ICCEPM 2020
The 8th International Conference on Construction Engineering and Project Management
Dec. 7-8, 2020, Hong Kong SAR

Real-time Knowledge Structure Mapping from Twitter for Damage Information Retrieval during a Disaster

Jiu Sohn[1]*, Yohan Kim[2], Somin Park[3], Hyoungkwan Kim[4]

[1] *Department of Civil and Environmental Engineering, Yonsei University, Seoul, Korea,* E-mail address: jiujohn@yonsei.ac.kr
[2] *Department of Civil and Environmental Engineering, Yonsei University, Seoul, Korea,* E-mail address: homez815@yonsei.ac.kr
[3] *Civil and Environmental Engineering, University of Michigan, Michigan, United States of America,* E-mail address: somin@umich.edu
[4] *Department of Civil and Environmental Engineering, Yonsei University, Seoul, Korea,* E-mail address: hyoungkwan@yonsei.ac.kr

Abstract: Twitter is a useful medium to grasp various damage situations that have occurred in society. However, it is a laborious task to spot damage-related topics according to time in the environment where information is constantly produced. This paper proposes a methodology of constructing a knowledge structure by combining the BERT-based classifier and the community detection techniques to discover the topics underlain in the damage information. The methodology consists of two steps. In the first step, the tweets are classified into the classes that are related to human damage, infrastructure damage, and industrial activity damage by a BERT-based transfer learning approach. In the second step, networks of the words that appear in the damage-related tweets are constructed based on the co-occurrence matrix. The derived networks are partitioned by maximizing the modularity to reveal the hidden topics. Five keywords with high values of degree centrality are selected to interpret the topics. The proposed methodology is validated with the Hurricane Harvey test data.

Key words: Damage information retrieval, Deep learning, Disaster management, Knowledge structure, Twitter

1. INTRODUCTION

Disaster has caused severe impact on human society including damage to property, business interruption, and casualties [1]. These damages can lead to huge economic losses as well as restrictions on social activities. As the frequency and intensity of disasters increase due to climate change, the scale of social and economic damage caused by disasters is increasing [2]. It is required for an effective disaster management system to alleviate the enormous social and economic damage originated from disasters. Establishing a response plan is a particular step of disaster management that refers to a process of suppressing the spread of damage during a disaster. Since the spread of the damamge are directly correlated to the volume of the social and economic losses, an adequate response plan should be established in a timely manner. When it comes to organizing a timely response plan, the critical factor is to promptly obtain high-quality situational information. Thus there is a need for a channel where it constantly provides high-quality situational information

Twitter is one of the suitable channels to obtain abundant information persistently as it is characterized by the active sharing of situational information among users who are scattered across large areas in disaster situations. Although a larege amount of information is underlain on Twitter, the

information includes both relevant and irrelevant inforatmion to understand the damage situation. The process of classifying relevant information should be carried out to completely utilize Twitter as medium to recognize damage in the society. The relevant information classified by time contains several types of damage. It is crucial to detect the types of damage as they give a clue for which damage type should be focused on priority when establishing a response plan. Attemps have been made to detect topics shared on Twitter [3]. However, there were few studies that focused on the topic change in damage-related information despite this information is the basis for establishing a response plan according to time.

Knowledge structure mapping refers to the process of defining the interrelationships of concepts that make up a single knowledge. This concepts have been widely used to discover research trend in various domains such as library and information science, medical education [5, 6]. As well, this concept can be utilized to identify topics by keywords from a large amount of data produced on Twitter. Therefore, this paper proposes the knowledge structure mapping methodology that effectively identifies the topic change in damage-related information. Firstly, the damage-related information is classified by the state-of-the-art deep learning model, Bidirectional Encoder Representations from Transformer (BERT) [7]. Then the topics for damage-related information are identified with the keywords based on the network theory. The proposed methodology is validated through the Hurricane Harvey test data.

2. METHODOLOGY

2.1. Damage-related tweets classification through a BERT-based approach

In the early days when deep learning was applied for natural language tasks, models such as SVMs, Naïve Bayes, and CNN were dominant. However, theses models had limitations that they could not reflect word order information well. If the order information of words is not reflected, it is impossible to grasp the entire context. In other words, those models could not capture the different usage of the the same words in different sentences. Recurrent neural networks (RNNs) and long short-term memory networks (LSTMs) have been proposed to reflect word order information. These models are recurrent models, using the previous hidden state to compute the current output. However, they have a disadvantage in terms of forgetting prior information when the text is long. Transformer, unlike the existing RNNs or LSTMs, learns the dependency of input and output using attention mechanism instead of recurrence. This compensated for the loss of information when the preceding and subsequent words were far apart.

In this study, the initialized pre-trained parameters of BERT were fine-tuned to reflect the semantic representation of damage-related tweets via transfer learning. The BERT-Base model was employed to achieve the given task. The specific task was to classify a tweet into one of the four classes that are related to human damage, infrastructure damage, and industrial activity damage: “Dead, Injured people,” “Found people, Evacuation, Rescued,” “Infra, Industrial activity related,” and “Missing, Displaced, Trapped people”. The data downloaded from CrisisNLP [8] and CrisisLex [9] were used for the model fine tuing. The number of labeled tweets for each class is summarized in Table 1. In order to determine the optimal hyperparameters, the data was split into train and validation data by an 80:20 ratio. The optimal hyper parameters were determined when the loss of the validation dataset was minimized. The damage-related tweets in the Hurricane Harvey (August 25, 2017, 14:44~15:13) were downloaded from creative commons (open source data with no copyright) for the test data to evaluate the performance of the BERT model.

Table 1. Data description for model training and test.

Class	Training data	Test data
Dead, Injured people	338	6
Found people, Evacuation, Rescued	291	93
Infrastructure, Industrial activity related	274	181
Missing, Displaced, Trapped people	162	4

2.2. Knowledge structure mapping

The knowledge structure mapping refers to the process of discovering topics that are discussed in "Dead, Injured people," "Found people, Evacuation, Rescued," "Infra, Industrial activity related," and "Missing, Displaced, Trapped people" by keywords with the passage of time. The relationship between words is represented with nodes and edges based on the co-occurrence matrix. The node stands for a word and the edge stands for the number of the co-occurrence between words. The length of the edge indicates how strongly the nodes are related. The smaller the edge length, the stronger the node is connected. In other words, the edge lengh is small when the number of the co-occrurrence is large. Community detection is a way of partitioning the network on the basis of modularity, the criterion to measure the quality of the partition. This paper performed the community detection based on the algorithm suggested by Blondel et al. [10].

3. RESULT AND DISCUSSION

The values of optimal hyperparameters were as follows: 128 for sequence length, 30 for batch size, 7 for epochs, and 1e-4 for learning rate. Using the BERT model fine-tuned with the given data, the Hurricane Harvey test data was categorized into the 4 classes. Accuracy, precision, recall, and F1-score values were obtained for each class. The results are summarized in Table 2. The macro average of the F1 scores is 85% and the weighted average (considering the number of data in each class as weight) of F1 score is 96%.

The network was visualized with the open software Gephi [11], as shown in Figure 1. The size of the node is proportional to the frequency of nodes and the thickness of the edge reflects the closeness between words. The nodes that are clustered into the same community are expressed in the same color. The set of the words with high centrality values are the keywords that stand for the clustered community. In this study, the top five keywords with high centrality values were selected to interpret the topic of the community. The keywords that made up the communities in "Infrastructure, Transportation damage" are summarized in Table 3. The topics for the six communities were interpreted as follows: "Gas price rise" for community 1, "Airport damage due to storm" for community 2, "road and pump damage" for community 3, "The open of border patrol checkpoint" for community 4, "Interruption of communication" for community 5, "postponement of sporting Kansas City (KC) match" for community 6, and "nuclear power in damage" for community 7.

Table 2. Precision, recall, and F1 score for four classes.

Class	Precision	Recall	F1 score
Dead, Injured people	0.80	0.67	0.73
Found people, Evacuation, Rescued	0.99	0.95	0.97
Infrastructure, Industrial activity related	0.96	0.99	0.98
Missing, Displaced, Trapped people	0.75	0.75	0.75
Macro average	0.88	0.84	0.85
Weighted average	0.96	0.96	0.96

Table 3. The keywords for the communities in "Infrastructure, Industrial activity related".

Community	Keywords
1	gas, prices, texas, us, rise
2	storm, surge, airport, superstorm, expect
3	water, approaches, capacity, pumps, road
4	Open, patrol, border, checkpoints, stay
5	Stop, working, cell, phones, daysweek

6	Houston, postponement, match, kc, sporting,
7	Power, tx, threatens, nuclear, plant

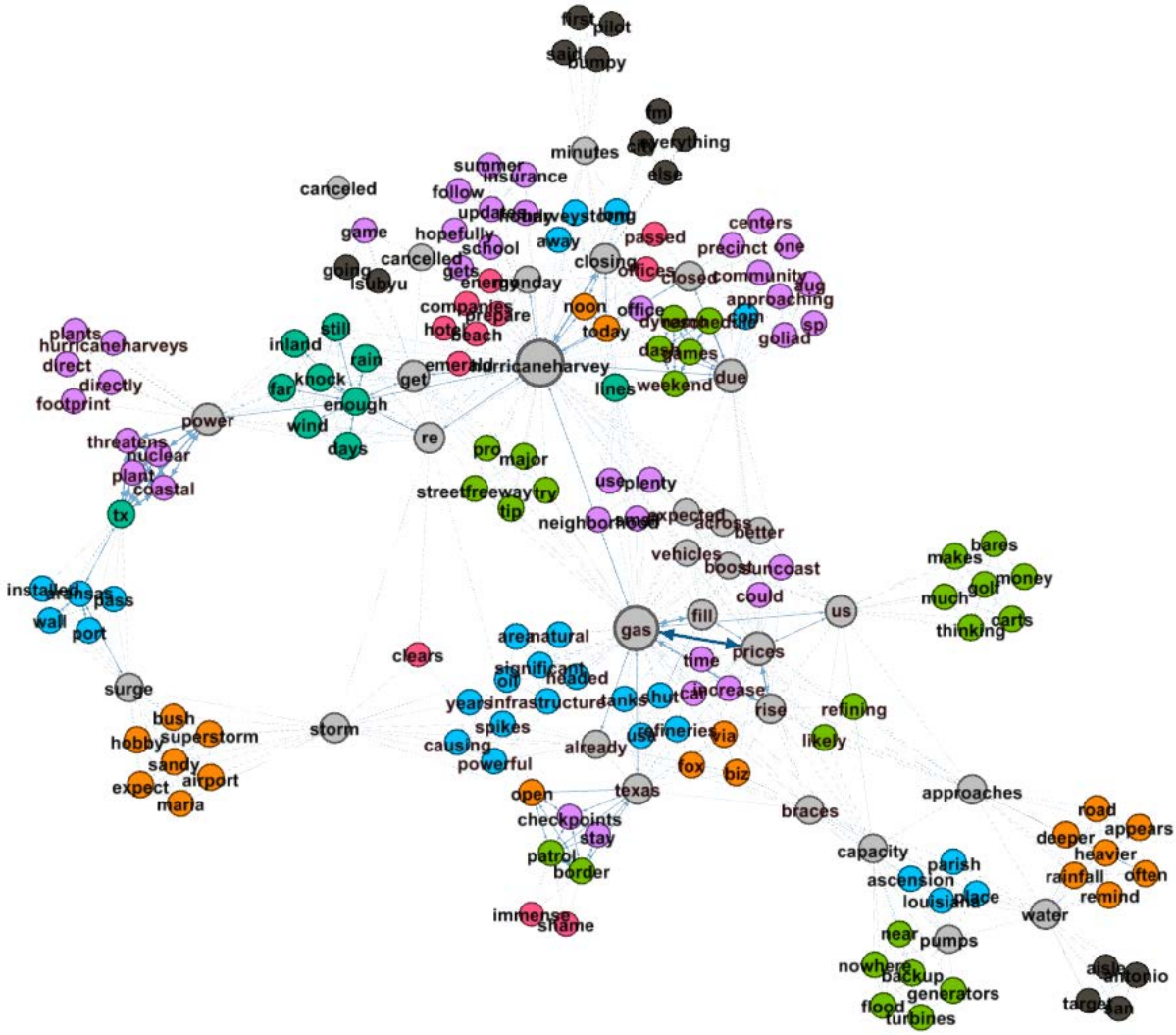

Figure 1. Knowledge stucture for “Infrastructure, Industrial activity related”.

4. CONCLUSION

The damage information shared on Twitter can be used for establishing a response plan in a timely manner. However, it is a challenging task to identify the damage-related topics in the stream of large data. To address this problem, this paper proposes an effective methodology to identify the damage-related topics in real-time by constructing a knowledge structure with the combination of the BERT-based classifier and the community detection techniques. A case study of Hurricane Harvey validated the applicability of the proposed methodology. The proposed model can be improved by collecting supplementary tweets to reflect the diverse expressions used in the damage-related tweets. The geospatial information could be used with the derived knowledge structure to specify the area where the specific damage happened. With the improvemet, the knowledge structure obtained from the study is expected to be used as a basis for preparing an efficient countermeasure to disasters.

ACKNOWLEGEMENTS

This work was supported by National Research Foundation of Korea (NRF) grants funded by the Ministry of Science and ICT (No.2018R1A2B2008600) and the Ministry of Education (No.2018R1A6A1A08025348).

REFERENCES

[1] V.Meyer, N.Becker, V.Markantonis, R.Schwarze, “Assessing the costs of natural hazards-state of the art and knowledge gaps”. Natural Hazards and Earth System Sciences, 13.5: 1351-1373, 2013.

[2] M.Coronese, F.Lamperti, K.Keller, F.Chiaromonte, A.Roventini, “Evidence for sharp increase in the economic damages of extreme natural disasters”. Proceedings of the National Academy of Sciences, 116.43: 21450-21455, 2019.

[3] S. Gründer-Fahrer, A. Schlaf, G. Wiedemann, G. Heyer, “Topics and topical phases in German social media communication during a disaster”, Natural language engineering, 24 221-264, 2018.

[4] Loh, Adrian Sin Loy, and R. Subramaniam. "Mapping the knowledge structure exhibited by a cohort of students based on their understanding of how a galvanic cell produces energy." Journal of Research in Science Teaching 55.6: 777-809, 2018.

[5] X. Chen, J. Chen, D. Wu, Y. Xie, J. Li, “Mapping the research trends by co-word analysis based on keywords from funded project”. Procedia Computer Science, 91, 547-555, 2016.

[6] M. Srinivasan, M. McElvany, M. Shay, J. Shavelson, C. West, “Measuring knowledge structure: Reliability of concept mapping assessment in medical education.”, Academic Medicine, 83(12), 1196-1203, 2008.

[7] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, “Bert: Pre-training of deep bidirectional transformers for language understanding.”, arXiv preprint arXiv:1810.04805, 2018.

[8] Imran, Muhammad, Prasenjit Mitra, and Carlos Castillo. “Twitter as a lifeline: Human-annotated twitter corpora for NLP of crisis-related messages.”, arXiv preprint arXiv:1605.05894, 2016.

[9] Olteanu, Alexandra, Sarah Vieweg, and Carlos Castillo. “What to expect when the unexpected happens: Social media communications across crises.”, Proceedings of the 18th ACM conference on computer supported cooperative work & social computing, 2015.

[10] Vincent D Blondel, Jean-Loup Guillaume, Renaud Lambiotte, Etienne Lefebvre, “Fast unfolding of communities in large networks.”, in Journal of Statistical Mechanics (10), P1000, 2008

[11] Bastian, Mathieu, Sebastien Heymann, and Mathieu Jacomy. "Gephi: an open source software for exploring and manipulating networks." Icwsm 8.2009 361-362, 2009.