Rapid Acquisition and Analysis of Disaster Information from Social Media

You-Rui Liu (1), Pai-Hui Hsu(2)

[1 2] Department of Civil Engineering, National Taiwan University,
No. 1, Sec. 4, Roosevelt Rd., Taipei 10617, Taiwan.
Email: r06521816@ntu.edu.tw; hsuph@ntu.edu.tw

KEY WORDS: Disaster, Social Media, Text Analysis

ABSTRACT: In recent years, the social network has flourished. Many sources of information are transmitted through the social network. In addition to being a tool for people to connect, it also plays an important role in disaster response. In the past, most users usually passively obtained information through the Internet. With the rise of social media, they can actively deliver messages and become information providers. In the event of disasters, using the social network to actively deliver information becomes a very beneficial tool.

With the development of technology, as long as the disaster information is transmitted through the Internet, search and rescue units can immediately grasp the situation of the disaster and make a more effective disaster response. There are many kinds of disasters. The disasters that occur frequently in Taiwan are mostly earthquakes, floods, etc. The most important thing is to get disaster information in real-time. The immediate acquisition of disaster information depends on the immediate return of people at the disaster site so that the search and rescue units can grasp the time to rescue and let the disaster not expand.

This research focuses on the application of social network to collect disaster information and analyze disaster information. Through the immediacy of the social network, it can quickly collect disaster information and immediately deal with it. For the disaster information, further textual relevance analysis can be used to identify relevant disasters that may be derived.

1. INTRODUCTION

In recent years, social networking has flourished. Many sources of information are transmitted through the social network. In addition to being a tool for people to connect, when disasters occur, social media also plays an important role. In the past, most Internet users usually passively obtained information through the Internet. With the rise of social networking sites, in addition to passively receiving information, they can actively publish and deliver messages and become information providers. In the event of a disaster, using the social network to actively publish information becomes a tool that is very beneficial to disaster relief.

Facebook (Facebook), PTT, Twitter (Twitter), etc. can all be called social networks. With the increasing use of the population, it is a trend to establish links through social networks. Compared with the traditional reward for disasters, people can only notify the disaster response center by phone. Concerning the immediacy of disaster relief, with the development of technology, as long as the disaster information is transmitted through the Internet, relevant search and rescue units can immediately grasp the disaster situation and adapt to local conditions, and make more effective disaster response measures. Therefore, the study focuses on the application of social network to collect disaster information and analyze disaster information. It is hoped that through the immediacy of the social network, it can quickly collect disaster information, determine the location of the disaster, and immediately deal with it. For the disaster information, further textual relevance analysis can be used to identify relevant disasters that may be derived.

Taiwan is located at the junction of the Eurasia plate and the Philippine sea plate and is located in the low latitudes of the western Pacific, so there are often natural disasters. The natural disasters that often occur in Taiwan are earthquakes and floods. In summer, it is the rainy season in Taiwan. Whether it is the typhoon intrusion from July to September or a large amount of rainfall brought by the southwesterly airstream, it often poses a great threat to Taiwan and causes heavy damage. When the windstorms occur, the most common voice expressed by the victims is that the government can quickly help them rebuild their homes and minimize the damage. Therefore, when disasters occur, immediate disaster relief has become a top priority. In the past, victims were often limited by resources underdeveloped, and they were unable to immediately report the disaster. They could only notify the disaster relief center by telephone and wait passively for rescue. If the disaster relief center is busy, it cannot be immediately It is often known that the disaster-stricken areas have delayed the disaster relief time. In today's society where communication software is popular, if you can use resources such as communication software or social networking, you can immediately report disaster information and use the social network. The ability to receive a large number of messages is expected to significantly reduce search and rescue time and minimize damage.

2. LITERATURE REVIEW

2.1 Text Analysis

Text analysis is to analyze non-structural texts, convert them from unstructured data into structured data, and obtain useful information from them. Commonly used analysis methods can be roughly divided into classification analysis, cluster analysis, and association rules analysis. With the development of technology, text analysis has also been developed using machine learning methods.

When processing unstructured data, we need to perform Pre-Processing. Taking Chinese text pre-processing as an example, the data pre-processing steps are to select target data, to break words and break sentences, and then to analyze and use the text through some statistical methods and algorithms to obtain the necessary information. As a reference for decision making.

2.2 Data Quantification

The text data is presented in numerical values, which can be used to analyze the data and convert the data into usable information. The process is called data quantification. There are many methods for data quantification, and it can also be combined with machine learning to quantify data. Word2vec is a tool for analyzing Chinese characters and quantifying data. It is also a part of machine learning.

This study uses Word2vec for text analysis. Word2vec is a tool developed by Tomas Mikolov on google in 2013. It is an algorithm that turns a single text into a feature vector. In addition to converting text into vectors, it can also read words in the word. Semantics, and links similar words, the following will explain the algorithm.

Word2vec converts "words" into "vector" form, which can simplify the processing of text content into vector operations in vector space, calculate the similarity in vector space, and use it to represent the similarity of text semantics. The distance between the word and the word is calculated according to the input "set of words", where Word2vec uses a three-layer neural network, and the size of the input layer and the output layer is the number of words of the trained text, and the size of the middle layer is for the size of the vector you want to compress, the schematic is shown in Figure 1 below.

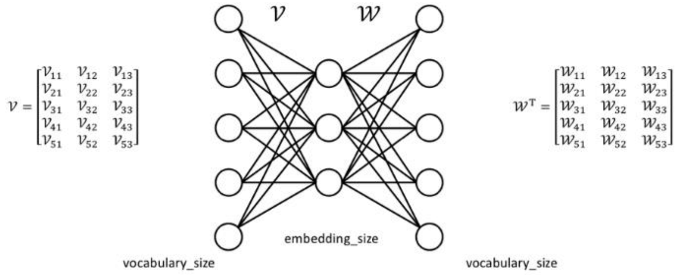
v
w
$v = \begin{bmatrix} v_{11} & v_{12} & v_{13} \\ v_{21} & v_{22} & v_{23} \\ v_{31} & v_{32} & v_{33} \\ v_{41} & v_{42} & v_{43} \\ v_{51} & v_{52} & v_{53} \end{bmatrix}$
$w^{\mathsf{T}} = \begin{bmatrix} w_{11} & w_{12} & w_{13} \\ w_{21} & w_{22} & w_{23} \\ w_{31} & w_{32} & w_{33} \\ w_{41} & w_{42} & w_{43} \\ w_{51} & w_{52} & w_{53} \end{bmatrix}$
embedding_size
vocabulary_size
vocabulary_size

Figure 1 Word2vec Calculation Diagram (Chen,2017)

The Word2vec input layer method can be divided into two models: Skip-gram and Continuous Bag of Words (CBOW). The Skip-gram model uses a word as input to predict the context around it. The CBOW model uses a word. The context is used as input to predict the word itself, that is, to use the words of the context as input to the neural network, and finally to predict the word of the target. The logic of Skip-gram is to input only one word at a time, and the output label is the text within a certain distance before and after it, so the same word will have multiple labels, and the distance between the front and back texts is also one of the skip-gram parameters; CBOW is The middle word of a sentence is treated as a label, the left and right characters are input words, multiple words are used as input values, and a label output value is produced, and the length of the sentence can be adjusted.

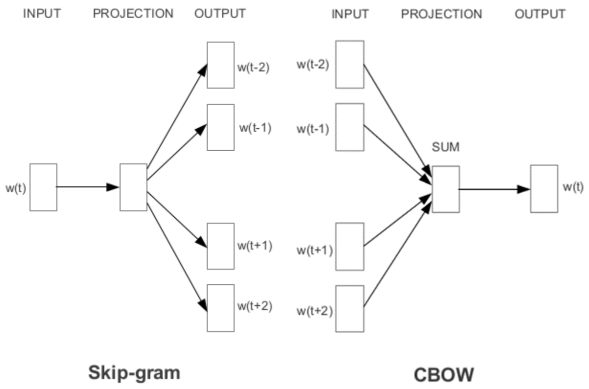

Figure 2 Skip-gram and CBOW Diagram

The output has two forms, the Hierarchical softmax model and the Negative sampling model. Hierarchical softmax uses hierarchical classification and expresses the results in a Huffman tree. Negative sampling uses the method of negative sampling. To obtain the word vector of each word, compared with the two output models, Negative sampling can be used to improve the training speed and improve the quality of the resulting word vector, using random negative sampling to greatly improve performance. Word2vec calculates the cosine of two words, which represents the relevance of two words. The judgment index is cosine similarity. The value range is between 0 and 1.

3. RESEARCH METHODS

The research process is roughly to use the social network PTT to collect disaster information in real-time. Through the information and messages posted by the public on the social network, the information and keywords can be extracted using information climbing, and the content is analyzed and the disaster is obtained. After the information such as the location, time and disaster description, the time and space analysis of the disaster was carried out.

3.1 Selecting a Community Site and Getting Information from the Site

Social network is booming and this study intends to use PTT as the main source of disaster data collection. PTT has many types of boards. PTT gossip board is used as the main board for obtaining data, and seismic information is obtained from it.

3.2 Grabbing Disaster Messages

After obtaining the articles related to the disaster, the information of the disaster must be captured. The main provider of the disaster information is the netizens on the PTT board, so it is necessary to capture the message provided by the netizens under the article. Taking the earthquake disaster article on April 18, 2019 as an example. Write the program, extract the message from the article on the board, save it and save it in Excel format, and store it according to the label category, user name, message information, date and time.

3.3 Using Jieba Segement to Analyze

Since the captured message material belongs to unstructured data, it is necessary to analyze the obtained message and convert the data from unstructured data into structured data. The main method of Chinese text analysis is the analysis of broken words. This study uses the Jieba word-breaking module to analyze the word-breaking words and count the number of occurrences of the words, which can be used to estimate the location information of the earthquake. The stuttering system can be divided into full mode, precise mode, search engine mode and the like. The full-mode hyphenation method is to scan and segment the words in the sentence that can be word-formed, and the segmentation speed is fast, but the ambiguity cannot be solved. The exact mode of the hyphenation method is to separate the sentences most accurately and is suitable for Contextual Analysis. The word breaker mode of the search engine mode is to segment the long words again based on the precise mode. The word-breaking method adopted by this research is the precise mode. Because the analysis of the article messages on the social network is part of the text analysis, it is more appropriate to use the accurate mode for analysis. After the word segmentation is completed, the number of occurrences of each word can be calculated together, and according to the location information in each word, the number of occurrences is determined, and the location of the estimated disaster is determined.

3.4 Performing Text Quantification Analysis

Since the message materials of the netizens on the social networking site are mostly unstructured, they must be converted into structured data using text analysis and obtain available information. This study uses the Word2vec tool for text analysis, and Word2vec is an algorithm that turns a single text into a feature vector to quantify the text data. This study uses the Skip-gram and Negative Sampling models in Word2vec to replace the traditional Huffman tree and use random negative sampling to greatly improve performance, find out the vocabulary related to disaster keywords, and analyze the correlation. It can interpret other disasters that may be derived from them. The analysis target can be roughly divided into two parts. The first part is to randomly analyze different numbers of data for the same earthquake case, and the second part is to analyze and compare different earthquake cases.

3.5 Data Visualization

Data Visualization presents data and data in a visual manner. Through the way of image, it is easier to distinguish the laws, trends, and relevance of data. This study presents a large amount of textual data graphically, while common visual methods include word cloud drawing, association drawing, vocabulary distribution drawing, etc.

4. EXPERIMENTS

4.1 Experimental Data

This experimental data was taken from the social networking site PTT as a source of information. From the PTT gossip board, articles about earthquakes were collected, and three articles were taken from the middle of the experiment as experimental analysis objects.

The first one was the earthquake that occurred at 8:44 pm on March 17, 2019. The article is available from https://www.ptt.cc/bbs/Gossiping/M.1552826711.A.6BB.html

The second part is the earthquake that occurred at 1:00 pm on April 18, 2019, in the Republic of China. The article is available from https://www.ptt.cc/bbs/Gossiping/M.1555563687.A.CD3.html

The third part is the earthquake that occurred at 6:44 pm on May 2, 2019 in the Republic of China. The article is available from https://www.ptt.cc/bbs/Gossiping/M.1556793906.A.188.html.

4.2 Data Capture

The article information of the social networking site is captured by the board. After the crawling, we can using Excel to save and see the amount of data crawled.The earthquake that occurred at 8:44 pm on March 17 grabbed 195 messages. The earthquake that occurred at 1:00 pm on April 18 grabbed 1,269 messages. The earthquake that occurred at 6:44 pm on May 2 grabbed 336 messages.

4.3 Data Analysis

After capturing a large amount of unstructured data, it is necessary to convert unstructured data into structured data to form information that is conducive to interpretation. The following analysis of the project and results.

4.3.1 Jieba Segement: The first step in converting unstructured data into structured data is to analyze the word-breaking. The word-breaking method adopted is the precise mode. Because the object of data analysis is the message of the article on the social network, it is a part of text analysis. Therefore, it is more appropriate to use the accurate mode for analysis. The segmentation method is to break a sentence into several meaningful words according to common words. After the hyphenation is completed, statistics are performed on each word data, and the number of occurrences of each word is calculated. After the statistics are completed, find out the vocabulary containing the location, such as the name of the city, and check the number of occurrences. According to the location information and the number of occurrences in each word, perform location analysis to identify the location of the estimated disaster or the impact of the disaster. Larger area. Table 1 below lists the names of the top five cities and the number of occurrences in the three pieces of earthquakes used in this experiment.

Table 1 The top five cities in each earthquake

03/17 Earthquake		04/18 Earthquake		05/02 Earthquake	
City	Times	City	Times	City	Times
Taipei	18	Taipei	85	Taipei	31
Yilan	5	Taichung	83	New Taipei	10
Taoyuan	2	Taoyuan	47	Yilan	6
Hsinchu	2	Hsinchu	34	Hsinchu	6
Tainan	2	NewTaipei	19	Taoyuan	5

4.3.2 Relevance Analysis: After completing the hyphenation, you can use the data after the hyphenation to complete the correlation analysis of the text, which is the quantitative analysis of the data. The method is Word2vec, and the concept is to convert the words into vectors and use the cosine. Cosine similarity is used for correlation analysis.

The concept of Word2vec correlation analysis is that if there are many similar sentences, the words in the same position in the sentence will be regarded as similar words, and other words related to them can be searched for specific words. Used in disaster prevention analysis of subsequent disasters. The range of similarity is 0 to 1. If the value is closer to 1, it means that the two words of the analysis are similar in sentence structure. The semantic meaning of the two words may be similar, and the correlation between the two is high. The closer to 0, the lower the relevance of the two words.

This experimental test can be divided into three major parts:

(1) For the same case, randomly extract the amount of data of different pens for analysis. The object of this experiment was analyzed by the earthquake that occurred on April 18, 2019. 250, 500, and 1269 data in the text were randomly selected for text analysis to discuss the same case and analysis caused by different data volumes. The difference in results.

(2) For the earthquake events of different scales, text analysis is carried out to analyze the differences between different time and different earthquake intensities.

(3) Conducting the relevance analysis of words, intending to input a keyword to be analyzed, and analyzing other words most relevant to the keyword, which can be used as a reference for subsequent disaster prevention.

In each earthquake case of this experiment, the analysis target is selected as the location vocabulary with the most occurrences and the other places to analyze the correlation, and judge whether the model is appropriate. According to the data obtained by the earthquake, the correlation analysis is carried out, the words to be analyzed are input, and the other words with relevance to the target words to be analyzed are calculated by Word2vec, which can be used as a reference for disaster relief. Table 2 shows the text analysis of different data quantities for the same earthquake. In the experiment of 250 data, because the selected threshold of the construction dictionary is the word with more than 5 words, the number of occurrences of Taichung is too low, so it cannot be included in the dictionary for analysis. In the cosine similarity analysis of Table 3, the cosine similarity analysis of the words is performed for each place where the occurrence is frequent. Since the threshold for setting the dictionary is the word with more than 5 words, the seismic data on 03/17 in the place where Taoyuan, Hsinchu, and Tainan appear less than 5 times, the vocabulary does not exist in the dictionary for vector analysis, so it is not within the scope of this analysis. In the correlation analysis of Table 4, the word "earthquake" is used as the keyword to be analyzed, and the first five words with the highest relevance to the "earthquake" are found, which can be used as the disaster correlation judgment.

Table 2 04/18 Seismic cosine similarity analysis (different data volume)

Item	250 data		500 data		1269 data	
Cosine Value	Taipei-Taichung	-	Taipei-Taichung	0.98909	Taipei-Taichung	0.99984
	Taipei-Taoyuan	0.26002	Taipei-Taoyuan	0.97869	Taipei-Taoyuan	0.99985
	Taipei-Hsinchu	0.27486	Taipei-Hsinchu	0.98449	Taipei-Hsinchu	0.99985
	Taipei-New Taipei	0.21598	Taipei-New Taipei	0.97638	Taipei-New Taipei	0.99982

Table 3 Cosine similarity analysis of each earthquake

Item	03/17 Earthquake		04/18 Earthquake		05/02 Earthquake	
Cosine Value	Taipei-Yilan	0.10587	Taipei-Taichung	0.99984	Taipei-New Taipei	0.49737
	Taipei-Taoyuan	-	Taipei-Taoyuan	0.99985	Taipei-Yilan	0.48733
	Taipei-Hsinchu	-	Taipei-Hsinchu	0.99985	Taipei-Hsinchu	0.28789
	Taipei-Tainan	-	Taipei-New Taipei	0.99982	Taipei-Taoyuan	0.27582

Table 4 Word relevance analysis

Item	03/17 Earthquake		04/18 Earthquake		05/02 Earthquake	
Associated words	Earthquake		Earthquake		Earthquake	
	Feel	0.15665	Hualien	0.99993	Taipei	0.40793
	Taipei	0.10007	Very big	0.99993	Rain	0.38608
	Illusion	0.07484	Taoyuan	0.99989	Feel	0.36397
	Yilan	0.03146	Taipei	0.99989	New Taipei	0.36182
	No feel	0.02299	Hsinchu	0.99988	Yilan	0.32970

4.3.3 Different Types of Disasters - Flood Analysis: In addition to analyzing earthquake disasters, it also analyzes different types of disasters. Taking floods as an example, the case was flooded in part of the heavy rains in Taipei on September 8, 107. The source of the article is the PTT gossip version, and the type of disaster is flooding. The article capture URL is as follows:
https://www.ptt.cc/bbs/Gossiping/M.1536401698.A.A2E.html

The steps for performing the analysis are the same as the previous steps, and the article is grabbed and analyzed. This article grabbed 803 comments, and analyzed and counted the first few words that appeared most frequently as shown in Table 5. After the word breaks, the word frequency statistics are counted, and the first seven most frequently appearing words are listed. After the statistics, the observations can be found, and the place names and political-related words appear more frequently.

Table 5 Word Statistic

Item	Words	Times
Words often appeared	Southern	103
	Flooding	95
	Taipei	40
	Ke Fans	39
	Puddle	38
	Eastern District	38
	Water Control	33

Then, the correlation analysis is carried out. Taking the word "flooding" as an example, the vocabulary related to the word is analyzed as shown in Table 6 below. The word similarity analysis was also carried out. The words "Taipei" and "flooding" were selected, as well as "Taipei" and "Eastern District". The results are shown in Table 7.

Table 6 Analysis of the word similarity (taking the word “flooding” as an example)

Item	The related words and similarities of the word “flooding”	
Related word analysis	Flooding	Similarity
	Southern	0.99986
	Ke Fans	0.99985
	Rainfall	0.99983
	Water Control	0.99981
	Eastern District	0.99981

Table 7 Analysis of the word similarity

Item	Words	Similarity
Related word analysis	Taipei-Flooding	0.99978
	Taipei-Eastern District	0.99977

It can be seen from Table 6 and Table 7 that in this flooding situation, the following netizens' messages are mostly based on politically relevant messages, making it difficult to determine the disaster situation. Taking the word "flooding" as an example, words related to flooding are location information (such as southern, eastern, etc.) and politically related information (such as Ke Fans). This situation can be seen in flooding. In the collection of disasters, in addition to location information, netizens discussed politically related issues, such as water drainage measures, construction, etc., and less real disaster information. It can be seen that the disaster data and analysis results obtained by different types of disasters are not the same. The acquisition of seismic data, netizens' messages are mostly disaster locations, feelings, magnitude and feeling of earthquakes, while the flood data is mostly politically relevant messages and discussions.

4.4 Data Visualization

4.4.1 Word Distribution Map: Using the natural language processing suite to draw a vocabulary distribution map, you can view the target keyword words to be analyzed, and the distribution in the text to make the data clearer. The vocabulary distribution map of this experiment was drawn by taking the earthquake that occurred on April 18, 2019as an example. The vocabulary analyzed is the vocabulary commonly used in earthquakes and the six municipalities in Taiwan. Observe the above words in textual materials. The distribution in the middle can quickly determine which areas of the disaster may cause more serious impacts.

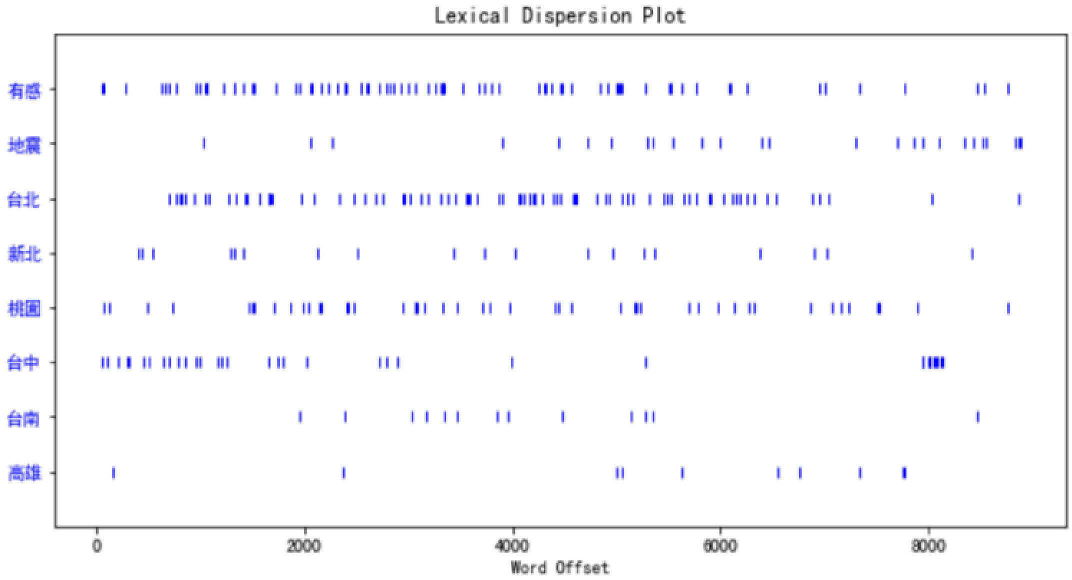

Figure 3 Vocabulary distribution map (take the April 18 earthquake as an example)

4.4.2 Making a Layered Color Map: Seismic data of 03/17, 04/18 and 05/02 will be analyzed for the occurrence of urban times, and a layered color map will be produced. It can be seen more clearly which cities appear more frequently and the color will be deeper. To make data interpretation more clear and intuitive.

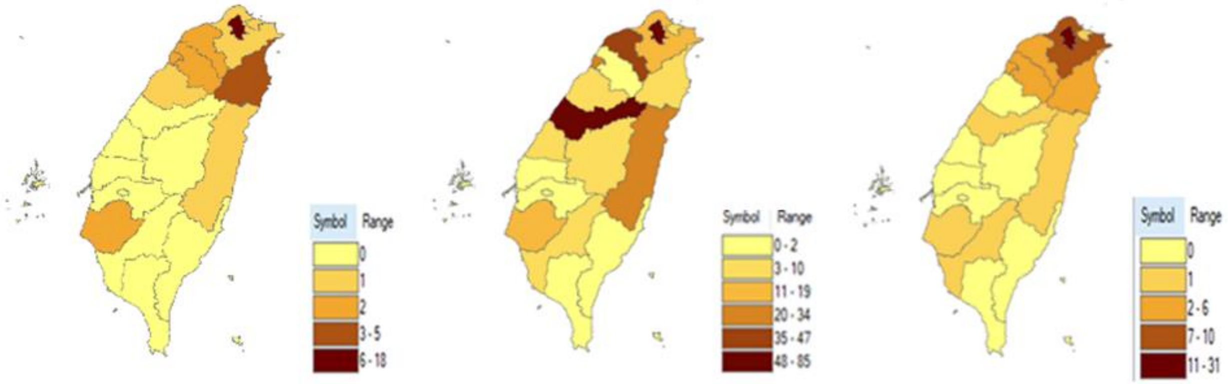

Figure 4 Layered color map of 3/17 (left), 4/18 (middle), 5/2 (right)

4.5 Result Comparison

This experiment analyzes the results of the three earthquakes and summarizes the results of the analysis. The conclusions are as follows. After counting the words, the number of words was counted. It was learned that the earthquakes occurred on 03/17, and the number of occurrences in cities such as "Taipei" and "Yilan" was relatively frequent. Therefore, it can be assumed that the earthquake should occur in North Taiwan. The 4/18 earthquakes occurred in the northern and central cities of Taiwan, and the number of times was too high. Therefore, it can be inferred that the scale of the earthquake should be large and affect all of Taiwan. In the earthquake of 05/02, the top five cities were located in the northern part of Taiwan. Therefore, it is estimated that the earthquake should occur in northern Taiwan, and the number of times is more than 03/17. It can be estimated that the scale of the earthquake should be greater than 03/17 earthquake.

The amount of message data obtained by the three earthquakes was the highest in the 04/18 earthquake, which was 1,269. The minimum number of data was 195 in the earthquake of 03/17. It can be inferred that the earthquake with the largest earthquake is the earthquake of 04/18, the earthquake of 05/02 is the second, and the earthquake of the smallest is the earthquake of 03/17. Comparing the earthquake reports issued by the Central Meteorological Administration after the earthquake, the earthquake information is shown in Table 8. It is known from the table that the earthquake occurred in 03/17, which was 4.2 in size, the smallest of the three cases, and the earthquake occurred in northern Taiwan. The earthquake that occurred on 04/18, with a scale of 6.1, is the largest in the case. The earthquake that occurred in 05/02, the size of the scale is 4.8, the scale is between the former two, and occurred in North Taiwan. It can be seen that the information in Table 8 is consistent with the results of this experimental analysis.

5. CONCLUSIONS

This study uses the social network PTT to collect disaster information in real-time, and to conduct data acquisition and analysis, and visualize the data to achieve the purpose of obtaining the time and place of the disaster and identifying the location of the disaster occurrence or impact. The main work in the future is as follows:
(1) Continuously collect disaster information or related information on social software
(2) Analysis, comparison, and verification for different types of disasters
(3) Collecting more different cases for the same type of disasters, analyzing and comparing them
(4) Analysis and comparison using different methods of text analysis
(5) In the event of an earthquake, the energy is quantified and the numerical results are compared with the accuracy of the government-published information.

6. REFERENCES

Nelli, F., 2015. Python Data Analysis, Apress Publishers, New York, pp. 103-105.
Hajba, G., 2018.Website Scraping with Python: Using BeautifulSoup and Scrapy, Apress Publishers, New York, pp. 32-33.
Tateosian, L., 2015. Python For ArcGIS, Springer Publisher, New York, pp. 13.
Al-Taie, M., 2017. Python for Graph and Network Analysis, Springer Publisher, New York, pp. 49.
Embarak, O., 2018. Data Analysis and Visualization Using Python: Analyze Data to Create Visualizations for BI Systems, Apress Publishers, New York, pp. 85-86.
Goyal, P., Pandey, S. and Jain, K., 2018. Deep Learning for Natural Language Processing: Creating Neural Networks with Python, Apress Publishers, New York, pp. 20-21.
Gerrard, P., 2016. Lean Python: Learn Just Enough Python to Build Useful Tools, Apress Publishers, New York, 20p.
Mukhopadhyay, S., 2018. Advanced Data Analytics Using Python: With Machine Learning, Deep Learning and NLP Examples, Apress Publishers, New York, pp. 10.
Srinivasa-Desikan, B., 2018. Natural Language Processing and Computational Linguistics: A Practical Guide to Text Analysis with Python, Gensim, SpaCy, and Keras, Packt Publishing, Birmingham, pp. 10-15.
Porcu, V., 2018. Python for Data Mining Quick Syntax Reference, Apress Publishers, New York, pp. 13-14