Comparison of Vocabulary Features among Multiple Data Sources for Constructing a Knowledge Base on Disaster Information

Megumi Yasuo[1*] and Mitsunori Matsushita[1†]

[1*]Graduate School of Informatics, Kansai University,
2-1-1, Ryozenji-cho, Takatsuki, 569-1095, Osaka, Japan.

*Corresponding author(s). E-mail(s): k290993@kansai-u.ac.jp;
Contributing authors: m_mat@kansai-u.ac.jp;
†These authors contributed equally to this work.

Abstract

This research aims to develop a framework for smoothly obtaining disaster information from multiple web services through a knowledge base of disaster information. In Japan, where natural disasters occur frequently, there is a need for a system that can utilize disaster information transmitted on the Web from various locations in disaster-stricken areas for rescue operations and disaster recovery when a disaster occurs. Since such information is posted to many web services, searchers must refer to multiple web services to obtain the desired information. In this study, we propose understanding the characteristics of disaster information posted on each web service and using them as a guide for searchers to obtain disaster information smoothly. To achieve this goal, we tried to construct a vocabulary set of disaster information by acquiring textual information from two different data sources and using word embedding and clustering. Comparison of the acquired disaster information revealed two points: The composition of disaster information categories differs among data sources. Even texts in the same category have different characteristics of words depending on the data source.

Keywords: disaster information, multiple data sources, word embedding, text clustering

1 Introduction

Japan is often called a disaster-prone nation, facing significant disasters yearly. When disasters strike, people in disaster-stricken areas disseminate information about disaster-related information, including damage information, rescue requests, and the delivery status of relief supplies, with various social media, such as Twitter and Facebook. Regarding the Osaka north earthquake in 2018, approximately 220,000 tweets in Japanese containing the word "Jishin (earthquake)" were posted in the first 10 minutes after the earthquake occurred[1]. A significant characteristic of these posts is that they originate directly from people in the disaster-stricken area. Information shared by those directly impacted is rapid, aiding in obtaining information about the disaster-stricken area faster than compilation by mass media. Moreover, information that mass media may not cover can be collected if someone shares it. To support rescue and disaster recovery activities quickly, it is essential to have a framework for efficiently collecting and organizing such information. Utilizing such shared information by individuals, one can swiftly acquire information that traditional newspaper articles, news reports, or local government websites might not capture.

To achieve this objective, it is necessary to identify differences in disaster information posted on various web services. Among the types of disaster information are requests for assistance, traffic updates, and data regarding the locations affected by the disaster. Collecting all these pieces of information from a single resource is difficult. Therefore, those searching for disaster information will focus on resources relevant to their tasks. Such information retrieval assumes that the searcher possesses prior knowledge about collecting disaster information. However, searchers needing this prerequisite knowledge must refer to various web services to obtain the information they seek.

Enabling people to access information easily will help them collect information more effectively. This paper aims to acquire metadata about the content in each web service and use it as a guide for searchers. As the first step, this study analyzes vocabulary acquired from different resources. It examines two aspects: whether there are differences in information acquired between web services and how the content of posted information differs among web services.

2 Related works

This research aims to acquire and store disaster information from the web and use it as knowledge. To achieve this goal, we extract disaster information from text information on web services and compare its contents to investigate the differences in the knowledge acquired from each resource. In this chapter, we review research on the use of information on the web in the event of a disaster and research on information extraction from text and define the position of this research.

2.1 Research on the use of information on the Web in the event of a disaster

Research using disaster information posted on web services has been attempted for various purposes, such as assessing the damage from disasters and collecting information on disaster-stricken areas[2][3][4]. In particular, there is a high demand for collecting disaster information from large-scale general-purpose social networking services such as Twitter, and several studies have been conducted so far[5][6][7]. One of the disaster information analysis systems based on Twitter information is "DISAANA," developed by the National Institute of Information and Communications Technology (NICT)[8]. The system analyzes information posted by Twitter users in real-time, extracts what is happening where, and includes a 5W1H search function and a "contradictory post" function for information whose facts are unclear.

When extracting disaster information from SNS, information unrelated to the disaster is often mixed in as noise. In particular, entertainment-related information, such as games, tends to be posted on SNS. A survey of tweets in Japanese reported that 60% of tweets posted during normal times contained entertainment information[1].

Some studies have analyzed noise postings on SNS from the task of extracting disaster information. Morino et al. extracted disaster information from SNS postings at the time of a disaster and analyzed the noise contained in results[9]. This study examined the types of noise in the mix and tested whether these noises could be separated from lexical features. It suggests that information about "games" is highly separable. On the other hand, we reported that it is difficult to separate noise with lexical features that are not biased toward specific contents, such as "merchandise items."

User-posted information has been considered unverifiable, unreliable, and unsuitable for use by large organizations in actual rescue operations. This point is discussed in some papers[10], including interviews conducted by Tapia et al. with NGO humanitarian organizations[11]. Tapia et al. suggest that the methods to obtain information with acceptable reliability from microblogs include extracting valuable data using text classification techniques and acquiring and automatically classifying information from major geographical areas. Concerning this problem, a disaster information collection system that considers the government's information collection process has been proposed[12]. This system is designed to enable users at the disaster site to report the location of damage smoothly via the web and to ensure the report's reliability by adding location information when the damage is reported.

2.2 Research on information extraction using text clustering

Acquiring textual features using word embeddings is one of the effective methods for information extraction. Word embedding is a method of acquiring distances between words with word vectorization in a set of documents. While this method is widely used for document classification, there are attempts to use these methods to acquire latent knowledge. Magno et al. attempted to identify cultural characteristics and values by country using 1.7 billion tweets posted on Twitter[13]. This attempt classifies the

[1] https://www.biglobe.co.jp/pressroom/release/2011/04/27-1

cultural characteristics of 59 countries based on 22 perspectives, including “religion” and “science.” The study revealed cultural characteristics and values by country, and correlations with actual cultural characteristics were pointed out in several indicators. Word embeddings were widely used in the research described in the previous section, which was oriented toward information recommendation. Park et al. attempted to recommend sightseeing routes according to the profile of tourists by using word-of-mouth information on travel sites for designing sightseeing tour routes[14]. This study analyzed reviews posted on TripAdvisor, a travel review site, and found that each reviewer's profile had unique challenges. Debanjan et al. attempted to extract more valuable reviews from many reviews posted on e-commerce sites by using word variance representation[15]. This study aims to link the subset of good review sentences that mention a product from multiple perspectives with the emotional polarity derived from the review sentences. While many general e-commerce sites provide a function to evaluate the usefulness of the review sentences themselves, the advantage of estimating usefulness scores directly from the review sentences is that it can also be applied to newly posted review sentences. An example of cross-domain use of word embeddings is the lyrics recommendation method by Han et al[16]. This study recommends lyrics similar to the user's environment in a tourist destination by sharing word embeddings of reviews of tourist attractions and song lyrics. Conventional song recommendation methods use meta-information such as genre and artist, but using word embeddings across domains, they recommend lyrics appropriate for the listening environment.

Previous studies on web-based information in the event of a disaster have focused on collecting damage and rescue requests, indicating a high need to obtain disaster information on the Web. On the other hand, several studies on web-based information resources about disasters have pointed out the need to ensure the reliability of web-based information for disaster recovery assistance and the need for knowledge acquisition support systems using text classification technology. Our research focuses on supporting the acquisition of disaster information. It aims to smoothly present information that matches the needs of a searcher who collects information about the disaster that needs to be dealt with quickly from multiple data sources. Previous studies using text clustering have shown that knowledge acquisition based on word embedding is helpful for knowledge extraction and information retrieval based on text. This paper examines building a word set by clustering words using these methods.

3 Collecting disaster information using a combination of data sources

When a large-scale disaster occurs, people post a variety of information about the disaster to web services. At this time, posters need to change the form and content of the information according to the specifications of each web service. For example, when posting to a web service such as Instagram, designed to post information with images attached, the contributor must prepare the images to be posted. Even in the same text media, there is a difference between sites where people share their impressions of an article (e.g., Hatena bookmark[2]), sites where contributors verbalize their claims

[2] https://b.hatena.ne.jp/

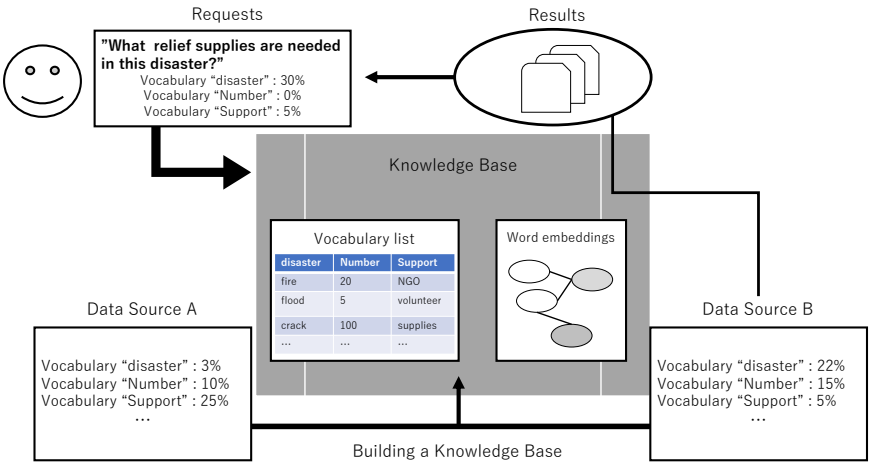

Fig. 1 Knowledge acquisition model for disaster information in multiple sources.

and interpretations in long sentences (e.g., note[3]), and sites where contributors tend to express their impressions and experiences in short sentences (e.g., Twitter), the tendency of the contents of the posts will be different. This difference in form creates a difference in the information stored in each web service. Even when attempting to extract information on disaster information, the information obtained will differ among web services. Therefore, selecting a web service that provides information in line with the collection intention as an information source when acquiring information that meets a specific purpose is necessary. Under the above consideration, this study aims to construct a knowledge base of disaster information to match search queries and target web services to be extracted.

There are several types of large-scale disasters, such as earthquakes, volcanic eruptions, and lightning strikes, but the actual damage caused by these disasters differs from case to case. For example, in the case of a major earthquake, a landslide may occur as a secondary disaster, or a large-scale fire may cause damage. Both of these can be considered earthquake-related disasters, but the countermeasures and rescue policies required are different. Therefore, to obtain the necessary information for each disaster, a search framework that considers the disaster's characteristics is necessary. Information retrieval through the knowledge base of disaster information can estimate possible secondary disasters based on the characteristics of the disaster to be retrieved and present them to the searcher.

The knowledge base for handling disaster information is built from a vocabulary set clustered by word meaning and a set of sentences labeled based on disaster-related categories. The vocabulary sets are clustered based on the semantic similarity of the words obtained from the disaster information. The vocabulary sets are obtained for each type of disaster. The set of sentences is constructed by dividing the text acquired from web services such as SNS and news articles into sentences and classifying them

[3]https://note.com/

as disaster information based on the content of the sentences. Once a distributed representation is obtained from the set of sentences, a sentence vector based on the vocabulary set is calculated for each sentence. The sentence vectors are associated with labels assigned to the original sentences. The knowledge base we aim to build in this research is statistical information on sentence vectors for each disaster category. With this knowledge base, it is possible to analogize the relevant disaster category from the input text's features and perform similarity search and information extraction based on semantic distance.

The process of acquiring information using the disaster information knowledge base is shown in Figure 1. This figure shows the relationship between the searcher's information request, the knowledge base, and each data source. When a searcher makes an information request through the search system, the knowledge base analyzes which cluster features are included in the information request and returns appropriate knowledge and data sources as search results according to the features. The advantage of this method over general query matching and similarity search is that it can present search results considering the semantic distance per data source. As mentioned above, the tendency of information stored in web services varies depending on the design of the web service. The semantic distance of a data source to an information request can be used to determine whether the information in the presented data source should be the target of a detailed search.

Since building a disaster knowledge base requires extracting knowledge from vast data, having more data as a resource is generally desirable. However, processing vast data to build a knowledge base requires enormous computational resources. To construct a helpful knowledge base, ensuring the diversity of information obtained from each data source is desirable. In this paper, we focus on the diversity of information in building a knowledge base of disaster information and verify that using multiple data sources together improves knowledge coverage.

4 Comparison of the nature of disaster information across data sources

In this paper, we obtained disaster information from two websites and analyzed their contents to clarify the differences in the information obtained from each site. Disaster information includes images and video data showing the damage, text data such as damage reports and requests for help, and numerical data such as location information. In this paper, text data is used as the target of analysis as the information to be acquired. As an analysis method, this paper uses a combination of word embedding and clustering to compare the characteristics of the words in each resource. This method is used to analyze trends in a dataset based on the semantic similarity of words. Text data on disasters are extracted from two web services, morphological analysis is performed to extract the part-of-speech of the words to be analyzed, a word distribution representation is obtained, and then clustering is performed to obtain semantically similar word clusters. The characteristics of the words in each data source are then compared to determine whether there are any differences between the data sources and their characteristics.

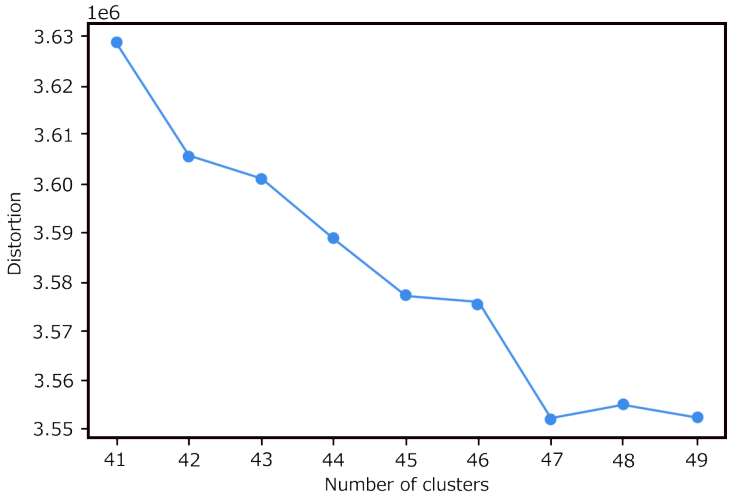

Fig. 2 Determining the number of cluster divisions using the elbow method.

Table 1 Examples of label names and words in each class

Label Name	Words
abbreviation	ROCK, ETC, DO, NR, RU, JET, SS
mood	subtle, hard, intense, grandiose, vivid, sufficient, crippled
operation	management, destruction, delivery, placement, removal, construction
number	7, three, 1, hundred, six, one, five, two

4.1 Extracting disaster-related words

This study focused on the July 2020 torrential rainstorm in Japan from July 3 to July 31, 2020 (hereafter, the Kumamoto torrential rainstorm disaster)[4]. This disaster caused extensive damage, mainly in the Kyushu region, where 84 people died. The data source used was 9,206 Japanese-language tweets about the July 2020 torrential rain collected from Twitter, which was manually verified for content after eliminating retweets and tweets with duplicate content. In addition, among news articles reported during the same period, articles containing "Kumamoto torrential rainstorm disaster" in the title were extracted from the Mainichi Shimbun Article Search Service[5]. As a result, 271 articles were extracted. Then, nouns, adjectives, and adverbs for analysis and generated word variance representations were extracted using the Japanese Wikipedia entity vector[6]. The data were analyzed by Mecab (Ver. 0.996)[17], a morphological analysis engine for Japanese, and mecab-ipadic-NEologd (Ver. 0.0.7)[7], a Japanese dictionary and the target words extracted. As a result, a vocabulary of 13,305 words and 7,362 words were obtained from Twitter data and news articles, respectively.

[4]https://www.data.jma.go.jp/kumamoto/shosai/kakusyusiryou/20200708_kumamoto.pdf
[5]https://mainichi.jp/contents/edu/maisaku/
[6]http://www.cl.ecei.tohoku.ac.jp/~m-suzuki/jawiki_vector
[7]https://github.com/neologd/mecab-ipadic-neologd

The total number of unique words, excluding duplicates from each data source, was 17,896. The acquired words were subjected to cluster divisions using the k-means++ method for cluster numbers ranging from 2 to 70, and the progression of the sum of squared errors (SSE) was calculated. Figure 2 shows the results. Based on this figure, the elbow method was applied to estimate the optimal cluster number, revealing that a cluster number of 47 was determined to be an appropriate choice for the division.

Appropriate cluster names were then assigned to the clusters created manually. The labeling task was performed by four university students from the informatics faculty (hereafter, labeler). Thirty words were randomly selected from the acquired words classified into each cluster. Four labelers were requested to assign the most appropriate cluster name to represent words in each cluster. Words in the collected responses that were identical for two or more labelers were allowed to overlap with other clusters and were used as label names. Items that were not uniquely defined were not assigned a label name and were excluded from the analysis. Table 1 shows examples of the label names and some words included in each class. Finally, label names were assigned to 42 of the 47 clusters.

4.2 Analysing disaster-related words appeared

The proposal in this paper is to use multiple data sources together to build a knowledge base of disaster information, aiming at acquiring knowledge that is difficult to obtain from a single data source. To verify the proposal's validity, it is necessary to experimentally demonstrate that the nature of the information obtained from different data sources is different and that by using them together, information that cannot be obtained from a single data source can be collected. To verify the proposal, we observed the text acquired from each data source and assigned labels to them qualitatively. We compared their composition ratios to observe whether there was a difference in the type of information acquired across data sources.

First, to observe what kind of information each data source contained, we divided the data sources and assigned each sentence a classification label based on its textual content (hereafter, sentence label). For this process, we divided the tweet data into sentences, and the news article data was divided based on the punctuation points.

The nine sentence labels assigned are "disaster information," "traffic information," "support information," "human damage reports," "weather & warning information," "physical damage reports," "evacuation information," and "others." Table 2 shows the criteria for the assignment of sentence labels. To clarify whether there is a difference in the expression of disaster information among data sources, the composition ratio of the assigned sentence labels was compared among the data sources. Finally, we validated 30 news articles (including 426 sentences) and 1436 tweets for about 10% of the data used to build the lexical set.

Figure 3 shows the composition ratios of sentence labels assigned to each data source. The top 5% of all the sentences in each data source were labeled "support information," "human damage report," and "physical damage report" for the news articles group, and "disaster information," "human damage report," and "evacuation information" for the tweets group.

Table 2 Labels attached to the sentences and their criteria

Label	Contents
disaster information	an overview of the disaster, such as an outline of the disaster.
traffic information	public transportation and the availability of public roads
support information	relief supplies and volunteers
human damage reports	injuries, isolation, and other damage caused by the disaster
weather & warning information	weather information and warnings issued or lifted
physical damage reports	damage to homes and public facilities
evacuation information	evacuation centers, evacuees and evacuation status
others	contents that fall outside the above categories.

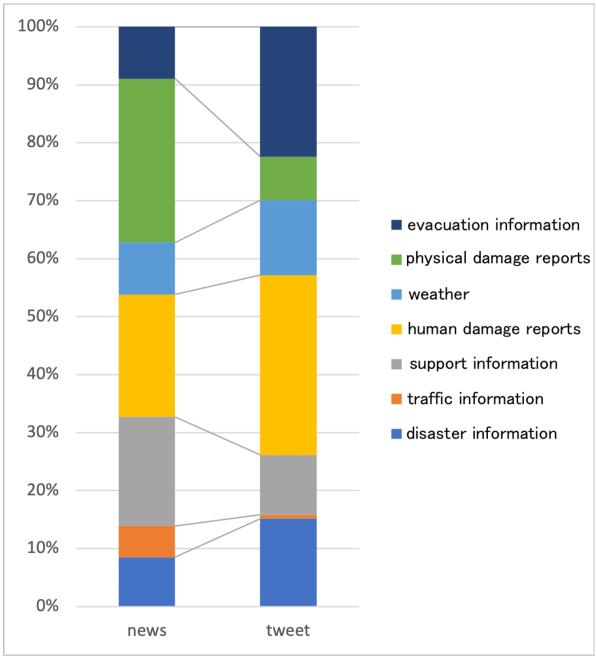

Fig. 3 Sentence label composition ratio among data resources.

Next, we analyzed the relationship between sentence labels and each cluster to reveal the differences in word characteristics between data sources. We calculated the proportion of words corresponding to each cluster for each data source, normalized the number of corresponding words by the number of sentences, and then calculated the differences between data sources. The calculated results are presented in Table 3. Larger values indicate a higher frequency of vocabulary usage in tweets for the corresponding cluster, while smaller values suggest a higher frequency of vocabulary usage in news articles. Based on this data, we qualitatively examined items falling within the 3-σ range of the standard deviation of the calculated results and considered the observed features unique to each data source.

Words in the cluster labeled “abbreviation” were frequently found in the tweet data labeled “support information:” The words in the “abbreviation” cluster were observed

Table 3 An excerpt of correlation biases between word cluster labels and disaster categories

label	abbreviations	mood	Kanji	operation	number
disaster information	0.4561	0.2578	0.1533	0.0847	-0.1602
traffic information	0.4167	**-0.5000**	**0.6667**	**0.5833**	-0.1667
support information	**0.6282**	0.3034	-0.0557	0.0420	**-0.4164**
human damage reports	0.4397	-0.0071	-0.1773	0.0780	**-0.6809**
weather & wearing information	0.2076	0.0466	0.1076	0.1746	-0.3788
physical damage reports	0.4542	0.2997	0.2502	0.2890	-0.3819
evacuation information	0.2392	-0.1353	-0.1275	0.2490	**-0.5245**
ohers	0.2254	0.1056	0.0230	-0.0149	-0.1874

in 34 of the 47 sentences labeled "support information." An example sentence is as follows:

> "*KDDI* and Okinawa Cellular are implementing support measures for customers in areas where the Disaster Relief Law has been applied due to the recent heavy rains in Kyushu. We sincerely hope for the earliest possible restoration of operations. For more information, please click here → *[URL]*"

These sentences included abbreviations of company names and URLs.

Words in the cluster labeled "mood" appeared frequently in the news data labeled "traffic information:" The words in the "mood" were observed in 10 of the 12 sentences labeled "Traffic Information." An example sentence is as follows:

> "When the reporter entered the Issachi area in the center of the village, which had become inaccessible due to the severing of National Highway 219 and other roads, he was *left speechless*."

In these sentences, descriptions of the personal opinions and impressions of the describer were confirmed.

Words in the cluster labeled "number" appeared in the news data labeled "support information," "human damage report," and "evacuation information:" An example sentence is as follows:

> "The number of evacuees has risen to *2,099* in four prefectures, including Kumamoto."

In these sentences, the number of evacuees and damaged houses were included in the sentence.

5 Discussion

In the experiment in the previous chapter, we obtained disaster information from two different data sources and compared their contents to clarify "whether it is possible to obtain disaster information of different nature from multiple data sources" and "how the obtained disaster information differs among data sources." The differences in the composition of the sentence labels indicate that the nature of the disaster information available among the data sources is different. The results show that news articles provide mainly information on relief and material damage, while tweets provide information on general disasters and evacuation. This result indicates that the type of

knowledge stored in each web service is different, meaning that by using multiple resources together to build a knowledge base of disaster information, it is possible to acquire knowledge that cannot be acquired with a single resource. The analysis of lexical features among data sources revealed that each data source has its characteristic descriptions. In particular, a comparison of data labeled "traffic information" suggested that the available information differs among the resources, even when the same label is used. This result indicates that the same event contains references from different perspectives. This result indicates that, when searching for disaster information, it is essential to use different data sources according to the problem to be solved.

This method has the limitation that it cannot be used for content that includes other modalities, such as images and videos. Especially for tweet data, we often observe references to other URLs or postings that assume the user is viewing the attached image. By considering methods for acquiring knowledge from such complex information, the construction of a more practical knowledge base should be considered. In addition, the data in this paper was extracted based only on the period and keywords, but it includes automatic posting by bots and reportage articles. In building a knowledge base, constructing a knowledge base with less noise should be considered by combining more advanced data cleansing methods.

6 Conclusion

This study aimed to facilitate access to disaster information by constructing a knowledge base on disaster information. It examined the significance of constructing a knowledge base using multiple data sources. Textual data were obtained from two different data sources, and the differences in the knowledge obtained from each data source were analyzed using word embedding and clustering. The analysis revealed that the composition of the available data differs between the data sources and that the knowledge acquired differs between the data sources, even for references to the same event. These results suggest the appropriateness of using multiple data sources on the Web to build a knowledge base of disaster information. It also suggests that it is possible to supplement the knowledge acquired from a single data source with information from different perspectives.

References

[1] S. Yamada, K. Utsu, and O. Uchida, "An analysis of tweets during the 2018 osaka north earthquake in japan -a brief report," in *2018 5th International Conference on Information and Communication Technologies for Disaster Management (ICT-DM)*, 2018, pp. 1–5.

[2] J. B. Houston, J. Hawthorne, M. F. Perreault, E. H. Park, M. Goldstein Hode, M. R. Halliwell, S. E. Turner McGowen, R. Davis, S. Vaid, J. A. McElderry, and S. A. Griffith, "Social media and disasters: a functional framework for social media use in disaster planning, response, and research," *Disasters*, vol. 39, no. 1, pp. 1–22, 2015.

[3] M. Gerald and K. Yamamoto, "Flood disaster management system for situation awareness and response using twitter data," in *Information Technology in Disaster Risk Reduction*, J. Sasaki, Y. Murayama, D. Velev, and P. Zlateva, Eds. Cham: Springer International Publishing, 2022, pp. 35–48.

[4] Q. Cui, K. Shoyama, M. Hanashima, and Y. Usuda, "Early estimation of heavy rain damage at the municipal level based on time-series analysis of sns information," *Journal of Disaster Research*, vol. 17, no. 6, pp. 944–955, 2022.

[5] D. E. Alexander, "Social media in disaster risk reduction and crisis management," *Science and Engineering Ethics*, vol. 20, no. 3, pp. 717–733, 2014.

[6] S. Chair, M. Charrad, and N. B. B. Saoud, "Towards a social media-based framework for disaster communication," *Procedia Computer Science*, vol. 164, pp. 271–278, 2019.

[7] T. Ishii, H. Nakayama, R. Onuma, H. Kaminaga, Y. Miyadera, and S. Nakamura, "A framework for promoting the experience of novices in examining articles that alert dangers of disaster on social media," in *2022 International Conference on Computational Science and Computational Intelligence (CSCI)*, 2022, pp. 2081–2085.

[8] J. Mizuno, M. Tanaka, K. Ohtake, J.-H. Oh, J. Kloetzer, C. Hashimoto, and K. Torisawa, "WISDOM X, DISAANA and D-SUMM: Large-scale nlp systems for analyzing textual big data," in *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: System Demonstrations*, 2016, pp. 263–267.

[9] Y. Morino and M. Matsushita, "Investigation of contamination by entertainment content in disaster information gathering," in *Information Processing Society of Japan, Special Interest Group on Entertainment Computing (IPSJ-SIGEC)*, vol. 2022-EC-65, no. 33, 2022, pp. 1–2, in Japanese.

[10] B. M. Alajmi and O. Khalil, "The extent of and motivation for disaster information seeking behavior via social networking sites," *Journal of Electronic Resources Librarianship*, vol. 34, no. 3, pp. 219–244, 2022.

[11] A. Tapia, K. Bajpai, J. Jansen, and J. Yen, "Seeking the trustworthy tweet: Can microblogged data fit the information needs of disaster response and humanitarian relief organizations," *Proceedings of the 8th International ISCRAM Conference*, 01 2011.

[12] M. Yasuo, S. Kitamura, and M. Matsushita, "Basic study on information sharing system for gathering damage situation in large scale disaster," in *Proceedings of Human Communication Symposium 2018*, no. B-6-2, 2018, in japanese.

[13] G. Magno and V. Almeida, "Measuring international online human values with word embeddings," *ACM Transactions on the Web*, vol. 16, no. 2, pp. 1–38, 2021.

[14] S.-T. Park and C. Liu, "A study on topic models using lda and word2vec in travel route recommendation: Focus on convergence travel and tours reviews," *Personal and Ubiquitous Computing*, vol. 26, no. 2, p. 429–445, 2022.

[15] D. Paul, S. Sarkar, M. Chelliah, C. Kalyan, and P. P. Sinai Nadkarni, "Recommendation of high quality representative reviews in e-commerce," in *Proceedings of the Eleventh ACM Conference on Recommender Systems*. Association for Computing Machinery, 2017, p. 311–315.

[16] Y. Han, R. Yamanishi, and Y. Nishihara, "Music retrieval focusing on lyrics with summary of tourist-spot reviews based on shared word-vectors," in *2020 International Conference on Technologies and Applications of Artificial Intelligence*, 2020, pp. 73–78.

[17] T. Kudo, K. Yamamoto, and Y. Matsumoto, "Applying conditional random fields to Japanese morphological analysis," in *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 2004, pp. 230–237.