DISASTER TWEET CLASSIFICATION FOR DAMAGE ASSESSMENT AND ITS IMPROVEMENT WITH FEATURE ANALYSIS

Yuto OIKAWA[1], Michal PTASZYNSKI[2] and Fumito MASUI[3]

[1]Dept. of Computer Science, Kitami Inst. of Tech.
E-mail: f1712200362@std.kitami-it.ac.jp
[2]Associate Professor, Dept. of Computer Science, Kitami Inst. of Tech.
E-mail: michal@mail.kitami-it.ac.jp
[3]Professor, Dept. of Computer Science, Kitami Inst. of Tech.
E-mail: f-masui@mail.kitami-it.ac.jp

(Koencho 165, Kitami, Hokkaido 090-8507, Japan)

In extracting tweets useful in rescue missions during disasters, previous research have focused on extracting tweets containing specific addresses or locations. We assume that tweets without addresses can also be useful for disaster relief as the location can be inferred or written indirectly. However, to do that it is firstly necessary to be able to differentiate between tweets describing direct experiences from opinions, and other unrelated contents. Therefore, in this study, we used BERT(Bidirectional Encoder Representations from Transformers) to classify tweets into three classes depending on the directness of information they express, based on the assumption that the tweets posted during disasters that are expressing direct experiences, when provided to rescue teams, can be useful for evaluating the disaster situation. Additionally, We confirmed the assumption that frequent words in the two of the three categories prevented correct classification, and improved the classification efficacy. The results were satisfying enough to be considered for application in efficient information extraction during disasters.

Keywords *: Disaster Information, Rescue Request, Twitter, BERT, Document Classification*

1. INTRODUCTION

Out of the various social network services (SNS), Twitter[*1] is characteristic for its high immediacy of information proliferation, with its users tending to post their experiences in an immediate way in the form of short messages [8, 1, 9]. Messages posted on Twitter (later: “tweets") have a character limit of 280 characters (or 140 characters in Japanese, Chinese, and Korean), and this limitation amplifies the immediacy [11, 5]. It is these characteristics that make Twitter suitable for acquiring real-time local information from a large number of people, which is especially useful in time of disasters. Moreover, in the case of a disaster, information can be both sent and received without being affected by the dropout of power or the congestion of telephone lines, as long as there is any access to the Internet. It is a unique advantage that is not found in traditional media such as television and radio.

For this reason, it has been pointed out that collecting and transmitting information on Twitter is an effective means of saving lives in the time of disasters. [3, 12, 15, 10, 9] This method was already applied during the typhoon disaster in November 2019, Nagano Prefecture, Japan, and resulted in successfully responding to about 50 rescue request tweets with actual rescue missions. This example provides a real world evidence for the effectiveness of Twitter in the times of disaster.

Additionally, there have been a number of studies on the extraction of useful tweets during disasters [16, 25, 23, 20, 18] and on the application of rescue-related tweets [19, 21, 26, 22, 17]. In the former research, the focus is on the definition and on the way of extracting the useful tweets, but the specific usability scenarios are not thoroughly evaluated. The latter studies focus primarily on tweets that include addresses, but it is questionable whether the address information is sufficient in rescue operations, since providing address information during disasters is not widespread.

We argue, that there are some tweets that are valuable in a disaster relief event, even when they do not contain

[*1] https://twitter.com/

specific addresses. An example of this includes tweets from people who have directly witnessed a person in need of rescue. A tweet like this is useful to indirectly estimate the whereabouts of the victims. In addition, tweets from users who have directly experienced the disaster are helpful not only for rescue teams but also for people outside the disaster area in assessing the damage. For example, when a disaster occurs in a region where one's parents reside and the safety of one's family members is unknown, being able to obtain the tweets with direct descriptions of the situation around the area in question is useful in assessing the potential danger to the close ones.

The above-mentioned tweets often appear when a disaster occurs. This is because a disaster is an extraordinary event, which makes such tweets easily noticeable for users who use Twitter regularly. Although it has to be noted that in the case of a disaster of a large scale that requires evacuation, tweets revealing one's situation should be sent after the appropriate evacuation has been completed, while in the case of a disaster not requiring immediate evacuation, such tweets could be sent immediately when the user is exposed to damage.

In this study, we focus on a situation when a disaster has occurred, and aim to make the tweet logs easily available to rescue teams and people outside the disaster areas to help them understand the situation and make informed decisions. In order to do this, in this research we construct a machine learning classifier for tweets by using the classification criteria defined by Fukushima et al. (2014) [3] (Section 2.), report on the initial performance of the constructed classifier (Section 3.), analyze the properties of the errors in training data (Section 4.) and use this analysis to further improve the classifier performance (Section 5.).

2. CLASSIFICATION CRITERIA AND TRAINING DATA

(1) Classification criteria

Classification criteria for the tweets are shown in Table 1. For the purpose of judging the damage situation during disasters, the following three types of tweets are used: tweets from users who directly experienced the disaster, the factual information, and the decisive expressions. These three types of tweets are referred to as the primary information, or the information provided by someone who directly saw, heard or experienced the situation described in the tweet.

In addition, it is necessary to exclude the influence of cognitive bias[6] in order to make appropriate judgments about situation using the tweet logs. In particular, it is important to exclude the influence of the anchor effect[13] and confirmation bias[7]. In order to do this, tweets that may cause cognitive bias are classified separately to the primary information. In this study, we consider tweets that contain elements of an opinion, emotional expressions, expressions of intentions, and call to action to be sources of such cognitive bias. These four types of tweets are referred to as the sesquiary information (the `Other` category contains tweets such as conversations with someone, soliloquies, greetings, etc., which are included in sesquiary information due to high similarity with such). The applicability of the sesquiary information was studied before in the context of decision making during elections, shopping, etc.[3].

Table 1 Classification criteria for tweets.

Sub category	Category		
	Primary	Sesquiary	Secondary
Direct experience	○		
Factual Information	○		
Decisive expressions	○		
Opinions		○	
Emotional expressions		○	
Expression of an intention		○	
A call to action		○	
Other		○	
Expressions indicating a rumor			○
News article			○

Finally, tweets that contain expressions indicating rumors, and tweets that contain references to news articles, represent the type of information that is provided as not first-hand, and are referred to as the secondary information.

a) Challenges in previous research

The classification performance of Fukushima et al.'s[24] approach is shown in Tables 2 and 3. It can be seen from the tables that the majority of tweets were predicted as primary information. This comes from bias in training data. Hence, to improve the classification performance of tweets based on the criteria of Fukushima et al., there is a need to 1) collect more training data to decrease the bias and 2) collect a wider variety of data to increase the diversity of information included in the training data and help the classifier achieve better generalization.

(2) Training data

An overview of the training data is shown in the Table 4. In this table, the top three are the tweets collected by Fukushima et al. (the tweets about the Great East Japan Earthquake are the tweets provided by Twitter Japan, Inc. for the Big Data Project). Such tweets are more than 6 years old, therefore we considered them as one type of data. The tweets about heavy rain and typhoon were collected by the

Table 2 Classification performance (evaluation index) in Fukushima et al.'s efforts.

	Precision	Recall	F1-score
Primary	0.39	0.93	0.55
Sesquiary	0.67	0.33	0.44
Secondary	1.00	0.13	0.24
Macro Avg.	0.69	0.47	0.41

Table 3 Classification performance (confusion matrix) in Fukushima et al.'s efforts.

		Predicted		
		Primary	Sesquiary	Secondary
Actual	Primary	28	2	0
	Sesquiary	20	10	0
	Secondary	23	3	4

author. The number of tweets in these three categories is 350 per information type.

Regarding the annotation, the data by Fukushima et al. was annotated by several people, while tweets about heavy rain and typhoon were annotated by the author himself. In cases where a tweet falls within more than one category, an ordering of priority was given as follows `secondary > primary > sesquiary`. The reason was that it is appropriate to consider tweets that contain elements of secondary information as such information and filter it out as potentially containing rumors and second-hand information, while the usefulness of the primary information is defined to be greater than that of the sesquiary information, thus even if a tweet contains its elements, the priority is given to the primary information [3].

Annotation of the subcategories was also performed by the author of the paper himself. The distribution of each subcategory was shown in the Table 5. It can be seen from this table that there is a certain number of tweets that could cause cognitive bias in the event of a disaster. In addition, since the Other subcategory accounts for more than half of the tweets (573 in total), the category of sesquiary information may in practice be biased towards the Other subcategory. However, this is not considered a problem because in the occurrence of a disaster sesquiary information is not primarily used in decision making.

In the actual classification of tweets, the tweets are not classified into subcategories, but into the three main categories: primary, sesquiary, and secondary. As a preprocessing of the data, the URLs are replaced with the string [URL] using a regular expression, because all the URLs in tweets collected through the Twitter API are usually shortened to a pattern: `http://t.co/random_string`.

3. PRELIMINARY EXPERIMENTS

In this section, we describe a classification experiment using a pre-trained "BERT"[2] language model for Japanese created with the Transformer[14] architecture, which can take into account the whole sentence structure. The BERT (Bidirectional Encoder Representations from Transformers) model is a language model that can be applied to existing natural language processing tasks through the use of transfer learning.

(1) Method

The data described in the previous section was split into two parts, 80% of which was used as training data in the process of fine-tuning the BERT language model, and 20% was used as test data. The data was stratified in such a way that the ratio between the number of categories and the number of types of data was equal. To classify the test data, we transformed the training data into a distributed representation and let BERT learn the features of the categories. BERT then automatically classified the test data into the categories with the most similar features.

(2) Result and discussion

The evaluation results of the classifier are shown in Table 6. The reason for the high classification performance of the secondary information was that the URL in the subcategory "the news articles" often functions as a distinctive feature. The reason why the classification performance of the primary information and the sesquiary information is the same was that these two types of information often appear together in one tweet, and thus there is no characteristic feature to easily separate them, making it difficult to discriminate between them. Hence, in order to clarify the distinctive features and mixed features, we analyzed the frequently occurring words in the tweets containing the primary and the sesquiary information.

4. ANALYSIS OF FREQUENT WORDS

In the previous section, we noticed that there was a lack of distinctive features in tweets containing primary and sesquiary information, which was a factor in the degradation of classification performance. Therefore in this section we describe the results of a further exploration on frequently occurring words in each subcategory and data type in order to identify the words that are characteristic for primary and sesquiary information.

(1) Method

The subcategories with more than 90 tweets in two or more data types were targeted, and the top 20 most frequently occurring words were identified. Only content words (verbs, nouns, adjectives, and adverbs) were included in the analysis, and words used as search queries (Mt. Ontake, typhoon, heavy rain, and election) were excluded. For the analysis, we tokenized the training data

Table 4 An overview of the training data

Data types	Number	Period	Search query
Fukushima et al._Great Earthquake	297	2011.3.11～3.17	※
Fukushima et al._Mt. Ontake	522	2014.9.27～10.6	#御嶽山 (#Ontake-san / Mount Ontake)
Fukushima et al._Lower House Election	231	2014.12.2～12.14	#総選挙 (#sōsenkyo / general elections)
Heavy_rain	1050	2020.7.4～7.8	豪雨 (Gōu)
Typhoon	1050	2020.9.2～9.6	#台風，#台風 9 号，#台風 10 号 (#taifū / typhoon, #taifū9gō / typhoon no.10)

Table 5 Distribution of subcategories

Subcategories	Data types		
	Fukushima et al.	heavy rain	typhoon
Direct experience	35	174	330
Factual Information	143	35	5
Decisive expressions	172	140	15
Opinions	51	27	28
Emotional expressions	26	112	94
Expression of an intention	14	20	17
A call to action	20	37	32
The others	239	155	179
Expressions indicating a rumor	160	46	22
News article	190	304	328

using GiNZA[4], a model for the analysis of Japanese, and counted the occurrence frequency of lemmatized words. The lemmatization allows us to aggregate the conjugations of the words into a single dictionary word form.

(2) Results and discussion of analysis

The results of the analysis for each subcategory are shown in Figures 1-3. Note that “Decisive expressions" are not shown because there were no occurrences of this category in the the results.

From these tables, we can see that the six words, namely, 怖い (scary), やばい (bad [colloquial]), すごい (great [colloquial]), 強い (strong), 寝る (sleep), and 停電 (power outage) frequently appear in both primary and sesquiary information in tweets about typhoons. Such words are considered to be one of the causes that make it difficult to discriminate between primary and sesquiary information. Hence, in the next classification experiment we verified whether the tweets containing these words actually made it difficult for the model to distinguish between the two categories.

5. CLASSIFICATION EXPERIMENT

In this section, we describe the results of a classification experiment after refining the training data based on the results of feature analysis in the previous section.

(1) Method

Any samples in training data containing any of the six words mentioned in the discussion in Section 4. were excluded from one of the categories, and replaced with another tweet. For this case, we focused on the meanings of the words, and considered “blackout" as the word with the high primary information, and “scary," “bad," “great," “strong," and “sleep" as the words with the high occurrence in sesquiary information. We excluded 26 tweets containing “blackout" from the sesquiary information and 105 tweets containing any of the remaining 5 words from the primary information and replaced them with other tweets from the same category.

(2) Results and discussion

In Table 7, we show the classification results after the replacement, while in Tables 8-9 we show the confusion

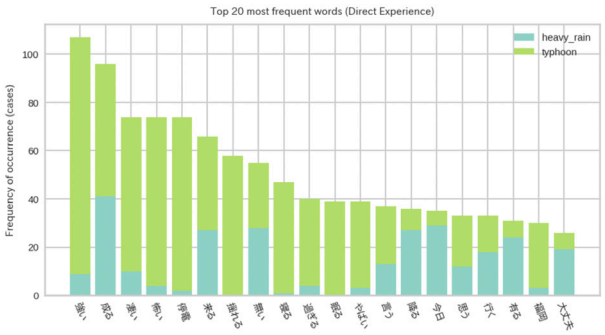

Fig.1 Frequent words : Direct experience

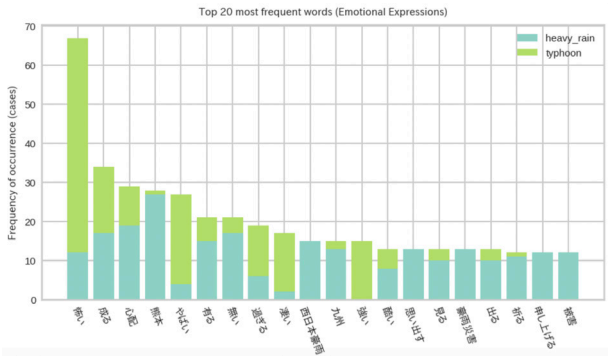

Fig.2 Frequent Words : Emotional expressions

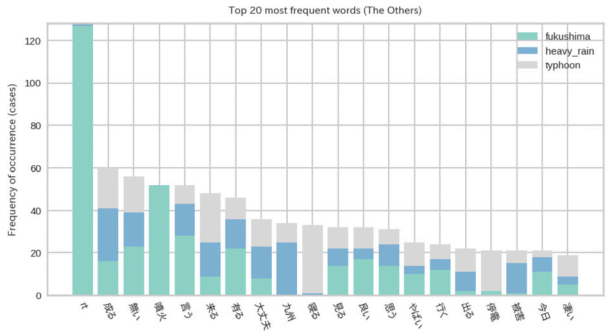

Fig.3 Frequent Words : Other

matrices before and after the improvement. Comparing Table 6 and Table 7, we can see that the classification performance of sesquiary information has improved and the value of Recall of the primary information was improved by 10%. Moreover, comparing Table 8-9, we can see that the number of tweets predicted to be sesquiary information decreased for tweets with primary information. This result suggests that the training data containing one of the words in the question may have caused the original misclassification.

Nevertheless, when Cochran's Q test was performed on the confusion matrices before and after the improvement, there was no significant difference in the performance ($Q(2)=0$, $p = 1$). Hence, it is not possible to conclude the cause of discrimination difficulty from this result alone, but it is possible that a significant difference will be found with more data in the future.

Table 6 Classification performance (before improvement).

	Precision	Recall	F1-score
Primary	0.78	0.70	0.74
Sesquiary	0.76	0.73	0.75
Secondary	0.80	0.91	0.86
Macro Avg.	0.78	0.78	0.78

Table 7 Classification performance (after improvement).

	Precision	Recall	F1-score
Primary	0.77	0.80	0.78
Sesquiary	0.80	0.75	0.77
Secondary	0.83	0.86	0.85
Macro Avg.	0.80	0.80	0.80

Table 8 Confusion matrix (before improvement).

		Predicted		
		Primary	Sesquiary	Secondary
Actual	Primary	147	38	25
	Sesquiary	34	154	22
	Secondary	7	11	192

Table 9 Confusion matrix (after improvement).

		Predicted		
		Primary	Sesquiary	Secondary
Actual	Primary	167	27	16
	Sesquiary	33	157	20
	Secondary	17	13	180

6. CONCLUSIONS

In this paper, we focused on the analysis of tweets by users who have directly experienced a disaster, and attempted to automatically extract them for the use in determining the damage assessment during disasters. Consequently, we achieved a performance of about 80%, which can be considered as sufficiently high for smooth information support. Moreover, feature analysis of the training data and additional classification experiment after replacing erroneous samples in the data suggested that training data containing words that frequently appear in the two categories may be the cause of preventing correct classification.

In future work, we plan to use the constructed classifier to analyze more data to see what kind of emotion each category is correlated with.

ACKNOWLEDGMENTS This research has been partially supported by Kitami Institute of Technology Research Center for Strategic Assistance in the Prevention of Floods, Earthquakes and Regional Hazards (SAFER), research project on “Application of information triage methods in disaster prevention."

REFERENCES

1) Das, S., Dutta, A., Medina, G., Minjares-Kyle, L., and Elgart, Z. Extracting patterns from twitter to promote biking. *IATSS Research*, 43(1):51 – 59, 2019.
2) Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.
3) Fukushima, Y., Masui, F., Ptaszynski, M., Nakajima, Y., Watanabe, K., Kawaishi, R., Nitta, T., and Sato, R. Macroanalysis of microblogs: An empirical study of communication strategies on twitter during disasters and elections. In *2014 AAAI Spring Symposium Series*, 2014.
4) Hiroshi, M. and Masayuki. 短単位品詞の用法曖昧性解決と依存関係ラベリングの同時学習. **言語処理学会第 *25* 回年次大会**, 2019.
5) Hull, K. and Lewis, N. P. Why twitter displaces broadcast sports media: A model. *International Journal of Sport Communication*, 7(1):16–33, 2014.
6) Kahneman. "subjective probability : A judgement of representativeness". In *Cognitive Psychology 3*, pages 430–454.
7) Kahneman, D. *Thinking, fast and slow*. Macmillan, 2011.
8) Kaplan, A. M. and Haenlein, M. The early bird catches the news: Nine things you should know about microblogging. *Business Horizons*, 54(2):105 – 113, 2011.
9) Pourebrahim, N., Sultana, S., Edwards, J., Gochanour, A., and Mohanty, S. Understanding communication dynamics on twitter during natural disasters: A case study of hurricane sandy. *International Journal of Disaster Risk Reduction*, 37:101176, 2019.
10) Son, J., Lee, J., Oh, O., Lee, H. K., and Woo, J. Using a heuristic-systematic model to assess the twitter user profile’s impact on disaster tweet credibility. *International Journal of Information Management*, 54:102176, 2020.
11) Stieglitz, S. and Dang-Xuan, L. Political communication and influence through microblogging–an empirical analysis of sentiment in twitter messages and retweet behavior. In *2012 45th Hawaii International Conference on System Sciences*, pages 3500–3509. IEEE, 2012.
12) Stokes, C. and Senkbeil, J. C. Facebook and twitter, communication and shelter, and the 2011 tuscaloosa tornado. *Disasters*, 41(1):194–208, 2017.
13) Tversky, A. and Kahneman, D. Heuristics and biases: Judgement under uncertainty. *Science*, 185(4157):1124–30, 1974.
14) Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008, 2017.
15) Zhang, C., Fan, C., Yao, W., Hu, X., and Mostafavi, A. Social media for intelligent public information and warning in disasters: An interdisciplinary review. *International Journal of Information Management*, 49:190–207, 2019.
16) 宮永真央 and 広兼道幸. 人物評価手法に基づく twitter 情報の情報判断に関する研究. **ファジィシステムシンポジウム講演論文集**, 28:1161–1166, 2012.
17) 佐藤翔輔 and 今村文彦. 2018 年西日本豪雨災害における 「# 救助」 ツイートの実態: 2017 年 7 月九州北部豪雨災害との比較分析. **自然災害科学**, 37(4):383–396, 2019.
18) 泉翔太, 堀太成, 山根達郎, 全邦釘, 藤森祥文, and 森脇亮. Deep learning を用いたマイクロブログ投稿文の災害情報分類. ***AI*・データサイエンス論文集**, 1(J1):398–405, 2020.
19) 相田慎, 新堂安孝, and 内山将夫.「東日本大震災関連の救助要請情報抽出サイト」 による救助活動支援. **自然言語処理**, 20(3):405–422, 2013.
20) 湯沢昭夫, 小林亜樹, et al. 災害時における現地情報 tweet 抽出手法. **第 *79* 回全国大会講演論文集**, 2017(1):459–460, 2017.
21) 藤代裕之, 松下光範, and 小笠原盛浩. 大規模災害時におけるソーシャルメディアの活用―情報トリアージの適用可能性. **社会情報学**, 6(2):49–63, 2018.
22) 内田理 and 宇津圭祐. 災害時のソーシャルメディア利活用. **電子情報通信学会 基礎・境界ソサイエティ** *Fundamentals Review*, 13(4):301–311, 2020.
23) 馬場正剛, 鳥海不二夫, 榊剛史, 篠田孝祐, 栗原聡, 風間一洋, 野田五十樹, and 大橋弘忠. ソフトクラスタリングを用いた災害情報の分類. In **人工知能学会全国大会論文集 第 *29* 回全国大会** *(2015)*, pages 2B3NFC02a3–2B3NFC02a3. 一般社団法人 人工知能学会, 2015.
24) 福島裕斗, 桝井文人, and Michal, P. マイクロブログを対象とした情報トリアージに関する研究. In **北見工業大学 修士論文**, 2015.
25) 北島良三, 上村龍太郎, 内田理, and 鳥海不二夫. ニューラルネットワークを用いた tweet データの分類に関する研究. In **人工知能学会全国大会論文集 第 *29* 回全国大会** *(2015)*, pages 2B3NFC02a2–2B3NFC02a2. 一般社団法人 人工知能学会, 2015.
26) 凌摩, 川., 光範, 松., 晨潔, 宋., and 裕之, 藤. Twitter からの救助要請の抽出と検証―2018 年 7 月の西日本豪雨災害ツイートを対象として―. *DEIM Forum 2019 I7-5*, 2019.