algorithm to Twitter

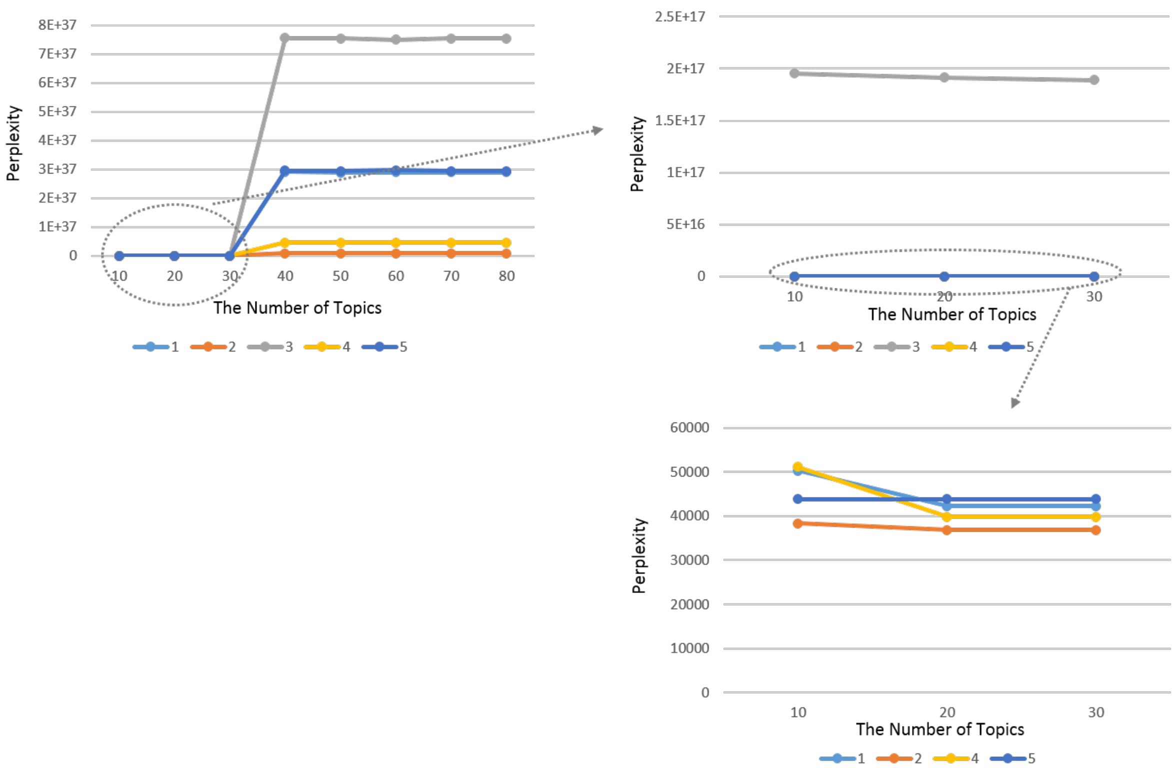

Figure 4.5: Perplexities of the test data for the models fitted with LDA. Each line corresponds to one of the folds in the 5-fold cross-validation

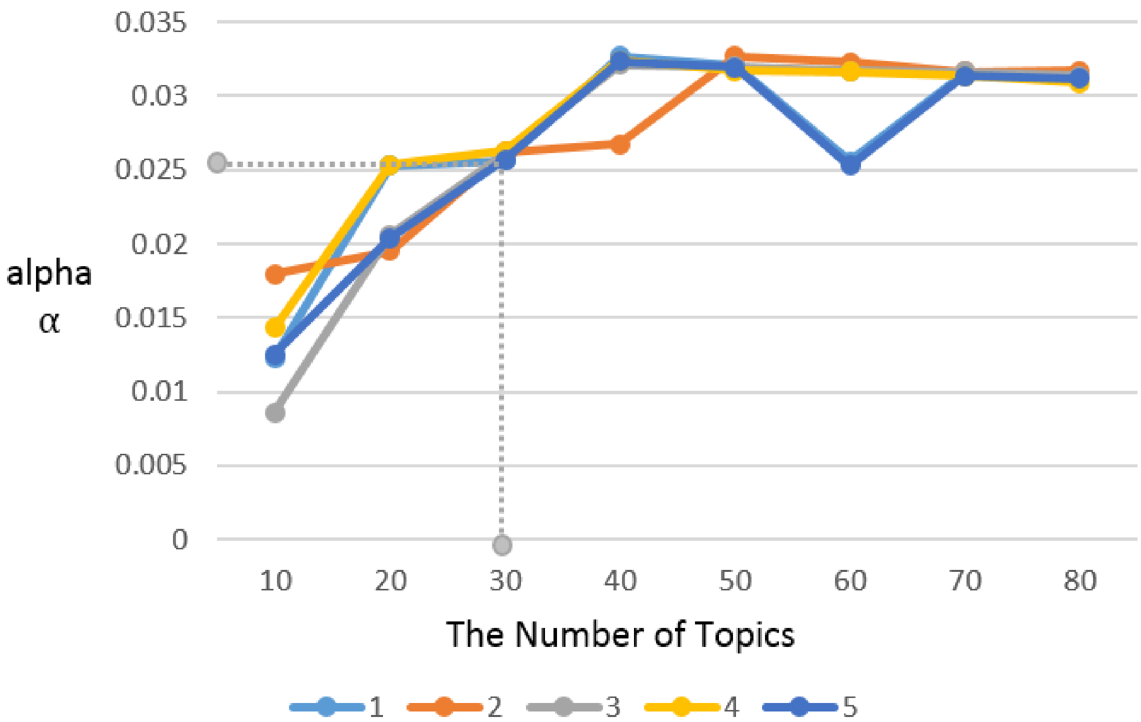

Figure 4.6: Estimated α values for the models fitted. Each line corresponds to one of the folds in the 5-fold cross-validation

data which collected in Great East Japan Earthquake. As explained, each topic is a probability distribution over a finite vocabulary of words, topics in a document are represented with words in order of the probability, which word w having probability in topic z, $p(w_n|z_n, \beta)$. Table 4.2 shows the 10 highest probability For example, 10 terms having highest probability for topic 1. are:

- japan
- you
- google
- nhk
- blog
- KEIKAKU TEIDEN (planned outage)
- CHIBA (chiba), KAIJI (start)
- and
- news
- TOKYODENRYOKU (Tokyo Electricity Company)

With those words, we estimate the topic as planned outage which caused by lack of electricity from Tokyo Electricity Company. When words occurs exactly same, such as CHIBA (chiba) and KAIJI (start), both of terms are recored. When we made the document as a collection of Tweets for 12 hours, it naturally contains a lot of retweets which means same sequence of the words belong to the corpus. So, if more than 2 words are same probability such as CHIBA (chiba) and KAIJI (start), those are the words which belongs to a tweet that retweeted a lot and not occur solely.

Also, each document is consists of mixture of topics θ, θ defines the topic mixture of the document, so intuitively θ_i (i represent topic) is the degree to which $topic_i$ appears in the document. Figure 4.7 shows topic changing from March 11 to 17. The vertical axis of the graph indicates the order of the degree of topics in a document and horizontal axis is documents from March 11 to 17 which represents time. We color topics that show a clear increasing and decreasing trend. Most of first and second highest degree θ_i topics follow this trend. It represents

Table 4.2: The 10 Highest probability words for each of 30 Topics

	Topic 1	Topic 2	Topic 3	Topic 4
1	japan	ヤシマ作戦ウエシマ作戦 発動名称倶楽部上島 リアクションギャグ資源品物	宮城岩沼下野郷新拓	ガソリン
2	you	隊員各位ヤシマ作戦 ウエシマ作戦発動模様 倶楽部資源品物	tag	ｱﾀﾞﾙﾄ
3	google	japan	earthquake	news
4	nhk	yomiuri	iphone投稿	how
5	blog	you	携帯電池ビックカメラ 無料開放飲料無料開放	japan
6	計画停電	the	仙台地震外危険 仙台トラスト避難	english
7	千葉開始	ヤシマウエシマ	japan	mau
8	and	大丈夫	大丈夫	tweeting
9	news	twitpic	google	quake
10	東京電力	news	車津波龍和工 株式会社倉庫トラック	地震一般情報要請 避難安否確認
	Topic 5	**Topic 6**	**Topic 7**	**Topic 8**
1	japan	フランス恋人太陽電池備え	大丈夫	公式拡散拡散
2	news	中国気	nhk	自衛隊
3	english	yomiuri	命危険日本国民 感謝気持ち	フジテレビ自重
4	protect	the	alchemist	japan
5	quake	自衛隊山被災死者不明目 予想遥か赤ちゃん自衛笑顔	放送報道中心お子様 アニメ子供番組経由 配信よろしけれお子様	東電職員
6	tweeting	よかっ	tweeting	yomiuri
7	how	大丈夫	japan	josef
8	person	南安否不明生存避難確認	tell	国民感情一覧
9	nhk	japan	yomiuri	you
10	ソフトバンク皆様無料 詳細被災方々冥福 方々お祈り	命危険日本国民感謝気持ち	media	chodo
	Topic 9	**Topic 10**	**Topic 11**	**Topic 12**
1	japan	itmedia	japan	大好き
2	nhk	日本全国皆さんやさし人気 持ち不安周り人不安気持ち ウルトラ	the	new
3	よかっ	nhk	you	person
4	person	関東炊飯電力供給電力命 関東主婦皆さん命電気	yomiuri	for
5	大丈夫	よかっ	宮城黒川大和牧場 牛乳無料譲り	live
6	itmedia	finder	and	nhk
7	震災方々冥福 お祈り国内無料	yomiuri	are	from
8	食料水	radio	被災報告震災実情	you
9	医者必要	tweeting	npo	this
10	命危険日本国民 感謝気持ち	命危険日本国民感謝気持ち	love	lifedeco

	Topic 13	Topic 14	Topic 15	Topic 16
1	yomiuri	news	the	命危険日本国民 感謝気持ち
2	大丈夫	衛生悪化感染発展	ｱﾀﾞﾙﾄ	japan
3	japan	炊き出しマップ	dvd	卒業撮影先生黒板 卒業生全員似顔絵
4	show	東京電力	you	親心配子供 情報ラジオネット
5	ヤシマ作戦ウエシマ 作戦発動名称倶楽部 上島リアクションギャグ 資源品物	地震一般情報要請 避難安否確認	and	are
6	for	yomiuri	out	chodo
7	are	被災男性笑顔大丈夫大丈 夫 津波体験再建声記者仕方 笑顔隣婦人格好	lifedeco	you
8	アメリカ	english	yomiuri	大丈夫
9	you	japan	good	nhk
10	and	world	wi白企画応募完了	news
	Topic 17	**Topic 18**	**Topic 19**	**Topic 20**
1	japan	ガソリン	無駄電気ウエシマ作戦 アオシマ作戦	奇跡スマイル
2	よかっ	twitter	the	ガソリン
3	this	ｱﾀﾞﾙﾄ	you	your
4	tbs	japan	love	japan
5	福島原発	奇跡スマイル	不謹慎関係拡散 ナカジマ作戦	衛生悪化感染発展
6	放射線原発関連パニック 原発周囲放射線基準 非常低く設定身体 影響パニック不安 避難被害心配落ち	衛生悪化感染発展	フジテレビ中継	自分被災
7	people	your	リラックマェ	and
8	foreigners	are	news	nhk
9	love	world	綾小路きみまろ お腹自分身体重く 荷物自分お腹	news
10	video	news	よかっ	茨城東茨城 茨城下石崎

	Topic 21	Topic 22	Topic 23	Topic 24
1	japan	自衛隊	all	the
2	まわり水没身動き 年寄り多くヤバイ	生存名簿	計画停電	you
3	nhk	マジ背中年寄り腕年寄り 年寄り普通	東北地方太平洋沖地震	never
4	tbs	フジテレビ自重	japan	walk
5	people	japan	東北地方太平洋沖地震 非常有効活用関東地区 計画停電自体影響想定活用	japan
6	福島原発	yomiuri	福島原発	your
7	you	エロゲ板	for	嬉し涙
8	東京電力発電電力 供給需要見通し発表	西日本東日本関係 西日本平均 電気月西日本	よかっ	and
9	良かっ	you	blog	love
10	and	よかっ	bloomberg	ヤシマ作戦ウエシ マ作戦 発動名称倶楽部上 島 リアクションギャグ 資源品物
	Topic 25	Topic 26	Topic 27	Topic 28
1	person	自衛隊	大丈夫	大丈夫
2	you	東電職員	person	車津波龍和工 株式会社倉庫トラッ ク
3	宮城岩沼下野郷新拓	フジテレビ自重	you	岩永美保
4	よかっ	comic	nhk	google地震あと 連絡家族友人
5	iphone投稿	japan	earthquake	finder
6	google地震あと 連絡家族友人	国民感情一覧	google災害情報	フォロワ拡散全国 非難場所一覧災害 掲示板ウィルコム 災害掲示板
7	google	生存名簿	よかっ	iphone投稿
8	車津波龍和工 株式会社倉庫トラック	お父さん原発避難原発人 母さん原発人自分必死	命危険消防電話 優先各地電話回線 混合安否確認電話	person
9	ドコモ	西日本東日本関係西 日本平均電気月西日本	japan	maps更新
10	japan	マジ背中年寄り腕年寄り 年寄り普通	発電栃木電気常用 発電電気アウト首都 電気使用最低限会話	警告大震災最後 最大災害悪化人災

	Topic 29	Topic 30		
1	事件現場応援コジマ作戦	nhk		
2	yomiuri	how		
3	無駄電気ウエシマ 作戦アオシマ作戦	日本全国皆さんやさし人気持ち 不安周り人不安気持ちウルトラ		
4	japan	tweeting		
5	事件現場コジマ作戦	大丈夫		
6	仙台市民緊急車両給油	日ユ募金かなり額懐中間搾取		
7	フジテレビ中継	命危険日本国民感謝気持ち		
8	roadracer	are		
9	google	just		
10	itmedia	yomiuri		

that as the situation changes with natural disaster, sequence of nuclear accidents and secondary earthquakes, the public focused on risks which catches their attentions.

Here, we explain what kind of topic actually observed on 7 days. Afternoon on March 11, when the first earthquake occurred, Topic 28,25 and 3 is prominent.

- Topic 28 Information about Damaged area (requesting rescue) from Earthquake and Tsunami
- Topic 25 Information about Damaged area (requesting rescue) from Earthquake and Tsunami (Miyagi)
- Topic 3 Information about Damaged area from Tsunami (Sendai)

We can see that people's concerns are targeting the disaster area that directly damaged by earthquake and tsunami. They exchanged information about safety confirmation of victims who lost contact or isolated.

On March 12, the topics cover more wide subjects such as a lack of electricity, which is the risks affecting whole east Japan In addition, topics which contains emotional feeling 'anxiety' are also observed.

AM March 12

- Topic 27 Information about Damaged area (Safety Confirmation) and Unstable Electric Power Supply
- Topic 21 Information about Damaged area (Safety Confirmation) and Unstable Electric Power Supply

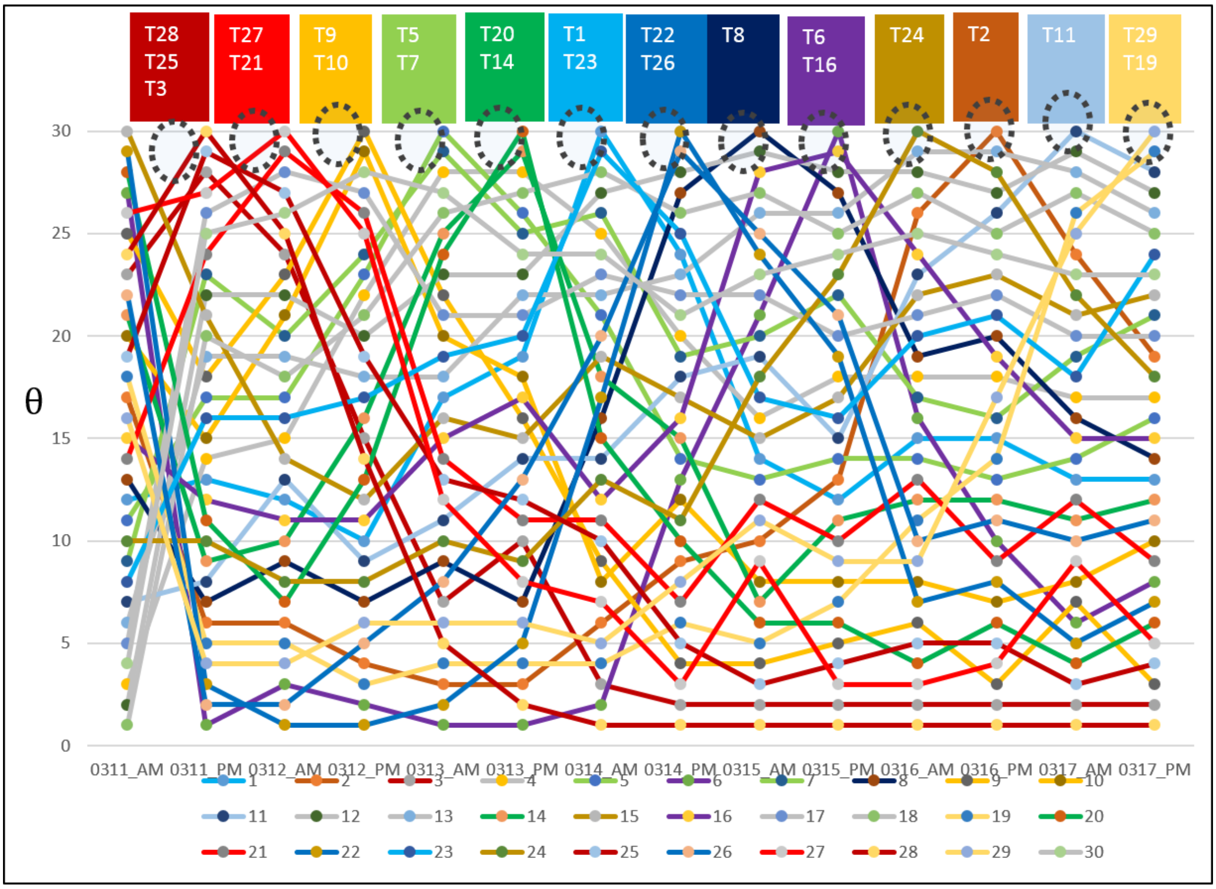

Figure 4.7: Topic Changing from March 11 to 17 (frequently changing topics)

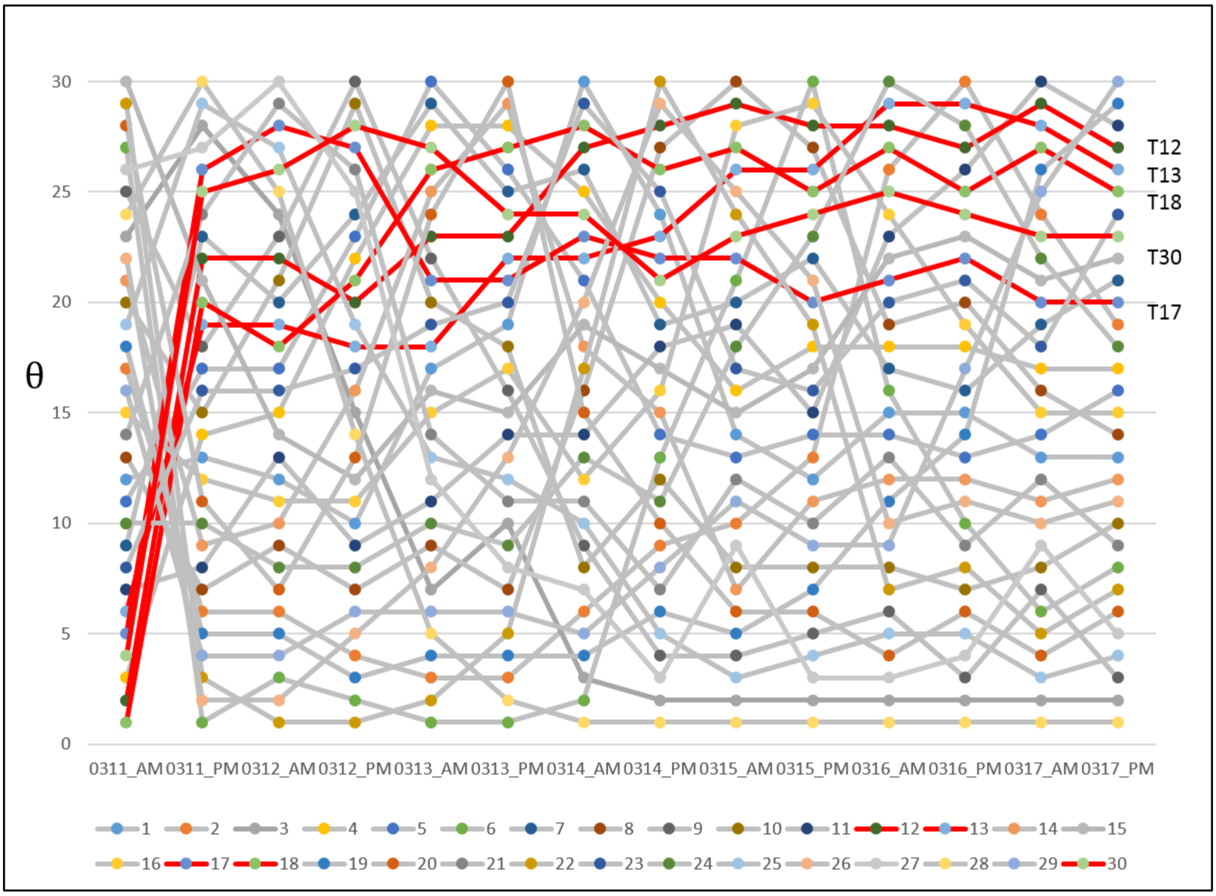

Table 4.3: Topic Changing from March 11 to 17 (not frequently changing topics)

PM March 12

- Topic 9 Information about Damaged area (Lack of goods and doctor)
- Topic 10 Anxiety about the Society and Thankfulness

Both of positive and negative emotional terms are observed, as positive terms 'Thankfulness' and negative terms 'anxiety'. As the situation went worse after the accidents of Nuclear power plant, people's concerns were diversified which nearly lead society to panic comes with anxiety.

On March 13, people are focusing on information regarding disaster areas especially about sanitary conditions in there:

AM March 13

- Topic 5 Information about mobile communication infra (free wireless internet data service)
- Topic 7 Thankfulness

PM March 13

- Topic 20 Information about Damaged area (Lack of goods and Bad sanitary conditions)
- Topic 14 Information about Damaged area (Bad Sanitary Conditions and Safety Confirmation)

On March 14, new topics comes up such as a planned outage or committing Japan Self-Defence Force troops to disaster areas

AM March 14

- Topic 1 Planned outage (East Japan)
- Topic 23 Planned outage (East Japan)

PM March 14

- Topic 22 Committing Japan Self-Defence Force Troops and Electricity Conditions in West Japan
- Topic 26 Committing Japan Self-Defence Force Troops and Electricity Conditions in West Japan

On March 15, the public is newly focusing on the nuclear power plant accident and Response of Foreign Country.

AM March 15

- Topic 8 Committing Japan Self-Defence Force Troops

PM March 15

- Topic 6 Response of Foreign Country
- Topic 16 Thankfulness

On March 16 and 17, people discuss power saving issue, which compared with power saving behavior (YAJIMA operation) in famous animation in Japan. AM March 16

- Topic 24 Power-Saving Movement
- Topic 13 Power-Saving Movement

PM March 16

- Topic 2 Power-Saving Movement
- Topic 13 Power-Saving Movement

On the other hand, Figure 4.3 shows some of topics having not frequent changing degree in those 7 days. Topic 12, 13, 18, 30 and 17 are never be prominent topics in 24 documents but maintain the degree higher than 20 during 7 days. Especially, Topic 30 and 17 represents 'anxiety' regarding nuclear power plant or whole Japanese society.

- Topic 30 Anxiety regarding Whole Japanese Society
- topic 17 Anxiety regarding Nuclear Power Plant Accident and Societal Panic

Those topics clearly represent that the public have an underlying emotions during the disaster and it is an anxious about whole Japanese Society threatened by nuclear power plant accident.

Even though people's concerns are changing and diversify as time goes, we could find that there are an underlying emotions people have regardless of situation changing. Topics representing emotions ('anxiety') is observed (Topic 30, 17 and 10), for the first time, on 12 March, when significant explosion from the Fukushima Daiichi nuclear power plant occurred. And it lasted util 17 March. So, it means that the 'anxiety' of the public has an intrinsic attribute that not a temporary changing but continuous.

4.6 Implication

This analysis shows that societal unstability is observed in the early stage of the disaster. Nuclear accident on 12 March could be considered as a direct reason for it. Risks being exposed to radiation was strongly perceived by the public. And it seems it does not disappear immediately. On the surface, the concerns that the public have are easily changed as new events occur, even the new event has less risk than former. But, more importantly, there are also fundamental concerns regardless of the events in disasters. In that sense, once the public perceived the risk, it is rarely changed or disappears. And according to the review in chapter 2., the public's risk perception is strongly related to potentials or impact to future generations not the fatality of this stage. Announcement toward the public usually contains only the objective facts that explain the disaster situation of the stage. However, for the public it sometimes intractable to understand the contents especially when they have to handle extremely complicated risk such as nuclear plant accident. And also that's why the public rely on the announcement of the organization such as the government. Organizations as a provider of the risk information in disasters, it is crucial that they provide information embracing the concept of potential risks to conduct the crisis communication effectively.

4.7 Conclusion

To facilitate efficient communication in the event of a future catastrophe, it is important to clarify concerns of each participant of communication. In this chapter, we investigate the concerns of the public in The Great East Japan Earthquake who is one of the important participants in communication. We shows variety concerns and how those concerns are changing in disaster. We develop the application to figure out what kind of concerns the public have during the disaster.

For applying LDA model to Twitter data. With LDA model, we examine the contents of utterances in the Twitter corpus by applying LDA algorithm, which identify topics of Twitter corpus by estimating the latent variables of data collection.

Various kinds of concerns have been observed during 7 days. People's concerns diversify as time goes, especially from March 12 when significant explosion from the Fukushima Daiichi nuclear power plant occurred. The public's concerns changes as the disaster situation changes, but in the same time, there are underlying concerns that unlikely changes which is 'anxiety' regarding radiation issue.

There is still a range of limitations needed to be solved in the future works. For more accurate observation, it would be necessary to examine a longer period to know the feature of fundamental concerns 'anxiety'. It is necessary to investigate how the 'anxiety' changes after the catastrophic condition to subside. To solve this problem, we would like to improve the result with applying multiple LDA models to investigate longer period data. For concatenate several LDA models of each stage, Dynamic Topic Models could be applied, which consider the prior distribution from previous stage in the model [Blei and Lafferty, 2006][Fujimoto et al., 2013].

Bibliography

[1] David M Blei, Andreq Y Ng, and Michael I Jordan. Latent Dirichlet Allocation. *Journal of Machine Learning Research*, pp. 993-1022, 2003.

[2] David M Blei and John D Lafferty. Dynamic topic models. *Proceedings of the 23rd international conference on Machine learning (ICML 2006)*, pp. 113-120, 2006.

[3] Kevin R Canini, Lei Shi, and Thomas L Griffiths. Online Inference of Topics with Latent Dirichlet Allocation. *Proceedings of the International Conference on Artificial Intelligence and Statistics*, (1999), pp.65-72, 2009.

[4] James Dickey. Multiple Hypergeometric Functions: Probabilistic Interpretations and Statistical Uses. *Journal Of The American Statistical Association*, 78(383), pp.628-637, 1983.

[5] Hiroshi Fujimoto, Takahiro Hara, and Shojiro Nishio. Online Topic Analysis with Online Optimal Smoothing and Dynamically Evolving Vocabulary. (5), pp.1212-1221, 2013.

[6] Thomas L Griffiths and Mark Steyvers. Finding scientific topics. *Proceedings of the National Academy of Sciences of the United States of America*, (101) Suppl, pp.5228-5235, 2004.

[7] Bettina Gru n and Kurt Hornik. topicmodels : An R Package for Fitting Topic Models. *Journal of Statistical Software*, 40(13), pp.130, 2011.

[8] Toshiko Kikkawa. Risk Communication. *Japanese Journal of Health Physics*, 35(4), pp.485-490, 2000.

[9] kiyoshi KOBAYASHI. Unforeseen Risk and Planning Perspectives. *Journal of Japan Society of Civil Engineers, Ser. D3 (Infrastructure Planning and Management)*, 69(5), I 1I 14, 2013.

[10] Arthur Lupia and Mathew D. McCubbins. *The Democratic Dilemma: Can Citizens Learn What They Need to Know? (Political Economy of Institutions and Decisions).* Cambridge University Press, 1998.

[11] Paul Slovic. Perception of risk. *Science*, 236(4799), pp. 280-285, 1987.

[12] Paul Slovic. The risk game. *Journal of Hazardous Materials*, 86(1-3), pp.17-24, 2001.

Chapter 5

Measurement of Disaster Anxiety of the public

5.1 Introduction

This chapter investigates the role of sentiment analysis, which has recently become popular in the field of crisis communication research. To review disaster management policy and foster proper communication in large-scale natural disasters, it is important to determine the publics perceptions of and sentiments about natural disaster events. Sentiment analysis is essential to numerous pursuits in knowledge engineering and numerous application areas embracing business, the health sector, the Internet, social networking, politics, and the economy.

To facilitate efficient communication in the event of a future catastrophe, it is important to understand the contents of utterance of the public who is the one of key players in crisis communication. Specifically, we focus on the perception of risk during a disaster. By utilizing 'anxiety' which is a key variable that affect a social disorder, we evaluate the 'anxiety' which is changing on 7 days with index we propose in this study.

This chapter is organized as follows. 5.2 explains by means of the corpus the concept of crisis communication containing anxiety, and highlights related studies. 5.3 explains the data. 5.4 describes the methodology and application proposed in this study. 5.5 presents corpus-based contents analysis of crisis communication during the disaster, investigating the anxiety levels proposed in 5.4. 5.6 explains implication. Finally, 5.7 describes conclusions and future work regarding this research.

5.2 Basic Idea

5.2.1 Anxiety as risk perception of the public

Crisis communication includes evidence of participants thoughts, behaviors and emotions. Therefore, several indexes could be considered for evaluating evidence that would be helpful in understanding crisis communication. For example, attention could be one of the indicators explaining how people react when considering risk in the disaster. According to Lupia et al. [Lupia et al., 1998], since the proper choice of information relates closely to gaining higher utility in the future, it is desirable to make effective use of attention as an indicator for measuring how people make judgments against various and ever-changing risks. For investigating attention, a real-time basis indicator presenting the behavior of the public could be applied. Nowadays there are various data sources on a real-time basis, such as Google Trends, a real-time daily and weekly index of the volume of queries that users enter into Google [Choi and Varian, 2012]. The query index is based on query share, the total query volume for the search term in question within a particular geographic region divided by the total number of queries in that region during the time period under examination [Choi and Varian, 2012]. The maximum query share in the time period specified is normalized to be 100, and the query share at the initial date being examined is normalized to be zero.

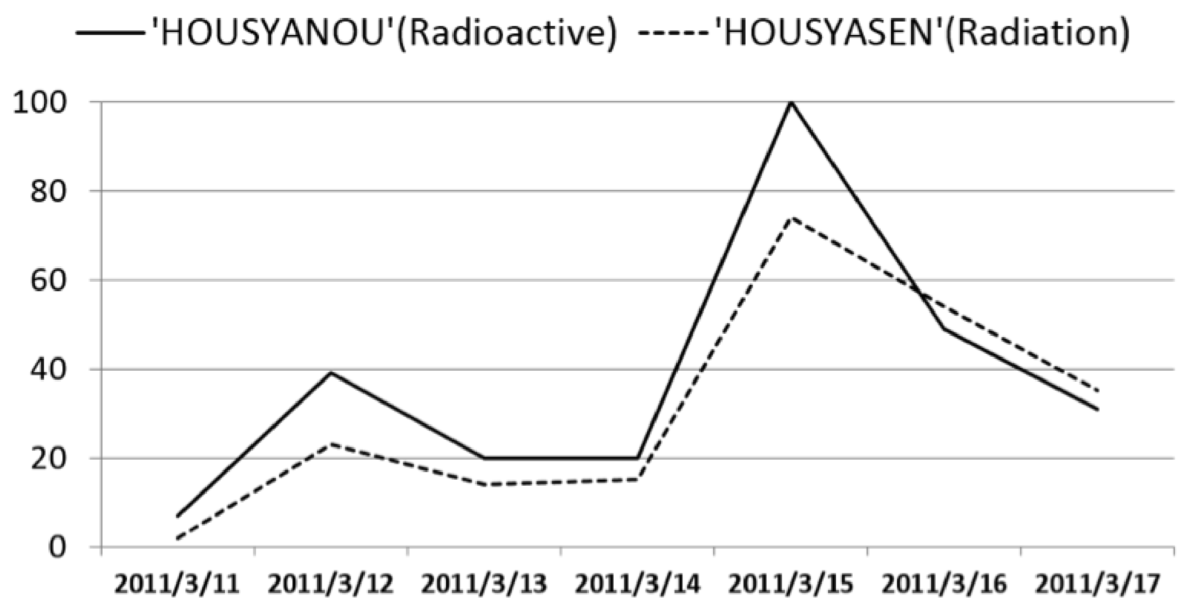

Figure 5.1: Time Series of Variation of Volume of Queries (Google Trends)

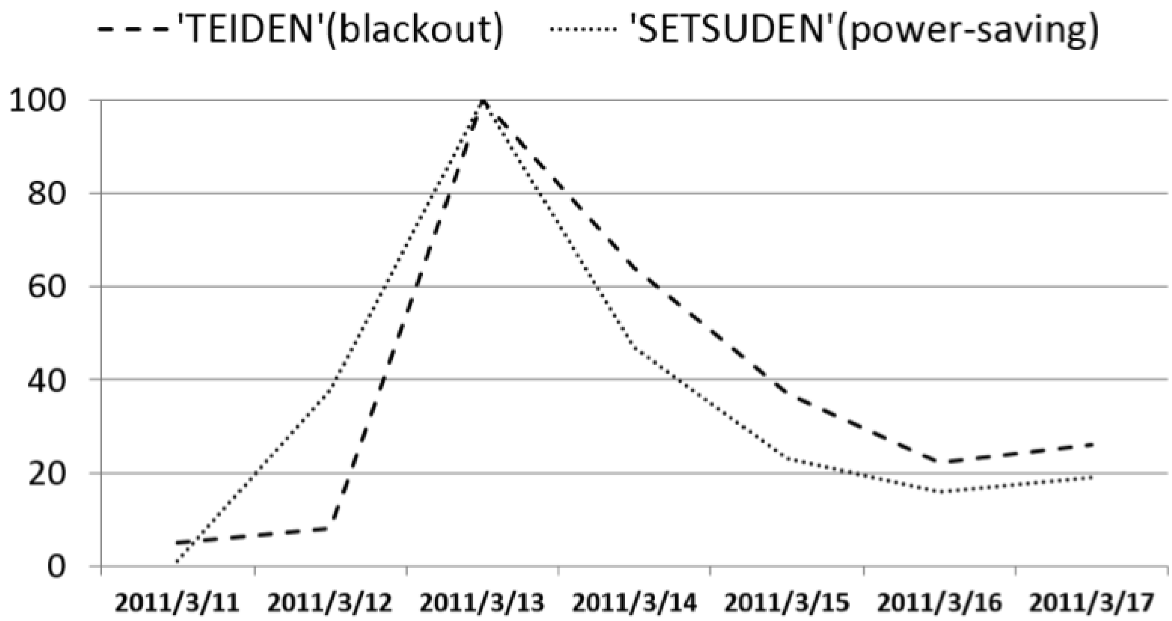

Figure 5.2: Time Series of Variation of Volume of Queries (Google Trends)

Figure 5.1 shows the time series variation associated with volume searching for HOUSYANO (radioactivity) and HOUSYASEN (radiation) in Japan during the week following the disaster. Both show a trend similar to that observed above: increasing on March 12, and then again on March 15 in contrast to March 13 and March 14 which represents relatively low. Interestingly, March 13 and 14 was the day when people paid attention to risks such as TEIDEN (blackout) and SETSUDEN (power-saving) (Figure 5.2). This observation indicates that people paid attention to risks closely related to the occurrence of accidents at the time. This implies that public attention is temporarily fixed, following accidents or problems.

Besides, as an indicator for measuring the level of attention, fastness was suggested, which refers to how fast people react to considerable risks [Baek et al., 2013]. The authors showed that people pay more attention to risks such as radiation after a nuclear power plant accident. Those implications can be evaluated by frequency-based methodology, as shown above.

But in a crisis, when people make a risk perception, they not only pay attention to the accident itself, but also perceive the overall disaster situation by examining the weight of the risks and their ability to control them. If they judge there are no considerable threatening elements, they feel relief, but if it turns out that the situation is out of their control, they feel anxiety.

In this study, the meaning of anxiety is limited to a certain range that is not an emotion related to a social event followed by activation of the sympathetic nervous system, such as increasing heart rate or blood pressure, but a negative feeling which is not only temporary [Yamazaki et al., 2004]. Anxiety, in psychology, refers to unpleasant emotions that include concern, distress, worry, and uncertainty about the results of an event, situation, or deadly catastrophe, vague anticipation of danger, or presentiment that danger or pain could happen in the future, with relatively high ambiguity or uncertainty as to what will happen, and when [Seiwa, 1999] [Tsuru, 1981]. These definitions commonly highlight the intrinsic feature of anxiety as containing the possibility of a future encounter with danger. Research studies on rumor have agreed that anxiety is an important factor in understanding how information spreads in a disaster. They argue that anxiety and informational ambiguity are key variables that affect a rumor mongering condition in abnormal communication under extreme events such as natural disasters [Stieglitz and Dang-Xuan, 2013][Oh et al., 2010]. In particular, Anthony [Anthony, 1973] discusses the link between the transmission of rumor and the level of anxiety. The results of experiments show that rumor is more frequently transmitted in high anxiety groups than in low anxiety groups. As rumor is interpreted as a collective transaction in which many people offer, evaluate, and interpret information, and from which they predict something [Oh et al., 2010], it is a type of linguistic expression of risk perception through the rationalizing of ambiguous information. Therefore, it is important to estimate the level of anxiety in order to investigate contents of crisis communication. Crisis communication contains evidence of risk perception indicating interpretation of peoples own capability to deal with the crisis and the emotions arising from thoughts such as anxiety. In the Great East Japan Earthquake, social media tools were used widely for individual levels of crisis communication. With the advent of social media, a place came into being where individuals could conduct crisis communication and convey their perception of risk during disasters.

5.2.2 Utility of the Twitter corpus as data for evaluating public sentiment

Twitter is responsible for a large stream exchanging subjective information between its users regarding disaster risks in crisis. Since it contains the individual context of communication, the corpus of Twitter data has been highlighted in this study.

For clarification, Twitter is a form of social media that allows its users to send short messages (140 characters or less) to others [Kireyev et al., 2009]. Its user can decide that the contents are visible to only a limited group of users or in general. It is a new type of chat service based on real-time platforms, and strict sorting of the vast amount of information produced based on the social relationships of users is not required. Starbird et al. [Starbird et al., 2010] pointed out that unlike Wikipedia, content passed through Twitter is short-lived; therefore, it cannot be discussed, verified, or edited. While most social media have places for interaction, interaction on Twitter occurs in and on the data itself, and through its distribution, manipulation, and redistribution. Information is part of a life cycle of generation, derivation, synthesis, and innovation that combines skills with information production to shape the information space [Baek et al., 2013]. Because of the unique characteristics of information interaction, information diffusion is determined by its users who decide what is valuable and what is not.

The societal context of risk perception measures during the disaster was examined based on Twitter data from public who addressed their conditions through social media. It contains both cross-societal and local context. With regard to local context, Twitter corpus includes information regarding locations of users and times when users tweet with their mobile communication devices. Twitter provides information valuable in understanding the areas affected by a disaster. Baek et al.[Baek et al., 2013] pointed out that risk information disseminated by the government and news media tends to ignore the local context, as evidenced by the fact that crisis communication geared to the local level had not been observed in prior disasters. Communication by means of social media can overcome this problem. Another feature of Twitter corpus is that it contains very little lexical redundancy in a single tweet and the distribution of information is rapid [Kireyev et al., 2009]. By investigating the contents of Twitter, it enables researchers to capture the early phases of crisis communication. This feature also encourages the government to utilize for distribution of information via Twitter (i.e., real-time risk information). Twitter is a social media containing subjective assessment created by individuals, while newspapers and Wikipedia are inter-subjective social media that publish information through cross-validation. Therefore, this study considers Twitter corpus as an important means of information exchange within crisis communication.

In the Great East Japan Earthquake, the status of disaster area was disseminated rapidly. Just nine minutes after the first tremor of the earthquake, the Division of Emergency Management in Kesennuma City, Miyagi prefecture (Twitter account name: bosai_kesennuma) disseminated information via Twitter. Messages were spread by Twitter users who received messages. Immediately, a large-scale chain of communication was created. In this way, the situation regarding affected areas became known and discussed instantly throughout the world [Kiyono, et al., 2013].

5.3 Data

Table 5.1: The Outline of Sample Data

The Total number of Tweets							
11-Mar	12-Mar	13-Mar	14-Mar	15-Mar	16-Mar	17-Mar	Total
26099910	27315744	23065384	25404539	25338601	24368393	22465493	174058064
The number of Tweets which include the word "HOUSYA (Radiation)"							
11-Mar	12-Mar	13-Mar	14-Mar	15-Mar	16-Mar	17-Mar	Total
72986	378322	124573	141684	482370	372543	221830	1794308
The Total number of Literatures (TEPCO)							
11-Mar	12-Mar	13-Mar	14-Mar	15-Mar	16-Mar	17-Mar	Total
11	30	28	27	25	22	32	175

Twitter data was provided by Twitter Japan via Project 311 (The Great East Japan Earthquake Big Data Workshop Project 311, 2015). It was collected over the seven days from March 11 (about an hour before the earthquake occurred) to March 17, 2011. Tweets posted during those 7 days and also those written in Japanese were included. The data comprises tweet IDs, user IDs, time and tweet contents. In this analysis, a 5% sample from the provided data was extracted randomly. Table 1 shows the total number of tweets in the sample dataset. As HOUSYA (radiation) had been chosen as a word associated with risk perception in this analysis, Tweets including the word HOUSYA (radiation) were chosen from the sample data. We carried out the analysis only with tweets including HOUSYA (radiation). The number of tweets including the word is given as data in Table 5.1. As a representative for the government regarding the accident at the nuclear power plant, Tokyo Electric Power Company (TEPCO) was selected to prepare announcements. TEPCO's corpus was gathered through the webpage entitled 'Press Release' subpage of TEPCO [TEPCO, 2011] and transformed as text data.

Figure 5.3 shows the time series for variations in the number of tweets in Japanese that contained HOUSYA (radiation) during the 7 days after the earthquake. Following the explosion accidents from the Fukushima Daiichi nuclear power plant, the number of tweets that included the term HOUSYA (radiation) was significant on March 12. Tweets containing the expression

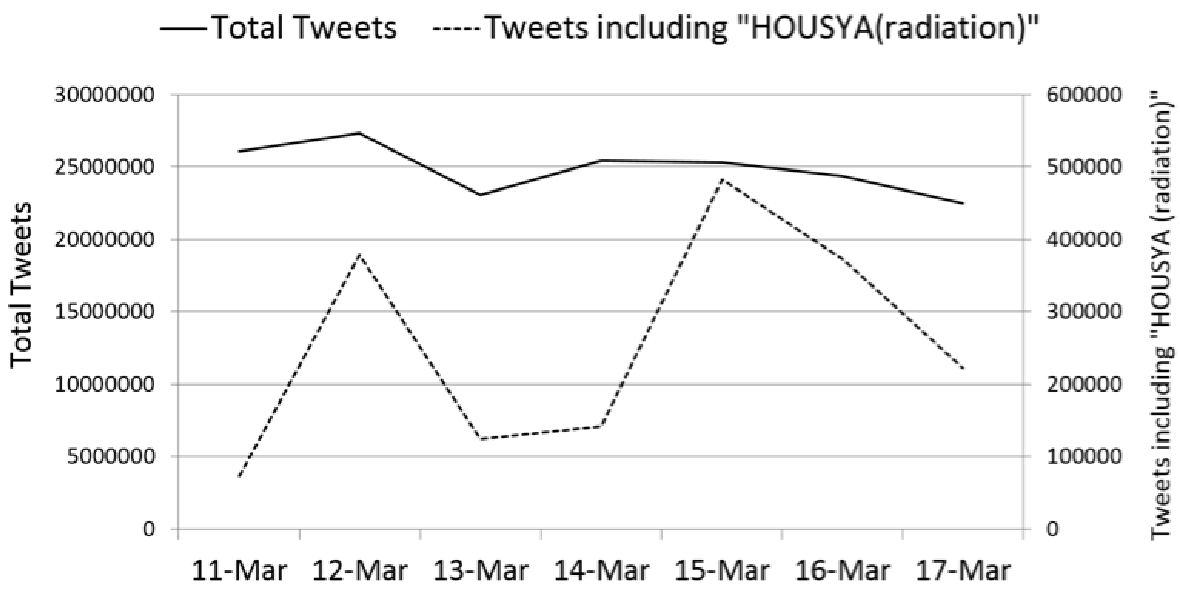

Figure 5.3: Time Series of Variation of Volume of Tweets including HOUSYA (radiation)

again increased sharply on March 15, when a series of three explosion occurred at the nuclear power plant. On the other hand, the total number of tweets did not change comparatively.

5.4 The Methodology and The Application

In this study, the application has been developed using the corpus of Twitter to measure a level of anxiety by investigating risk perception during a disaster (Figure 5.4). To examine anxiety levels of people and societies suffering from disasters, it is important to investigate how people perceive risks during disasters. As emotional perception contains an investigation of states and attitudes toward a specific subject, estimating semantic orientations of risk perception strategies could be used as an indicator for anxiety. When the corpus is very large, it is important to identify the semantic orientations of words automatically. In other words, it is necessary to determine whether the emotional risk perception is positive or negative automatically. In this study, level of anxiety was measured relative to risk perception determined by a list of semantic orientations for Japanese proposed by Takamura et al. [Takamura et al., 2005] according to use of the spin model (Figure 5.5). Terms representing risks or risk perception were extracted with text mining technology. In this analysis, co-occurrence terms occur with the word denoting the

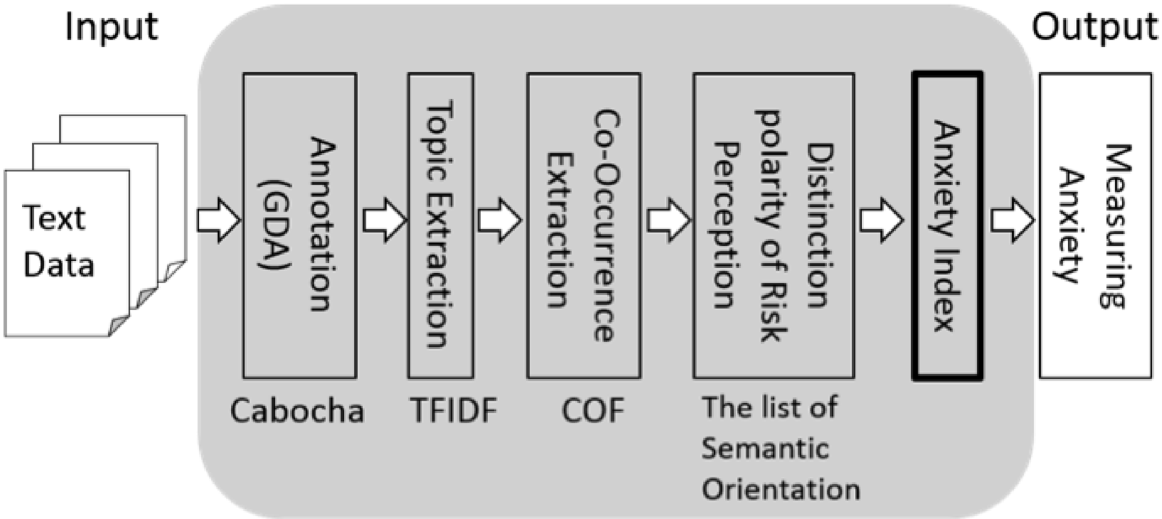

Figure 5.4: Outline of the Application

risk were regarded as risk perception. As mentioned, the polarity of each term of risk perception was weighed referring to the list of semantic orientations. Finally, Anxiety Index denoting negativity for perception of riska ratio for risk and risk perceptionwas proposed as a possible index for the quantitative measurement of anxiety. Thus, through a statistical analysis of data from actual communications, it would be possible to detect the degree of the spread of anxiety.

Figure 5.4 shows the outline of the application. First, it was necessary to tag every term in the text collection of natural language to apply computational linguistics or corpus linguistics. In this study, Cabocha (a Japanese dependency structure analyzer) was used to produce Global Document Annotation (GDA) tagging, which is processed language to be annotated and tagged by attributes.

Second, topics of collected data were decided according to the TFIDF. They could be defined as keywords or significant concerns of the public. Among terms with high TFIDF, HOUSYA (radiation) was chosen as a topic related closely to risks originated from the Fukushima Daiichi nuclear accident. A lot of Japanese people show high anxiety regarding the accident.

Then, the co-occurrence frequency of the term was extracted from the corpus to investigate expressions or perception regarding the risk. Co-Occurrence Frequency is the frequent occurrence of two terms from a text corpus. So, Co-Occurrence Frequency of terms such as radiation could be presented as a set of two elements, Co-Occurrence term and frequency, which is counted if it occurs with the term radiation in a sentence. The co-occurrence terms were categorized as positive, negative, and indistinguishable words according to the list of semantic orientations

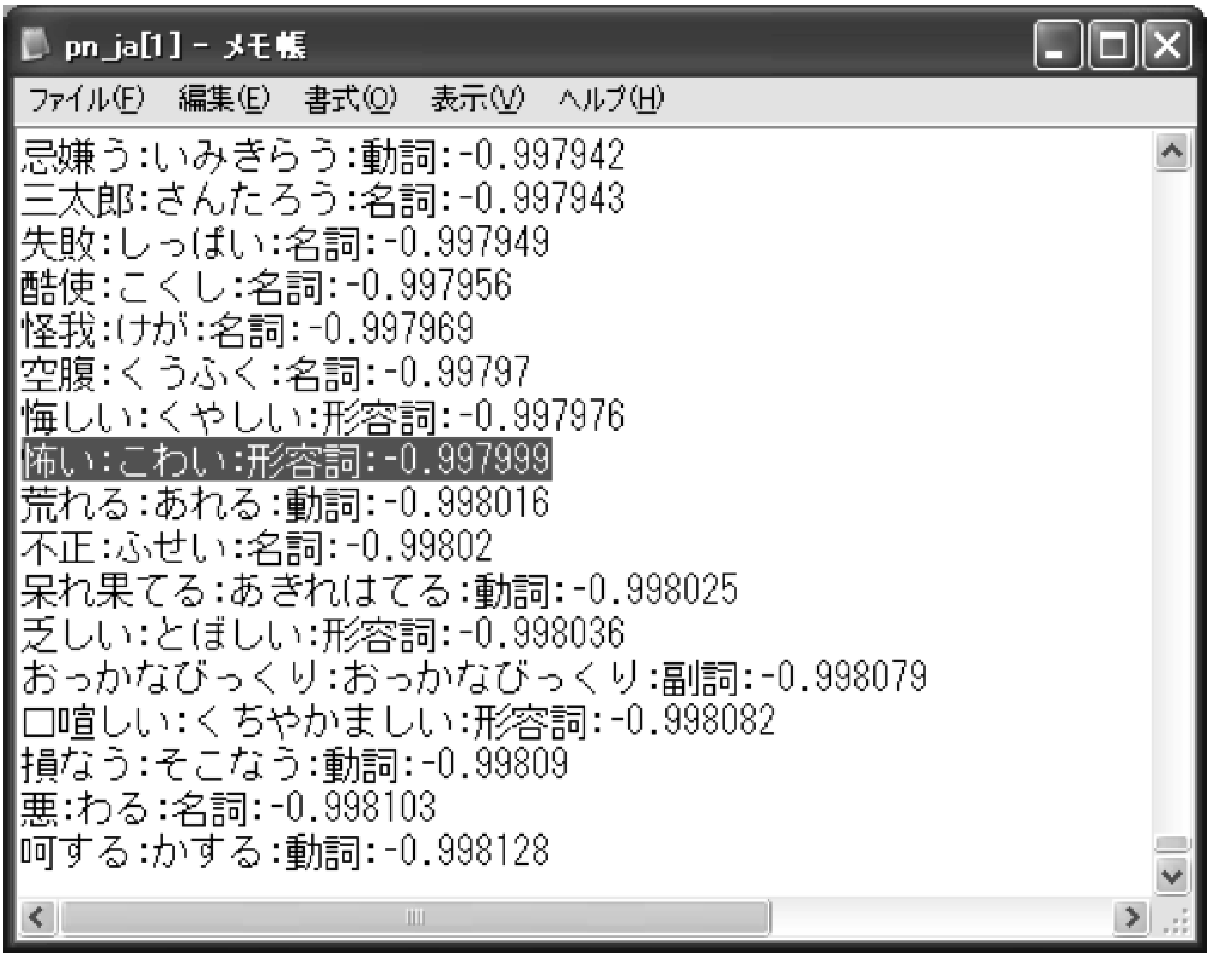

Figure 5.5: The List of Words and Semantic Orientations for Japanese

Source: Takamura et al.(2005)

[Takamura et al., 2005]. Terms labeled as negative words are considered as negative perception regarding the risk.

Finally, measuring anxiety was achieved by applying an Anxiety Index, represents how negatively people estimated the risk.

In the analysis, We apply the Anxiety Index to the corpus of governmental announcements and Twitter data to evaluate anxiety levels of governmental agencies and citizens.

5.5 Measuring Anxiety using Anxiety Index

To measure the level of anxiety using the Twitter corpus, an Anxiety Index has been proposed that represents the degree to which a risk (e.g., earthquake, tsunami, or radiation) is perceived negatively. The proposed Anxiety Index is defined with the Equation below 5.1 .

$$Anxiety\ Index = \frac{COF^N}{TF} \tag{5.1}$$

It is defined as the ratio of term frequency ($TF_{radiation}$) of a concerned risk to co-occurrence frequency of negative terms (COF_N). Co-occurrence terms are classified as a polarity (positive or negative) of terms according to semantic orientations[Takamura et al., 2005]. As co-occurrence terms are considered as a risk perception of a concerned risk, anxiety level is decided by quantity of negative co-occurrence terms per term frequency. Even if the crisis communication corpus were to include numerous references to HOUSYA (radiation) ($TF_{radiation}$) and negative perception terms (COF_N) were not great, people's anxiety levels for HOUSYA (radiation) would be interpreted as relatively low. On the other hand, even if the crisis communication corpus were to include only a few references to the term, a risk perception with a greater amount of negative terms would indicate that anxiety levels about HOUSYA (radiation) were relatively high.

By utilizing the Anxiety Index proposed, Figure 5.6 shows the co-relationship between anxiety and time of Twitter corpus.

In this study, COF_N with HOUSYA (radiation) were counted for each of seven days. To analyze the co-relationship between anxiety and time, we applied the Anxiety Index for anxiety variable y and set a time-serious variable $x \in (1, .., 7)$ as a day. Thus, the co-relationship could be stated formally as a linear regression model. In the figure, the regression coefficient of

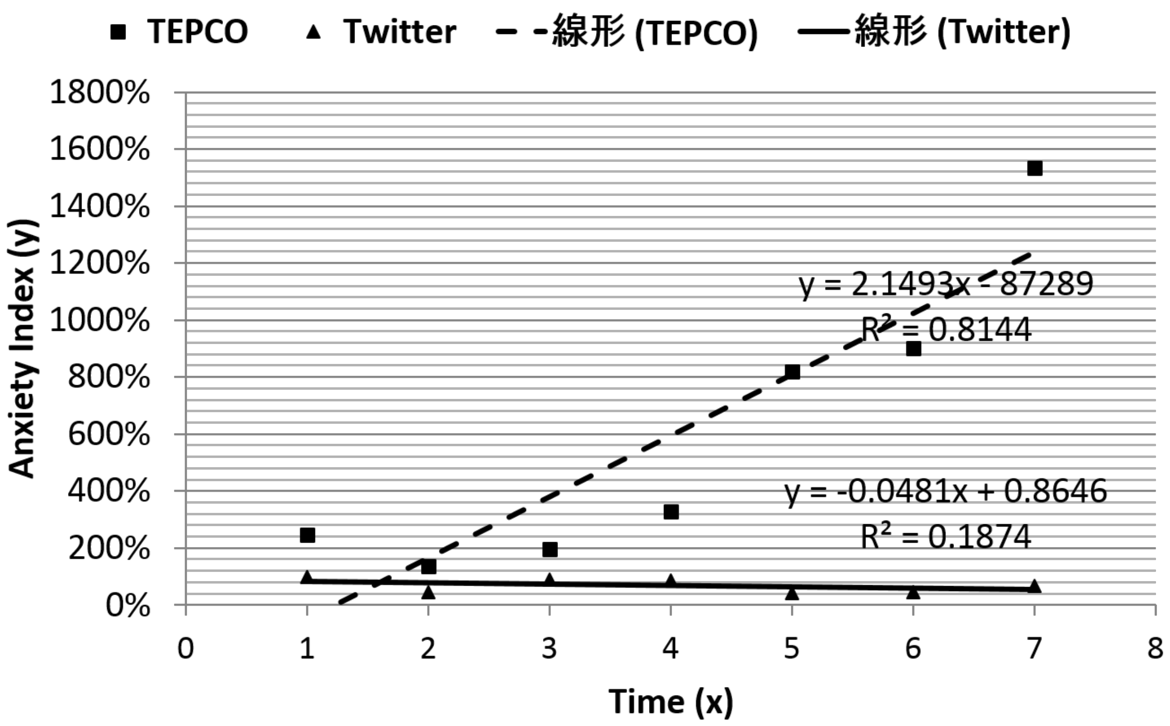

Figure 5.6: Time Series Variation of Anxiety

government (TEPCO) is 2.1493; it is positive and its coefficient of determination (R^2) is over 0.8. That means government (TEPCO) anxiety regarding 'HOUSYA (radiation)' increased as time passed. On the other hand, Twitter's regression coefficient was negative, but its R^2 was low. It means the co-relationship between anxiety levels of Twitter users and time was not significant.

To illustrate what terms were included in co-occurrence terms, Table 5.2 shows the 40 highest ranked co-occurrence frequency terms which is negative for both TEPCO and Twitter. In judging or assessing 'HOUSYA (radiation)', TEPCO tends to use scientific terms such as 'BUSHITSU (material object)', 'HANDAN (decision)', and 'CHOUSA (investigation)'. On the other hand, Twitter users tend to use terms related to emotional and safety, such as 'BYOUIN (hospital)', 'JINTAI (human body)', 'SINPAI (worry)', and 'KOWAI (feel fear)'. This table represents that anxiety was not conveyed through TEPCO announcements, even though negative terms were used. On the other hand, Twitter users' terms expressed their anxiety.

To deepen the investigation, terms were examined individually. Among the 40 highest ranked co-occurrence frequency terms, six terms were observed mutually: 'BUSHITSU (material object)', 'HOUSYUTSU (release)', 'EIKYO (influence)', 'SOKUTEI (measurement)', 'TSUJYO

TEPCO							
1	HATSUDEN (generating electricity)	11	SOKUTEI (measurement)	21	GOZEN (morning)	31	SU (prime (element))
2	TSUJYO (usually)	12	KYOUKAI (abuttal)	22	KAKUNOU (containment)	32	JISHI (implementation)
3	GENSHIRYOKU (nuclear power)	13	HOUSYUTSU (emission)	23	HASEI (emergence)	33	KANRYOU (completion)
4	BUSHITSU (material object)	14	GENSHI (atom)	24	ICHIBU (compartment)	34	EIKYOU (influence)
5	ATAI (value)	15	GAIBU (external)	25	KUUKI (air)	35	JISYOU (event)
6	JYOUSYOU (increasing)	16	ATSURYOKU (pressure)	26	ANZEN (safety)	36	TAISAKU (measures)
7	GOGO (afternoon)	17	YOUKI (flask)	27	KOUKA (drop)	37	CHOUSA (investigation)
8	SOCHI (taking measures)	18	HAIKI (exhaust air)	28	KAKUNIN (confirmation)	38	KITEI (regulation)
9	HANDAN (decision)	19	SYUUHEN (surround)	29	KOUNAI (in-plant)	39	SAIGAI (disaster)
10	SHIKICHI (site)	20	JIKEN (test)	30	OKUGAI (out of doors)	40	TOKUTEI (particular)
Twitter							
1	BUSHITSU (material object)	11	LEVERU (level)	21	SAGYO (work)	31	BAAI (case)
2	JISHIN (earthquake)	12	HISAI (disaster-affected)	22	KI (mind)	32	ASHITA (tomorrow)
3	JYOUHOU (information)	13	HIGAI (damage)	23	JIKO (accident)	33	KANSOKU (observance)
4	EIKYOU (influence)	14	KANOU (possible)	24	GENSHIRYOKU (nuclear power)	34	HITSUYOU (necessity)
5	TSUJYO (usually)	15	SOKUTEI (measurement)	25	KIKEN (danger)	35	GENZAI (present)
6	TEIDEN (blackout)	16	KOWAI (feel fear)	26	JINTAI (human body)	36	KAKUNIN (confirmation)
7	IMA (now)	17	NYUSU (news)	27	GIJYUTSU (technique)	37	BYOUIN (hospital)
8	HINAN (escape)	18	SINPAI (worry)	28	OSEN (pollution)	38	BUSHI (goods)
9	ONSEN (hot spring)	19	SU (prime (element))	29	KIBOU (hope)	39	HOUSYUTSU (emission)
10	KAKUSAN (spreading)	20	BAKUHATSU (blast)	30	HODO (press report)	40	NOUDO (concentration)

Table 5.2: 40 Highest Ranked Negative Co-Occurrence Frequency Words

(usually)' and 'GENSHIRYOKU (nuclear power)'. Among these terms, 'EIKYO (influence)' and 'SOKUTEI (measurement)' were selected excepting terms with similar meanings or grammatical co-occurrences with 'HOUSYA (radiation)', such as 'nuclear power' and 'material object', as well as terms not directly related to 'HOUSYA (radiation)', such as 'release' and 'usually'. Figure 5.3 and Figure 5.4 show the time series changes of co-occurrence frequency of these words. In TEPCO's announcement, co-occurrence frequencies of these terms decreased after March 14, while the daily changes in occurrence frequency showed a different pattern. TEPCO's anxiety level had not decreased necessarily, but it seems that TEPCO had become silent regarding 'HOUSYA (radiation)'. In contrast, co-occurrence frequency increased for both terms on Twitter after March 14 (see Figure 5.4) which is same pattern of the number of tweets including 'HOUSYA (radiation)'. It seems risk assessments of twitter users were influenced by Twitter users' utterance pattern about 'HOUSYA (radiation)'. In order to overcome the utterance pattern of Twitter, it is necessary to improve the measurement of anxiety levels. For instance, if tweets of users usually contain some words on a routine basis and use of these terms suddenly stops following a disaster, an expressive disaster-contributing anxiety can be identified.

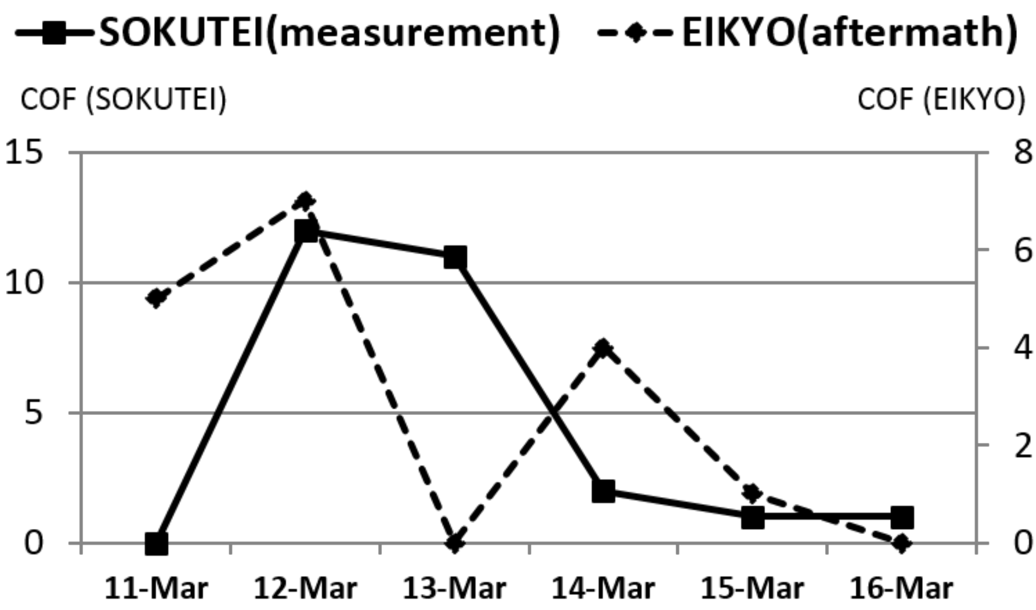

Table 5.3: The Time Series Variation of Co-Occurrence Frequency with 'HOUSYA (radiation)' (TEPCO)

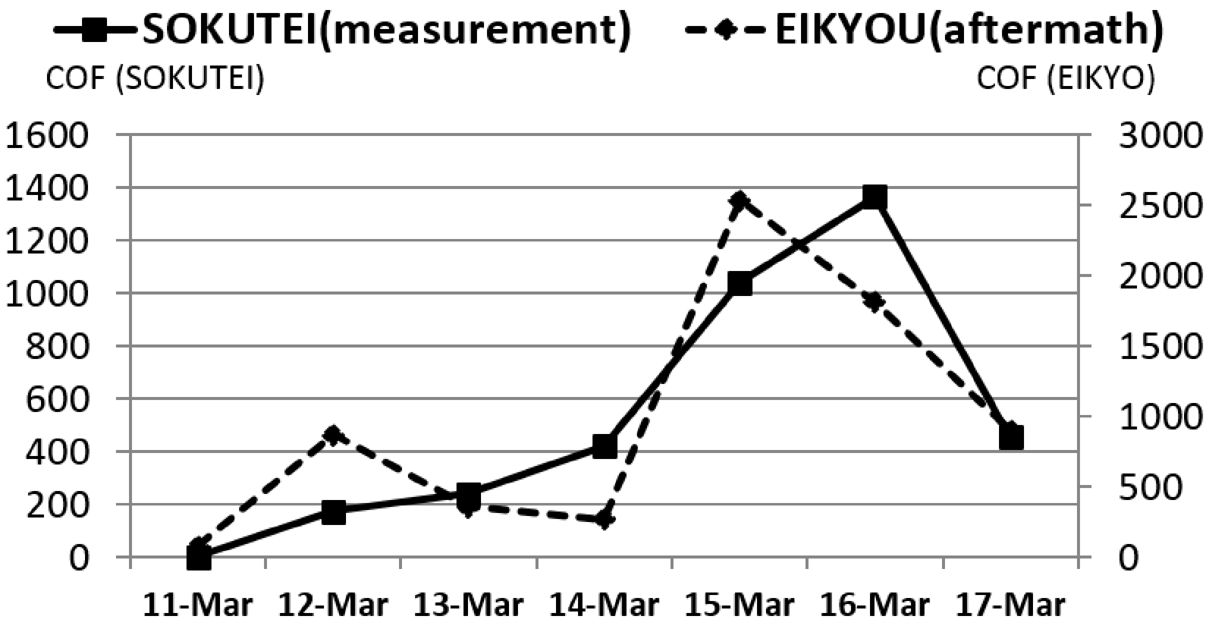

Table 5.4: The Time Series Variation of Co-Occurrence Frequency with 'HOUSYA (radiation)' (Twitter)

5.6 Implication

As explained, Anxiety is defined as an important emotional sensation in crisis communication, because anxiety is one of the significant factors that cause people to organize an unreasonable or irrational information process in disasters. Using quantified measures to perceive changes in anxiety levels, therefore, is essential to policy making. With the anxiety index, we could visualize the different pattern of the risk perception between the organization and the public. From the result of the analysis, risk perception of the public shows insensible changes compared to that of organization, even though the objective level of hazard rise to higher level. The risk perception of the public does not simply depend on the objective level of hazard. It is crucial that organization such as government who is a main information provider in disasters understand that the public who receive the message assess the risk based on their own risk perception which is not only based on objective facts.

5.7 Conclusion

This study has proposed a methodology and application for measuring anxiety in a disaster, based on corpus linguistics and a combination of text mining methods. In the analysis, the contents of crisis communication were investigated using Twitter data collected during early phase of the Great East Japan Earthquake. Measurement of anxiety was achieved by estimating the risk perception of Twitter users with the Anxiety Index, which was proposed in this study as an indicator of negative risk perceptions regarding the Great East Japan Earthquake. Anxiety levels of the public did not show clear fluctuations, while that of TEPCO's announcements changed noticeably.

By utilizing proposed methodology, the sentiment of the public could be evaluated. If the government perceives the public sentiment appropriately, it would help them to provide proper risk information to the public.

However, results of this study leave more to be investigated and answered. It is a limitation of our study that the analysis proceeded with Twitter data was restricted to only the seven days immediately following the disaster. For a more precise measurement of public anxiety, it would be necessary to examine a longer period. Also, our methodology needs improvement. First, data-mining techniques identifying topics and co-related terms are inadequate when applied to a corpus such as Twitter data, which is comprised of a superabundance of short sentences generated by different individuals. Second, the criteria used for determining the polarity (positive and negative) are established to apply to standard Japanese and do not cover the non-standard words (e.g., clipped words, repeated letters, or neologisms) used on Twitter. Many words are therefore omitted in the analysis.

Bibliography

[1] Susan Anthony. Anxiety and rumor. *The Journal of social psychology*, 89(1), pp.91-98, 1973.

[2] Seung Ji Baek, Hayeong Jeong, and Kiyoshi Kobayashi. Disaster anxiety measurement and corpus-based content analysis of crisis communication. In *Proceedings - 2013 IEEE International Conference on Systems, Man, and Cybernetics, SMC 2013*, pp. 1789-1794, 2013.

[3] Hyunyoung Choi and Hal Varian. Predicting the Present with Google Trends. *Economic Record*, 88(SUPPL.1), pp. 2-9, 2012.

[4] Kirill Kireyev, Leysia Palen, and Kenneth Anderson. Applications of topics models to analysis of disaster-related twitter data. *NIPS Workshop on Applications for Topic Models: Text and Beyond*, 2009.

[5] Kiyono, J., Hujii, S., Kobayashi, K., Jeong, H., Shaw, R., Takeuchi, U., & Hirayama, N.. *Kyodai Saigai to Ningen no Anzenhosyou (Catastrophe Disaster and Human Security)*, Puyoushobousyupan (In Japanese), 2013.

[6] Arthur Lupia and Mathew D. McCubbins. *The Democratic Dilemma: Can Citizens Learn What They Need to Know? (Political Economy of Institutions and Decisions).* Cambridge University Press, 1998.

[7] Onook Oh, Kyounghee Hazel Kwon, and H. Raghav Rao. An Exploration of Social Media in Extreme Events : Rumors Theory and Twitter During the Haiti Earthquake 2010.*Thirsty First International Conference on Information Systems, St. Louis*, pp. 231, 2010.

[8] Stefan Stieglitz and Linh Dang-Xuan. Emotions and Information Diffusion in Social MediaSentiment of Microblogs and Sharing Behavior. *Journal of Management Information Systems*, 29(4), pp.217-248, 2013.

[9] Kate Starbird and L Palen. Pass it on?: Retweeting in mass emergency. *Proceedings of the 7th International ISCRAM Conference*, pp.110, 2010.

[10] Seiwa, H.. *Huan (Anxiety). In Y. Nakajima et al, Sinrigakujiten(The dictionary of Psychology)*, p. 738, Yuhikaru, 1999.

[11] Tokyo Electric Power Company Web Page, http://www.tepco.co.jp/cc/press/index-j.html

[12] Tsuru, H.. *Huan (Anxiety). In H. Tsuru, Shinrigakujiten (Psychological Dictionary)*, p. 740, Heibonnsya, 1981.

[13] Hiroya Takamura, Takashi Inui and Manabu Okumura. Extracting semantic orientations of words using spin model. *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics - ACL 05*, pp.133-140, 2005.

[14] Mizuki Yamazaki, Toshiko Kikkawa, and Hedeyuki Horii. a Psychological Model for Evocation of IndividualS Anxiety Related To Social Affairs : the Case of Highly Pathogenic Avian Influenza. *Sociotechnica*, 2, pp. 379-388, 2004.

Chapter 6

Conclusions and Future Research

6.1 Conclusions

In this research, we examine the risk assessment of the public who is one of the key participant of the crisis communication during the Great East Japan Earthquake by means of Twitter corpus. We apply Topic models to clarify the contents of the communication. In chapter 4, the result shows that people's concerns are changing and diversify as the situation changes, but in the same time, we could also find that there are an underlying emotions people have regardless of situation changing which is 'anxiety' regarding the risk 'radiation'. In chapter 5, we evaluate the 'Anxiety' of the public using 'Anxiety Index' that we propose. And we examine the co-relation of Anxiety and time, the results shows insensible changes compared to that of organization, even though the objective level of hazard rise to higher level. It implies that the risk perception of the public does not simply depend on the objective level of hazard.

So, Announcement toward the public usually contains only the objective facts that explain the disaster situation of the stage. However, for the public it sometimes intractable to understand the contents especially when they have to handle extremely complicated risk such as nuclear plant accident. And also that's why the public rely on the announcement of the organization such as the government.

Organizations as a provider of the risk information in disasters, it is crucial that they provide information embracing the concept of potential risks to conduct the crisis communication effectively.

In a brief manner, we summarize every chapter in this dissertation as follows.

In chapter 2, we examine the necessity of the sentiment analysis to investigate the crisis communication and availability of Twitter data as a representative social media in the recent disaster. since the new type of news media contains the local context and individuals level of communication which never observed in pre-existing communication, it have changed the traditional structure of communication. According to the several researches related to crisis communication with Twitter, it could have dual-aspect. One is reinforcing the communication between the stakeholders in our society most importantly the public and the government. On the contrary, it also have a aspect of disturbing their interaction. From the point of view of considering Twitter as a tool for information distribution for the public, especially, in emergencies, the 'emotion' turns out that the most crucial factor motivating the public to make the emotional atmosphere which is the stream of gathering and interacting in society such that was shocked by catastrophe.

For investigating the emotion of the public using Twitter data, we explained some unique characteristic of information interaction in Twitter and its adaptability to analysis of crisis communication. We reviewed other researches that developed the methodologies and application tools to demonstrate the co-relation of emotional statements and information diffusion using data collected from social media. As we deal with the text data, we explained topic models such as TFIDF and LDA which is promising method for analysis of disaster-related Twitter data.

In chapter 3, we investigate the actual crisis communication in The Great East Japan Earthquake and The Fukushima Daiichi Nuclear Disaster. People who are in disasters, use Twitter to describe the events or accidents that is happening or happened just before. In addition, it it observed that they apply twitter as a tool for delivery of information that originally created by group of experts or official organization. It indicates that as they retweeted, they distributed the information by themselves without the media. As for government organization who play a role as a information provider in communication, the contents of information they provide usually depend on concern risks that they are encounter with. So, the contents of central government covered disaster risks such 'radiation' which affected broadly across wide areas, while local government focusing on specific risks such as 'tsunami' that they are faced with. As there was no a clear correlation between the number of tweets and retweeted, it is necessary to consider other elements such as a preformed network which is constructed before the disaster occur to investigate process of delivery of information in Twitter.

In chapter 4, to facilitate efficient communication in the event of a future catastrophe, it is important to clarify concerns of each participant of communication.

In this chapter, we investigate the concerns of the public in The Great East Japan Earthquake who is one of the important participants in communication. We shows variety concerns and how those concerns are changing in disaster. We develop the application to figure out what kind of concerns the public have during the disaster.

For applying LDA model to Twitter data. With LDA model, we examine the contents of utterances in the Twitter corpus by applying LDA algorithm, which identify topics of Twitter corpus by estimating the latent variables of data collection.

Various kinds of concerns have been observed during 7 days. People's concerns diversify as time goes, especially from March 12 when significant explosion from the Fukushima Daiichi nuclear power plant occurred. The public's concerns changes as the disaster situation changes, but in the same time, there are underlying concerns that unlikely changes which is 'anxiety' regarding radiation issue.

In chapter 5, This study has proposed a methodology and application for measuring anxiety in a disaster, based on corpus linguistics and a combination of text mining methods. In the analysis, the contents of crisis communication were investigated using Twitter data collected during early phase of the Great East Japan Earthquake. Measurement of anxiety was achieved by estimating the risk perception of Twitter users with the Anxiety Index, which was proposed in this study as an indicator of negative risk perceptions regarding the Great East Japan Earthquake. Anxiety levels of the public did not show clear fluctuations, while TEPCO's announcements and the Twitter corpus, anxiety expressed by the government changed noticeably.

By utilizing proposed methodology, the sentiment of the public could be evaluated. If the government perceives the public sentiment appropriately, it would help them to provide risk information to the public.

6.2 Topics for Future Research

Each of analysis results of this study leave more to be investigated and answered. It is a limitation of our study that the analysis proceeded with Twitter data was restricted to only the seven days immediately following the disaster. For a more precise measurement of public anxiety, it would be necessary to examine a longer period. Also, our methodology needs improvement.

For more accurate observation in analysis on changing topics as time goes, it is necessary to examine a longer period to know the feature of fundamental concerns 'anxiety'. It is necessary to investigate how the 'anxiety' changes after the catastrophic condition to subside. To solve this problem, we would like to improve the result with applying multiple LDA models to investigate longer period data. For concatenate several LDA models of each stage, Dynamic Topic Models could be applied, which consider the prior distribution from previous stage in the model [Blei and Lafferty, 2006][Fujimoto et al., 2013].

The data-mining techniques identifying topics and co-related terms are inadequate when applied to a corpus such as Twitter data, which is comprised of a superabundance of short sentences generated by different individuals. Second, the criteria used for determining the polarity (positive and negative) are established to apply to standard Japanese and do not cover the non-standard words (e.g., clipped words, repeated letters, or neologisms) used on Twitter. Many words are therefore omitted in the analysis.

Bibliography

[1] David M Blei and John D Lafferty. Dynamic topic models. *Proceedings of the 23rd international conference on Machine learning (ICML 2006)*, pp. 113-120, 2006.

[2] Hiroshi Fujimoto, Takahiro Hara, and Shojiro Nishio. Online Topic Analysis with Online Optimal Smoothing and Dynamically Evolving Vocabulary. (5), pp.1212-1221, 2013.