Time and Location Topic Model for analyzing Lihkg forum data

Ao Shen
The University of Hong Kong
Faculty of Engineering, Department of Comupter Science
Hong Kong, China
sao@cs.hku.hk

Kam Pui Chow
The University of Hong Kong
Faculty of Engineering, Department of Comupter Science
Hong Kong, China
chow@cs.hku.hk

***Abstract*—Open Source Intelligence (OSINT) is a choice for collecting information today for law enforcement to monitor illegal activities and allocate police resources effectively. However, massive amounts of public information cannot be analyzed by humans alone and so automatic pre-processing must be performed in advance. In traditional text analysis, the common word segmentation tools do not match the needs in special fields and special words (such as proper nouns, dialects, acronyms, metaphors, and so on). In the context of the Chinese language, we consider the problem of automatically determining the time and location of major public gatherings and demonstrations using public available information. As experimental scenario, we use the Lihkg online forum from August 1st to October 10th, 2019 as a corpus, and propose a topic vectorization method based on character embedding and Chinese word segmentation, using MLP (multi-layer perceptron) neural network as a location topic model. The result proves that the method and the model can correctly identify the time and the location of discussed activities by learning the existing location corpus.**

Keywords—Open Source Intelligence (OSINT), Lihkg, information extraction, neural network, topic modeling

I. INTRODUCTION

OSINT (Open source intelligence), which seeks and obtains valuable information from various public information resources, is an intelligence-gathering method for governments and intelligence services to analyze social opinion, cybercrime and cybersecurity [16]. It is currently the technology of collecting information, due to the low cost and low risk of open source intelligence. With the rapid development of information network technology, the role of open source intelligence is becoming increasingly important. However, the massive amount of public information cannot be analyzed by humans alone. Some pre-processing must take place in advance automatically, and finally analyzed and judged by experts to conclude. In this way, the open source information should include a series of activities such as collection, processing, analysis, distribution, inspection, and evaluation. Here, this paper focuses on the automatic processing, analysis and response of information. Based on the MLP (Multilayer Perceptron) neural network algorithm, this paper tries to automatically identify and extract relevant time and location information of discussed activities in open online forums.

The main source of OSINT is information from media and organizations such as books, newspapers, magazines, television, government, and social websites. In this study, we use the Lihkg forum as a corpus to model the topics of the demonstration time and location. The Lihkg discussion forum [1], called "連登" in Chinese, is a well-known multi-category discussion forum in Hong Kong. Users can post online to share their feelings and comments with others. Some activity organizers will post notification messages in advance on the forum, and supporters or regular users can immediately read, participate, and post their comments. In October 2019, discussions in a particular part of the forum were almost completely referring to public gatherings and demonstrations. Identifying the time and location of an unregistered demonstration is an important problem in police resource management. The purpose of this study is to establish an automatic recognition model of textual features of time and location from public posts. Of course, whether a particular activity constitutes a criminal offense depends on local laws and regulations, however, the situation in Hong Kong in October 2019 was rather dramatic since there were many incidents that hurt other people and public property [18] such that some organizers and defendants have been charged under the Public Order Ordinance [17]. Our motivation stems from the fact that similar conditions can apply in almost any country in the world.

Traditional crime classification methods analyze the textual data and detect crime evidence by keyword matching [19]. By comparing the corpus of crime-related words with the contents of the posts, all posts containing crime-related words can be extracted, and these posts are potential traces of the crime. However, this method has low flexibility, and it is easy to miss some keywords that are not on the list but have similar meanings. Deep neural networks are currently one of the most popular machine learning algorithms in the analysis and processing of textual data. At present, the method of automatically extracting keywords by using a neural network is manually labeling each word of each sentence with a feature (for example, person or location [2]). And then use deep learning to train, learn, and recognize the features after labeling a large number of sentences, which will cost a lot of time and energy. In order to improve this situation, we use MLP (multi-layer perceptron) [3] as a feedforward artificial neural network for topic modeling, to learn location words in a location corpus in order to extract locations from sentences. This research seeks answers to the following three questions: 1) How to do word segmentation of Chinese forum with the method of word embedding? 2) How to extract new words not in the existing corpus? 3) How to extract the details of discussed activities (such as time and location) from online forum data?

The rest of this paper is organized as follows. Section 2 overviews related work. Section 3 illustrates the structure of word segmentation and the location topic modeling. Section 4 introduces the experiment about the application of the model. The last section is the discussion and conclusion of this research.

II. RELATED WORK

This research involves many fields such as text processing and analysis, deep learning, crime investigation, and so on. At present in the field of text processing and analysis, Chinese word segmentation has been the research focus for a long time, especially traditional Chinese and Cantonese are more

difficult compared to simplified Chinese. Gao et al. [4] propose the importance of tokenizing known words and detecting different types of unknown words (for example, logical derivative words, named entities, and other unlisted words). They advocate the needs for multiple segmentation criteria. And different natural language processing applications are necessary for different Chinese word granularity. In order to reduce human effort, automatic word segmentation based on large corpora appeared. Qian et al. [5] propose an automatic word segmentation method based on HMM (Hidden Markov Model) model to tokenize Chinese classic text, and obtain good results on both the word segmentation and the part-of-speech tagging. HMM model has also been widely used in other word segmentation tools (such as jieba).

In the field of crime investigation research, applying machine learning and neural network to crime investigation based on a large amount of crime data can benefit the process of crime analysis. Andropov et al. [6] present a network anomaly detection method based on MLP neural network. They use MLP and real monitoring system to simulate, detect and classify some known network attacks. They show good results of their method in identifications of both known anomalies and new ones. AL-Saif et al. [7] try different features extraction techniques and different classification algorithms to detect and classify crimes based on Arabic twitter posts. They also find out that transforming words into feature vectors and using a machine learning classifier can construct a high-accuracy crime classification model.

Topic refers to the object or starting point of discourse, such as a person or event, and is generally composed of a character, a word, a phrase, or a short sentence [8]. Topic model is a cluster and statistics model that analyze the latent semantic structure of the text by machine learning algorithm [9]. It is mainly used for semantic analysis and text mining and analysis in Natural Language Processing. It can extract the required topic information according to the semantic needs of the analysis. Other researches on time topic model are mainly focused on data that have clear time stamps, such as the release time of the news [10], the creation time of the post on social network platforms [11], and so on. There are few time models that extract time from a sentence. Regarding location topic modeling, the easiest way to extract locations from text is to use a list of location names and then compare the words in the text with those names. This method is difficult to discover new locations, and it does not take the contexts into consideration. But the list of names is still a good starting point when training models to automatically recognize locations in the text. The challenge is not only to identify the place but to remove some ambiguity words that are related to the location topic. Lozano et al. [12] discuss the evolution of topics in social media and their geographical location over time. They propose that in crisis management, knowing the current geographical location of a particular topic can provide important information for resource allocation. And compared with the keyword matching method, the locations extracted by the location topic model is related to the actual location and can provide better geographical classification.

The above researches have promoted the Chinese word segmentation algorithm, and combined machine learning and crime investigation, which actually solved some problems in crime analysis. But these methods are based on supervised learning and require a lot of manually labeled data for training. Therefore, this study will integrate neural network classification methods into high-accuracy time extraction and location topic modeling on the premise of reducing human resources as much as possible.

III. TIME EXTRACTION AND LOCATION TOPIC MODEL

Most of the traditional Chinese text analysis is based on the tokenized data which means the text is already separated to words and phrases. They are limited to the word and sentence level, and topics are extracted from high-frequency words and feature words, so the information extracted is limited. Therefore, we think the researches on the character, sentence, and document level is more important before and after the word segmentation. The main process of topic modeling contains the following four steps: First, crawl the data from the Lihkg forum and enter the basic data processing stage, such as filtering out duplicate data. The second step is to perform word segmentation and extract phrases. We use three word segmentation methods to test, in order to obtain the best word segmentation effect, and can get more meaningful words. The third step is the extraction and analysis of time and location of the discussed topic. To extract time information, we use the regular expression information extraction method with the extraction rules defined by ourselves. The model automatically transforms all kinds of time expressions into a unified format with high accuracy. To extract locations, we use the classification extraction technology. Learning from a location corpus, the neural network algorithm can automatically extract relevant features and identify the other locations. The final step is to integrate and analyze the extracted time and location features.

A. Topic Vectorization: Character embedding and Word segmentation

Text vectorization is the basis of natural language processing, which is the process of changing text from words to numbers and then to vectors. Topic consists of a list of characters, so the topic vector is also started from the character vector on the semantic level. In this way, we discuss the topic vectorization from the following two aspects:

1) Character embedding

Before the Chinese word segmentation process, we first perform a character embedding calculation on the entire forum data and transform the character into vectors on the document layer based on the semantic meaning of the characters. Character embedding is a statistical language model based on neural networks. It can be used to specify the number of words in a window and calculate the probability of a character given its contents. It can also indicate the relationship between the target character and the other characters. Due to the flexibility of neural networks, the same method can be applied to characters, words, sentences, and documents. We adjusted the parameters and weights on the basis of the word2vec algorithm to build a character embedding model, turning each character into a vector, which is beneficial for subsequent calculations and analysis. For example, in the sentence "個/個/都/仲/係/向/維/園/方/向/行 (Every / one / of / them / are / heading / for / Victoria / Park)", the slash "/" is used as the delimiter symbol to split the sentence into a list of characters, and then the vector of each character can be calculated. The cosine similarity between two vectors can show the relationship between the two characters. If the similarity is high, the two characters may often appear together or have the same or similar meaning. For example, "元朗(Yuen Long)" is

a place in Hong Kong. The character of "元 (Yuen)" and "朗 (Long)" has a similarity of 0.6457, which is the highest of the similarities between "元 (Yuen)" and other characters.

Character embedding is a simple 3-layer neural network algorithm similar to word2vec [13]. It can calculate the probability of a character based on the surrounding characters within the window size range, so as to get the relationship between this character and other characters at the semantic level. When calculating the character vector, we set the window size to 5, because the number of characters in the location we need to extract is basically within 5. Locations with more than 5 characters are too detailed, and we only need the first 5 words to know the approximate place. The number of columns (i.e. the length of the vector) is defined by the user, and we set the length to 100. This is a relatively balanced length. It is not too short to distinguish from other vectors, and not too long for programming operations and running time. The larger the length of the vector, the more accurate the result may be, but the calculation time will be longer. Each character has a one-dimensional vector, so that the difference between the two characters can be distinguished based on the relationship between two adjacent characters. Later, we will treat each word, phrase or even sentence as it is composed of more than one character.

2) Chinese word segmentation

Word segmentation method is required to extract the words, phrases, or phrases related to location features. Therefore, we have considered the following three segmentation methods:

a) Jieba Tokenizer: Jieba is a commonly used Chinese word segmentation tool. It is based on a large number of corpora for semantic analysis, as well as algorithms like word maps, maximum probability path, and so on. It can automatically identify common words. However, the data used in this research is from the Lihkg forum in Hong Kong. Most of the texts are spoken Cantonese and traditional Chinese. Jieba works well based on the training results of a large number of simplified Chinese texts. It does not have strong adaptability in spoken Cantonese, so it is difficult to tokenize the words correctly, especially for some OOV (out-of-vocabulary) words. Therefore, we have added two other word segmentation methods to find n-grams words and phrases.

b) Co-occurrence frequency: We use the following formula of Gensim [14] phrases:

$$\text{intensity} = \frac{count(AB) - count_{min}}{count(A) \times count(B)} \times N > threshold \quad (1)$$

Here, $count(A)$, $count(B)$, and $count(AB)$ respectively represent the frequency of character A, character B, and word AB. N is the total number of characters in the corpus. $count_{min}$ represents the minimum frequency of words. Threshold represents the threshold for word segmentation. When the co-occurrence intensity of the word AB is greater than this threshold, the word AB can be divided into one word. After many experiments, we finally set $count_{min}$ as 20 and threshold as 0.5. Applying this formula in 5 times iteratively, we can get the words and phrases from 2-grams to 5-grams.

c) Co-occurrence strength: The method to calculate the co-occurrence strength of n-grams is as follows:

$$strength = \min\left\{\frac{P(ABC)}{P(AB)P(C)}, \frac{P(ABC)}{P(A)P(BC)}\right\} \quad (2)$$

Here, $P(A)$, $P(AB)$, and $P(ABC)$ respectively represent the probability of character A, word AB, and word ABC. The significance of calculating the co-occurrence strength of the characters A, B, and C is that if ABC can form a word, then the co-occurrence rate of AB and BC should be high. If AB and C have a little relationship, then the value of $P(AB)P(C)$ should be close to $P(ABC)$. The larger the value of $P(ABC)/(P(AB)P(C))$, the more likely it is that ABC appears as a word. Similarly, after multiple iterations, we can also get words and phrases from 2-grams to 5-grams.

After all the words and phrases are tokenized, they have the same method to calculate the vectors of words and phrases. The vector of a word or phrase is equal to the sum of all the vectors divided by the number of characters:

$$vector(w) = \frac{\sum_1^n v_i}{n} \ (v_i = [d_{i1}, d_{i2}, d_{i3}, \dots, d_{i100}]) \quad (3)$$

The word w consists of the characters $c_1, c_2, \dots, c_n$. And their vectors are $v_1, v_2, \dots, v_n$. When the three word segmentation methods find the same word, the vectors of the word are the same, so that the difference between the three segmentation methods can be more clearly distinguished in subsequent extraction of locations.

The content of a post can consist of one or multiple topics, thus forming a topic framework. After obtaining each topic and the corresponding vector, on the document level, the content of a post can form a topic matrix. Topic vectorization is to convert a topic matrix into a vector matrix, and each row in the matrix is a vector of a topic. We need to find location-related topics in this vector matrix.

B. Time topic and location topic model

In the October 2019, most of the posts in the Lihkg Current Affairs section were centered around the organizing of and comments on demonstrations and rally activities. Many of the posts that mention time and location at the same time were related to very specific assembly activity. We collected not only the main content of a post but also the reply of the main content. For example, there is a main post like "Date: 24/8/2019. Time: 1300 gathering. Location: Tsun Yip Street Playground". Except some spam replies like "Up up" to bump a post, there are also a lot of replies like "824 Kwun Tong"(The playground is in Kwun Tong). From our experience of reading the posts on Lihkg, all the posts in Current Affairs section with time and location were treated as demonstration-related posts and no further pre-selection was performed. Note that the time of the discussed event is not necessarily the same as the timestamp of the post.

With our focus on time and location we need two models to analyze the topic of time and location. As for time topics, we use information extraction technology based on regular expression. The extraction rules are formulated based on the features of time expression and the model can automatically perform matching. The location topic model is constructed based on the location names and then uses neural network algorithms to automatically learn and extract locations from the posts.

1) Time topic extraction

We need to extract time directly from the text, so the time topic model is built based on the format and characteristics of the time mentioned in the post content.

In the Lihkg posts, the time is mainly expressed in the following two ways: a) formal expression, with symbols

between the year, month, and day, similar to YYYY-MM-DD or YYMMDD, such as 2019-01-08, 08/01/2019, 190108, and so on. b) Abbreviated form, without year, similar to MMDD or MM/DD, such as 0928, 9/28, September 28, and so on. The format of time is limited and it can be extracted regularly, so we use regular expression to extract the time based on these rules. For example, for a date like 928, we can use the regular expression "(([1-9]|1[0-2])(0[1-9]|[1-2][0-9]|3[0-1]))" to extract the month and day and convert the date into the formal expression like " 2019-09-28 ". In this way, we can extract the activity time from the posts.

2) Location topic model

We use MLP neural network to build a classification model. It is a class of feedforward artificial neural network and consists of at least three layers of nodes: the input layer, the hidden layer of the fully connected layer, and the output layer. Fig. 1 shows the structure of the MLP. Every unit weights an input, a bias, and an activation function. The calculation process of MLP is that the data will go through the input layer. After the unit operation of the first layer, the result is the input of the second layer, and the operation is performed and the output of the second layer is the input of the third layer, and so on. The training process is to continuously adjust the values of weights and biases to achieve satisfactory results and meet classification requirements.

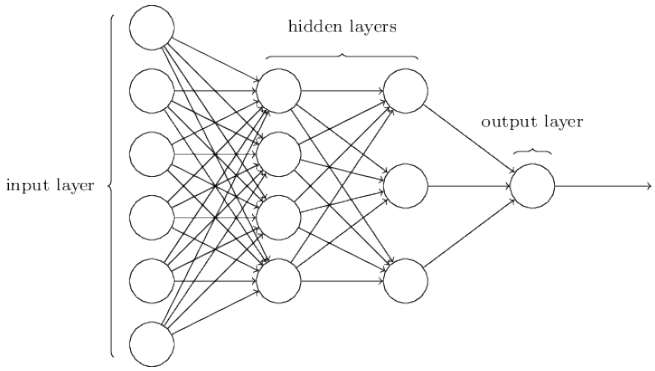

Fig. 1. Neural network structure of MLP

The MLP model in this research totally includes 5 layers including an input layer, an output layer, and 3 fully connected hidden layers in the middle. More layers mean more adjustment for the model, but it does not mean better classification and prediction results. The first layer is an input layer with a size of 100, and the second, third, and fourth hidden layers are fully connected layers with 64, 16, and 1 unit respectively. The last output layer has only one unit, which is used to indicate whether the input data is a location. For the activation function of each unit, we use the relu activation function for the first two hidden layers. And the sigmoid function for the third hidden layer, because the final goal is still a binary classification problem. The loss function we use is binary cross-entropy. It is often used in the case of binary classification. The output of sigmoid is sent to a binary cross-entropy loss function to continuously optimize the value of bias and achieve better classification results.

IV. EXPERIMENT

A. Data Collection and Processing

The data used in this research is a corpus of 18834 Hong Kong locations with full names mainly from Wikipedia [15], and all posts on the Lihkg Current Affairs section from 2019/08/01 to 2019/10/10. Totally there are 2153 titles and 302887 replies. All the posts are traditional Chinese and Cantonese. As Table I shows, we extracted the title, the creation time of the theme post, the reply, and the reply time of the post.

TABLE I. LIHKG POST COLLECTION

Title Creation Time	Title	Post Creation Time	Reply
8/23/2019 21:15	陳帆呼籲市民：寧願參加其他活動都好過堵塞機場(Chen Fan urges citizens: it would be better to participate in other activities than to block the airport)	8/23/2019 21:26	機場係正確目標(Airport is the correct target)
	陳帆呼籲市民：寧願參加其他活動都好過堵塞機場(Chen Fan urges citizens: it would be better to participate in other activities than to block the airport)	8/23/2019 21:33	民陣：831維園見 更正(Procession: See you on 831 at Victoria Park)

We train the word vector based on the content of the post. After character embedding and word segmentation of the posts mentioned in section III, we randomly extracted 7,930 high-frequency, common used, and nonlocation words and phrases. Then use the character vectors obtained previously to calculate the word vectors for Hong Kong locations and the nonlocation phrases. Examples of labeled-location and labeled-nonlocation data are shown in Table II.

TABLE II. EXAMPLES OF LOCATION AND NONLOCATION

Location	Nonlocation
中西區(Central and Western District)	依賴(rely)
灣仔(Wan Chai)	保衛(pretect)
南區(Southern District)	申請(apply)
東區(Eastern District)	阻擋(prevent)
觀塘(Kwun Tong)	團體(group)

There are locations in the location corpus and also mentioned in the post, such as "觀塘(Kwun Tong)", "元朗(Yuen Long)", and so on. But some locations are in the post but not in the location corpus, for example, overseas countries, cities in mainland China, public facility names such as "West Rail Station", and some shorthand locations. For example, there is only "Hong Kong International Airport" in the location corpus, but only "Airport" or "Hong Kong Airport" is mentioned in the post. Therefore, the main purpose of the location topic model is to find as many and accurate locations as possible, especially the location information that is not included in the location corpus.

B. Activity Location Vector Training

The location vector matrix is made up of location corpus vectors and labeled-location vectors. The nonlocation vector matrix consists of labeled-nonlocation vectors. With the location vector matrix and nonlocation vector matrix, we import the two matrix into the neural network model for training. The model randomly allocates 80% (21411 data) as the training set and 20% (5353 data) as the testing set. The training process has 20 epochs. Fig.2 shows how accuracy and

loss change over time during training and testing. The training accuracy is close to 97%, and the loss is controlled at around 0.08. The accuracy of the testing set is stable between 95% and 96%, and the loss is about 0.12.

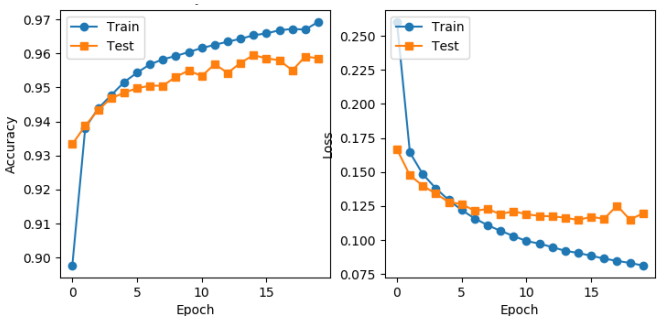

Fig. 2. Accuracy and loss of training and testing

In addition, we also calculated the recall, precision and F1 score of the prediction result of the testing set:

- Precision is the number of data correctly classified as positive divided by the total number of positive predictions. The equation is: $\frac{TP}{TP+FP}$ (4)
- Recall is the number of data correctly classified as positive divided by the total number of positive. The equation is: $\frac{TP}{TP+FN}$ (5)
- F1 score is the harmonic mean of precision and recall. The equation is: $2 \times \frac{precision \times recall}{precision+recall}$ (6)

In the equations, TP (True Positive) represents the number of data that is correctly predicted as that class. FN (False Negative) represents the number of data that is incorrectly predicted as the wrong class. FP (False Positive) represents the number of data that is incorrectly predicted as that class. Table III shows the calculation results of the precision, recall and F1 score. For the identification and prediction of the location, the precision is 98%, which means that 98% of all words predicted as locations are correct. The recall is 97%, which means that among the data in the location corpus, the model assigns 97% of the data to the location class. The results show that MLP has a good performance on binary classification, which means that this classification is also reliable for extracting and predicting location information.

TABLE III. EVALUATION RESULT

Class	Precision	Recall	F1-score	Support
Location	0.98	0.97	0.97	4007
Nonlocation	0.90	0.94	0.92	1346
Avg/Total	0.96	0.96	0.96	5353

C. Experiment Result

For the time extraction, we made a statistical chart of the date mentioned in the post and the post-release date. As shown in Fig.3, the blue line is the frequency of post-release dates, and the orange line is the frequency of dates mentioned in the post. Among the data indicated by the orange line, the lowest frequency is 11 on August 14 which is the riots in Sham Shui Po. The highest frequency is 1223 on September 1st and the airport was disrupted on that day. From the line chart, we can see that before and after the activity date, the number of posts on Lihkg will suddenly increase. And these dates in reality are quite close to the date of some big gathering events.

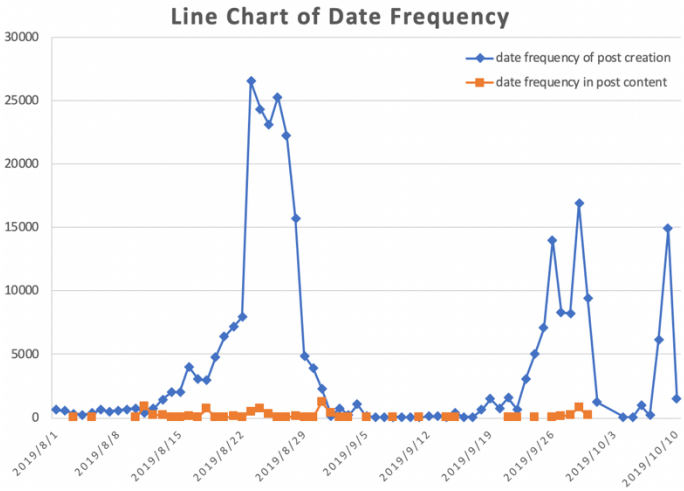

Fig. 3. Line chart of date frequency

We use the three segmentation methods in Section III to tokenize the posts and then go through the location topic model to identify the location words. The different results of the word segmentation methods are showed in Table IV. The result shows that Jieba still tokenizes a lot of useless words and phrases, and the model can filter out a large number of non-address words.

TABLE IV. PREDICTION RESULT

Segmentation Method	Predicted as location	Predicted as nonlocation	Total
Jieba	11945	96618	108563
Co-occurrence frequency	3025	27984	31009
Co-occurrence strength	1848	18646	20494

In order to better extract the locations, we output the words that are predicted as locations from all three segmentation methods, for a total of 668 words. Among them, there are locations that already exist in the location corpus, such as: "金鐘(Admiralty)", "落馬洲(Lok Ma Chau)", "九龍塘(Kowloon Tong)" and so on. There are also "new" addresses identified by the model, such as "大西北(Greater Northwest)", "大灣區(Great Bay Area)", "歐洲(Europe)", "愛沙尼亞(Estonia)", "港島區(Hong Kong Island)", etc. There are also some more precise location words, for example, some public facilities like "市中心(downtown)", "維園(V Park)" (abbreviation of Victoria Park), "賭場(casino)", "金拱門(Golden Arch)" (McDonald's nickname), "九龍東(Kowloon East)", "機場(airport)", "港鐵站(MTR Station)" and so on. In this way, we can extract a lot of locations related to the discussed activities, enrich the location corpus, and select some frequently mentioned locations as major monitoring areas.

After we add the time extraction function to the location topic model, we can automatically extract these activity feature information to test the ability to extract information. Specific examples are in Table V.

TABLE V. TOPIC MODEL RESULT

Post	Time	Location
8月5日 全港大罷工(Aug 5th General strike in Hong Kong)	['2019-08-05']	['全港'] (whole Hong Kong)
8月10日黃埔快閃都起碼至少有兩位義士被正式檢控(On August 10th, at least one of the men in the flash mob activity is charged.)	['2019-08-10']	['黃埔'] (Whampoa)
824 朝早 7 點機場 824 晏晝 1 點觀塘(August 24th , airport at 7a.m. August 24th, Kwun Tong at 1 p.m.)	['2019-08-24', '2019-08-24']	['機場','觀塘'] (Airport, Kwun Tong)
831太子恐襲 五大訴求 缺一不可(August 31, terrorist attack in Prince Edward. The five demands are indispensable)	['2019-08-31']	['太子'] (Prince Edward)

We also compare the extracted demonstrations with the timeline of real events. The extracted demonstrations are the events repeatedly mentioned in Lihkg dataset with their frequency in the brackets. The collection of real events is mainly from a citizen's website [20]. Table VI shows some comparison examples.

TABLE VI. TIMELINE COMPARISON RESULT

Extracted Events	Real Events
['2019-07-21', 元朗 Yuen Long] (223)	The 2019 Yuen Long attack was a mob attack that occurred on 21 to 22 July 2019, in Yuen Long, Hong Kong.
['2019-08-24',觀塘 Kwun Tong](124)	Aug 24 Kowloon East Kwun Tong Parade
['2019-08-31',太子 Prince Edward](93)	Aug 31 Prince Edward station attack
['2019-09-01', 機場 Airport](822)	Sept 1 Citizens launch "Airport Traffic Stress Test"
['2019-09-21', 屯門 Tuen Mun](30)	Sept 21 Recovering Tuen Mun
['2019-10-01', 荃灣 Tsuen Wan](38)	Hong Kong Demonstrations on the National Day of China
['2019-10-01', 黃大仙 Wong Tai Sin](46)	Hong Kong Demonstrations on the National Day of China

The result shows the location topic model can not only extract the location in the location corpus, but also the related newly appeared location. It also shows that the location topic vectorization model has high accuracy and versatility. For example, consider the extracted event in Yuen Long on July 21. The date is outside the time frame of our data extraction which is August to October, but we still extract the event from the post. The experiment proves that the ability to use regular expression and location topic model to respectively extract time and place is quite powerful, and the effect is quite good.

V. CONCLUSION

Unregistered public gatherings can easily threaten public safety. As a result, it is important for law enforcement to identify activity time and location of such activities to assign resources and have proper monitoring of the activity schedule. Providing only a location list, we use results from linguistics and deep learning theory, use regular expressions to automatically extract gathering time, and use a location topic model to automatically extract activity location. Our approach is not only suitable for forum data, but also can be used for other big data analysis, especially for data in Chinese. At present, it is limited to the automatic extraction of the date (when) and the location (where). Future work might investigate to extract the specific form features (how) and reasons (why) of the discussed topic in order to be helpful for law enforcement.

REFERENCES

[1] https://lihkg.com/category/1. (Access date: 02/09/2019)

[2] Chuanhai Dong, Huijia Wu, Jiajun Zhang, and Chengqing Zong. "Multichannel Lstm-crf for Named Entity Recognition in Chinese Social Media." Lecture Notes in Computer Science (including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics) 10565 (2017): 197-208.

[3] Kevin L. Priddy, Paul E. Keller, and Society of Photo-optical Instrumentation Engineers, Artificial Neural Networks, New Delhi: Prentice-Hall of India, 2007.

[4] Jianfeng Gao, Mu Li, Chang-Ning Huang, and Andi Wu. "Chinese Word Segmentation and Named Entity Recognition: A Pragmatic Approach." Computational Linguistics 31.4 (2005): 531-74.

[5] Zhiyong Qian, Jianzhong Zhou, Guoping Tong, Xinning Su. "Research on Automatic Word Segmentation and Pos Tagging for Chu Ci Based on HMM". 图书情报工作 58.04 (2014): 105-10.

[6] Sergey Andropov, Alexei Guirik, Budko Mikhail, and Budko Marina. "Network Anomaly Detection Using Artificial Neural Networks." 20th Conference of Open Innovations Association (FRUCT) 776.20 (2017): 26-31.

[7] Hissah Saif, and Hmood Al-Dossari. "Detecting and Classifying Crimes from Arabic Twitter Posts Using Text Mining Techniques." International Journal of Advanced Computer Science and Applications 9.10 (2018): 377-387.

[8] Qingye Tang, 语篇语言学 Discourse Linguistics, Shanghai University Press, 2009.

[9] Christos H.Papadimitriou, Prabhakar Raghavan, Hisao Tamaki, and Santosh Vempala. "Latent semantic indexing: A probabilistic analysis". Journal of Computer and System Sciences, 61(2), 2000, pp.217-235.

[10] Chong Wang, David Blei, and David Heckerman. "Continuous Time Dynamic Topic Models", UAI 2008 (2008): 520-529.

[11] Zhenfei Wang, Kaili Liu, Zhiyun Zheng, Fei Wang. "Research on Evolution Model of Microblog Topic Based on Time Sequence" 计算机科学 44.08 (2017): 270-73.

[12] Marianela García Lozano, Jonah Schreiber, and Joel Brynielsson. "Tracking Geographical Locations Using a Geo-Aware Topic Model for Analyzing Social Media Data." Decision Support Systems 99.SI (2017): 18-29.

[13] Tomas Mikolov, Wen-tau Yih, and Geoffrey Zweig. "Linguistic regularities in continuous space word representations". In Proceedings of NAACL-HLT, 2013, pages 746–751.

[14] Řehůřek R, Sojka P. "Gensim—statistical semantics in Python". Retrieved from genism.org. 2011. (Access date: 2019/12/6)

[15] https://en.wikipedia.org/wiki/Hong_Kong (Access date: 10/22/2019)

[16] Javier Pastor-Galindo, Pantaleone Nespoli, Félix Gómez Mármol, Gregorio Martínez Pérez: "The Not Yet Exploited Goldmine of OSINT: Opportunities, Open Challenges and Future Trends". IEEE Access 8: 10282-10304 (2020).

[17] Holmes Chan,'A catastrophe': Hong Kong police say 159 arrested during weekend chaos, 16 charged with rioting, Hong Kong Free Press, https://www.hongkongfp.com/2019/09/02/catastrophe-hong-kong-police-say-159-arrested-weekend-chaos-16-charged-rioting/. (Access date: 02/09/2019)

[18] AFP, Hong Kong police say man was set alight after arguing with protesters, Hong Kong Free Press, https://www.hongkongfp.com/2019/11/11/hong-kong-police-say-man-set-alight-arguing-protesters/. (Access date: 11/11/2019)

[19] Mohammad Reza Keyvanpour, Mostafa Javideh; Mohammadreza Ebrahimi. "Detecting and investigating crime by means of data mining: A general crime matching framework.". Procedia Computer Science, 2011, Vol.3, pp.872-880.

[20] 陳朗熹, Summary of key events in the Hong Kong protests, https://hkontheroad.org. (Access date: 20/03/2020)