Predicting the locations of unrest using social media

Shengzhi Qin, Qiaokun Wen, Kam-Pui Chow

Chapter 9

PREDICTING THE LOCATIONS OF UNREST USING SOCIAL MEDIA

Shengzhi Qin, Qiaokun Wen and Kam-Pui Chow

Abstract The public often relies on social media to discuss and organize activities such as rallies and demonstrations. Monitoring and analyzing open-source social media platforms can provide insights into the locations and scales of rallies and demonstrations, and help ensure that they are peaceful and orderly.

This chapter describes a dictionary-based, semi-supervised learning methodology for obtaining location information from Chinese web forums. The methodology trains a named entity recognition model using a small amount of labeled data and employs n-grams and association rule mining to validate the results. The validated data becomes the new training dataset; this step is performed iteratively to train the named entity recognition model. Experimental results demonstrate that the iteratively-trained model has much better performance than other models described in the research literature.

Keywords: Social media analysis, location extraction, named entity recognition

1. Introduction

Since 2019, large-scale protests by the Anti-Extradition Law Amendment Bill Movement (Anti-ELAB Movement) have occurred in Hong Kong [13]. Predicting the locations and scales of such protests can assist law enforcement in planning and mobilizing resources to ensure that the protests are peaceful and orderly.

Social media is often used to organize public events [16]. Online discussion sites, including web forums, have played key roles in organizing flash mobs, protest campaigns and demonstrations during periods of unrest [1]. The monitoring and analysis of public opinion in online discussions is an effective means for obtaining information that could assist public safety efforts.

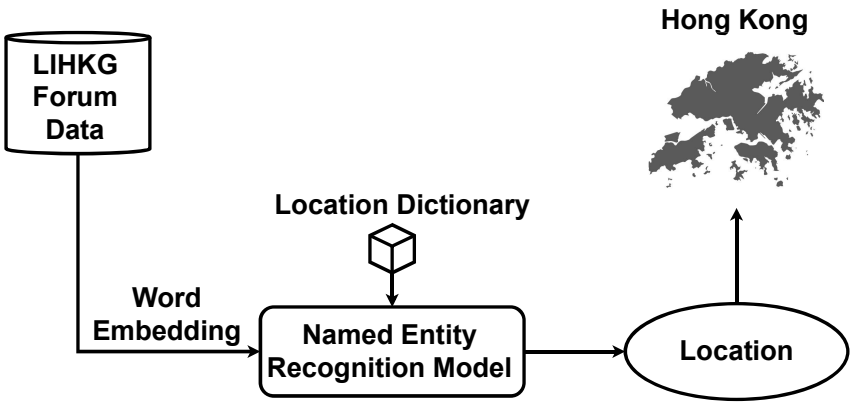

Figure 1. Location extraction methodology.

When discussing political activities in web forums, organizers often mention the locations of future rallies, such as the "818 Victoria Park Rally" [4]. Web forum users also share location information in their discussions. Therefore, it would be effective to focus on locations in web forum discussions to gain advance information about possible rallies.

This research focuses on the LIHKG public web forum [10], one of the most active platforms for discussing the Anti-ELAB Movement [8]. A dictionary-based, semi-supervised learning methodology was developed to automatically extract location information from harvested web forum data.

Figure 1 presents an overview of the location extraction methodology. The first step is to crawl a web forum to gather data. Next, the topic and post contents are processed by a named entity recognition model to identify location data. Note that the location dictionary is used in conjunction with the named entity recognition model to improve performance. Finally, the extracted locations are analyzed with other information to predict the locations where unrest may occur.

Empirical experiments have revealed that accurate location identification is the principal challenge. Another challenge is that the focus is on Chinese web forums. Unlike in English text, locations are difficult to extract because Chinese has neither word segmentation nor capitalization rules. Additionally, web forum data differs from data in structured sources such as newspapers and papers. Specifically, the language is irregular, often does not follow grammar rules and incorporates large amounts of slang and abbreviations. The dictionary-based, semi-supervised learning method developed in this research enhances the extraction of location information from Chinese web forums.

2. Related Work

Analysis of public opinion can provide law enforcement with alerts before incidents occur as well as important information about the incidents. Tang and Song [14] proposed a visual analysis methodology for high-frequency words and user responses on the Weibo social media platform in Mainland China. Yang and Ng [15] extracted and clustered key information from social networks to classify public opinion and enhance subsequent analysis. People exchange information about the times and locations of events using social media. Because time has a standard format, it can be extracted using simple methods such as regular expressions. However, location extraction is more difficult because it does not have a standard structure.

Machine learning and deep learning methods are used to identify entities, including locations, in sentences [9]. The tagging methodologies employ hidden Markov models [3], maximum entropy Markov models [12] and conditional random fields [7]. Hammerton [5] used a long short-term memory (LSTM) based neural network model for entity recognition. Collobert et al. [2] combined deep learning with data mining methods to achieve high accuracy. Huang et al. [6] proposed one of the best-performing English entity recognition models that uses a bidirectional long short-term memory and conditional random fields; the model is robust and does not have any special dependence on word embedding.

However, considerable differences exist between English and Chinese, especially with regard to sentence structure. In English, words are separated by spaces and proper nouns such as locations and names are capitalized; Chinese does not have such features. Another problem is that named entity recognition models focus on processing structured text such as reports and news, and are not applicable to short messages in web forums. Additionally, many expressions in web forums do not follow grammar rules. Furthermore, many posts contain non-standard expressions, including abbreviations and slang terms. These characteristics render manual labeling of training data and model training much more difficult.

Zhang and Yang [17] proposed the Lattice-LSTM model, which adds word segmentation results prior to word information and fuses word information into character information; experiments conducted with the Weibo platform yielded a 62.56% match ratio. Liu [11] designed an encoding strategy based on word-character long short-term memory that improves on the Lattice-LSTM model, but it cannot be batched due to its uneven structure. Zhu et al. [18] proposed a multi-task long short-term memory method for Chinese named entity recognition. It uses

Table 1. LIHKG web forum topic and post.

Topic: 2019-08-19 13:00:00,94, 神仙,56,反對修訂引渡條例大遊行集中討論
Post: 2019-08-18 15:02:00,108,241242, 明天旺角集合

Field	Value	English Translation
Topic Created Time	2019-08-19 13:00:00	
Topic Author ID	94	
Topic Author Name	神仙	Immortal
Topic ID	56	
Topic Title	反對修訂引渡條例大遊行	Anti-ELAB Protest
Post Created Time	2019-08-18 15:02:00	
Post Author ID	108	
Post Author Name	241242	
Post Content	明天旺角集合	Tomorrow Mong Kok meet up

three long short-term memory structures to model related information – one is based on character information for named entity recognition, the second is based on word information for word segmentation tasks and the third is based on a shared structure that models character and word information. The method achieved 59.31% accuracy on Weibo.

All the methods presented above require the manually labeling of large amounts of data. It is difficult to manually label forum data because web forum text has many spoken expressions and does not follow grammar rules, resulting in different people labeling entities differently. These differences require an additional verification step. In contrast, the proposed training method using dictionary-based, semi-supervised learning requires only a small amount of data to be labeled manually and uses prior location information to extract locations in web forum posts.

3. Location Extraction from Web Forum Data

This section describes the dictionary-based, semi-supervised learning methodology for extracting locations from web forum data.

3.1 Web Forum Dataset

This work has focused on LIHKG, a popular web forum website that is often referred to as the Hong Kong version of Reddit. LIHK was one of the most active platforms for discussing the Anti-ELAB Movement. Chinese language data was collected from the Current Affairs channel in LIHKG, the principal channel for discussing political events and protest movements. Table 1 shows a sample LIHKG web forum post.

Table 2. Locations in dictionary and web forum posts.

Dictionary Content	Web Forum Content
旺角 (Mong Kok)	旺角港鐵站 (Mong Kok MTR Station)
維多利亞公園 (Victoria Park)	維園 (Vi-Park)
皇后大道中 (Queen's Road Central)	皇后大道中168號 (Queen's Road Central #168)

The dataset used in this research comprised more than 300,000 LIHKG forum data items collected from 302,109 posts between August 18, 2019 to October 10, 2019. The posts, which had 42,258 distinct authors, contained 2,126 topics.

3.2 Dictionary-Based Semi-Supervised Learning

In real-world applications, dictionary-based keyword matching is often used to identify locations because it is simple, convenient and fast, and has low false positive rates. However, it cannot recognize locations outside the dictionary especially when the locations mentioned in web forums are not specified completely. Table 2 shows examples of locations used in web forum posts that do not completely match locations in the dictionary.

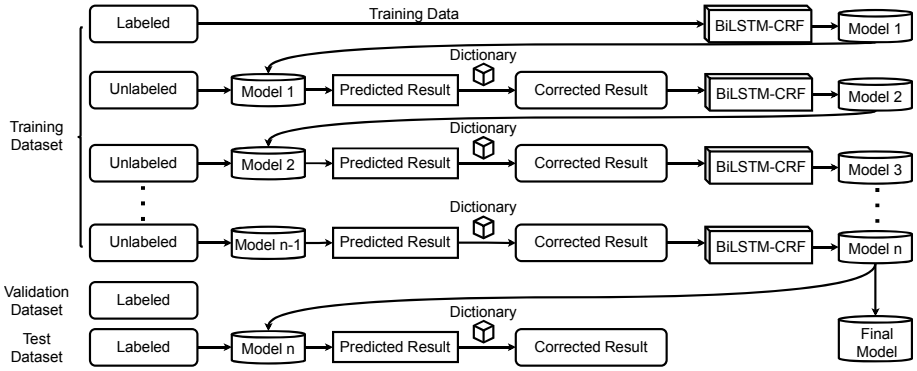

Figure 2. Dictionary-based, semi-supervised learning method.

Figure 2 shows the proposed dictionary-based, semi-supervised learning method, which combines named entity recognition and dictionary matching. The dataset was divided into three parts: a training dataset, a validation set and a testing dataset.

The training dataset was divided into several parts, named part_1, part_2, part_3, etc. Each part contained 50,000 posts. Only data in part_1 was labeled (i.e., only 50,000 posts were labeled). The validation set contained 2,000 posts for intermediate corrections. The testing dataset contained another 2,000 posts for final testing.

The labeled training data (part_1) was input to a bidirectional long short-term memory and conditional random field (BiLSTM-CRF) model for training. Model training was performed iteratively. Note that model_1 denotes the first trained model and model_i denotes the model trained after the i^{th} iteration. Each iteration involved three steps:

- **Step 1:** Apply $model_i$ to predict the 50,000 posts in part_{i+1}. The dictionary is applied in conjunction with the n-Gram-ARM algorithm (described below) to correct the training results.
- **Step 2:** Label the new corrected results using BIO labeling and re-input them as training data to model_i to obtain model_{i+1}.
- **Step 3:** Compare all the identified locations with the current dictionary. If a location is not in the dictionary, then it is added to the dictionary as a new location.

The iterations were repeated until all the parts of the training dataset were utilized. By continuously updating the model and dictionary, the dictionary was expanded and the named entity recognition model was improved iteratively.

3.3 BiLSTM-CRF Model

Dictionary-based, semi-supervised learning utilized the BiLSTM-CRF model for named entity recognition. In the data labeling process, words in a sentence were divided into two types – valid entities and invalid characters. Valid entities included time (TIM), location (LOC) and person (PER). The first character of a valid entity was labeled "B-XXX" and the remaining characters were labeled "I-XXX" where "XXX" is TIM, LOC or PER. Strings that were not valid entities were labeled "O." Table 3 shows examples of labeled text.

Upon being given a Chinese post, the BiLSTM-CRF model automatically identified the beginning and end of a location. Figure 3 shows the BiLSTM-CRF model structure. In addition to identifying location information in the post, the time and person entities were labeled in order to learn the relationships between entities and tags.

Word embedding was performed to convert text into a vector for input to the BiLSTM-CRF model. LSTM (long short-term memory) is

Table 3. Labeled text example.

English Translation	Tomorrow		Mong	Kok	meet	up
Chinese Post	明	天	旺	角	集	合
Labels	B-TIM	I-TIM	B-LOC	I-LOC	O	O

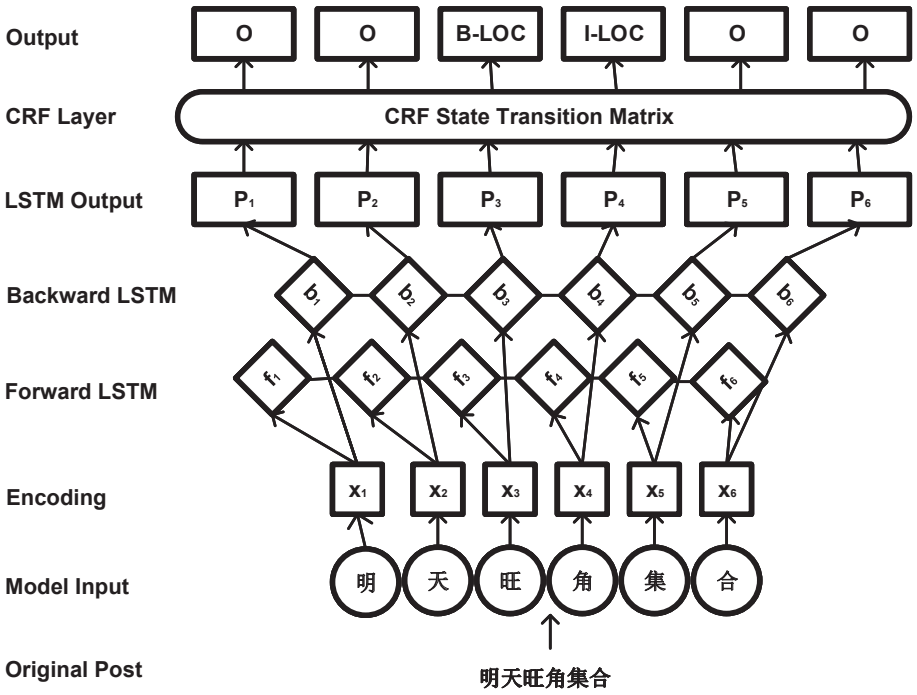

Figure 3. BiLSTM-CRF model structure.

a deep learning model that remembers information from the previous step. However, the LSTM model can only predict the output of the next layer based on information from the previous layer. In some problems, the output of the current layer is related to the previous state and also depends on the future state. For example, a missing word in a current sentence is predicted based on the previous paragraph as well the content that comes after the current sentence. To accomplish this, the BiLSTM (bidirectional long short-term memory) model incorporates an additional layer that uses the end of a sentence as the starting input of the sequence. The BiLSTM model processes the results of the two layers to improve named entity recognition.

Table 4. Possible situations leading to location recognition results.

Situation 1		
Locations labeled by dictionary only		
	Sample Data	**English Translation**
	明天旺角集合	Tomorrow **Mong Kok** meet up
Situation 2		
Locations labeled by dictionary and named entity recognition model		
Cases	**Sample Data**	**English Translation**
2a	明天[旺角地鐵站]集合	Tomorrow **[Mong Kok Railway Station]** meet up
2b	明天旺[角地鐵站]集合	Tomorrow **Mong [Kok Railway Station]** meet up
2c	明天黃[大仙]集合	Tomorrow **Wong [Tai Sin]** meet up
Situation 3		
Locations labeled by named entity recognition model only		
	Sample Data	**English Translation**
	明天[機場]集合	Tomorrow **[Airport]** meet up
Situation 4		
Locations not labeled by dictionary and named entity recognition model		

In the BiLSTM-CRF model, a conditional random field (CRF) layer is employed before the output instead of the traditional softmax method. This layer enables the relationships between labels to be learned. Some connections between entities with different attributes, such as time and place appearing consecutively, may be discovered. This is why entities other than locations are also labeled.

3.4 n-Gram-ARM Algorithm

As mentioned above, the proposed methodology utilizes the location dictionary to correct the model prediction results during the iterative training process. The results obtained after the iterative training were found to have many false positives and false negatives. Also, there were deviations between the dictionary labeling results and the named entity recognition results.

Table 4 shows the possible situations leading to location recognition results using the dictionary and named entity recognition model. Dictionary labels are underlined. Named entity recognition model labels are enclosed in square brackets. Correct labels are presented in boldface. Case 2a covers situations where the named entity recognition model labeling results are supersets of the dictionary labeling results. Case 2b covers situations where the named entity recognition model and dic-

Table 5. Solutions for specific situations.

Situation	**Solution**
1	Keep only dictionary-labeled parts
2a	n-Gram-ARM
2b	n-Gram-ARM
2c	Keep only dictionary-labeled parts
3	Simple n-Gram-ARM
4	Iterative training

tionary labeling results have non-empty intersections. Case 2c covers situations where the named entity recognition model labeling results are subsets of the dictionary labeling results.

Table 5 shows the solutions employed to reduce the false negative rates for the various situations. As discussed above, Situation 4 is handled by the iterative training process.

The locations labeled using the dictionary can be assumed to be correct, but the dictionary may not contain all the known locations and the new locations. This problem was addressed by applying the n-Gram-ARM algorithm, which has an association rule mining (ARM) part and an n-Gram part.

The association rule mining part of the algorithm used two indicators, Support and Confidence, to measure the goodness of the labels. Support is the frequency of the union of the words in the web forum corpus labeled in the dictionary and by the named entity recognition model. Confidence is the probability of the union of dictionary-labeled and named-entity-recognition-model-labeled words if the dictionary-labeled words exist. Specifically:

$$\text{Support} = freq(Dic \cup NER)$$

$$\text{Confidence} = \frac{freq(Dic \cup NER)}{freq(Dic)}$$

where Dic denotes the words labeled as locations by the dictionary, NER denotes the words predicted as locations by the named entity recognition model and $freq$ is the number of occurrences.

Support measures how often a word appears whereas Confidence measures how likely the two parts of the label appear consecutively. The greater the values of two indicators, the greater the probability that the word is a location. For example, if the dictionary extraction result is

Table 6. Association rule mining false positive example.

Sample Data	English Translation
明天[# 旺角地鐵站]集合	Tomorrow [# **Mong Kok Railway Station**] meet up

"Mong Kok" and the named entity recognition model extraction result is "Kok Railway Station" (Situation 2b in Table 4), then the union of the two extracted results is "Mong Kok Railway Station." The two indicators are computed as:

$$\text{Support} = \text{freq}(\text{旺角地鐵站})$$
$$\text{Confidence} = \text{freq}(\text{旺角地鐵站})/\text{freq}(\text{旺角})$$

If the Support and Confidence values are both greater than the association rule mining threshold, then the word is considered to be a location and the named entity recognition model label is retained. After training for the first time and using association rule mining, words that are not in the dictionary may be obtained. These words correspond to new locations identified by association rule mining.

The n-Gram method reduces the number of false positives. However, other false positives are possible. Specifically, words labeled by the two methods are incorrect and the labeled part of the named entity recognition model is a superset of the correct answers. Table 6 shows an example of a false positive obtained by association rule mining.

When the association rule mining validation result of the named entity recognition model output is lower than the association rule mining threshold, n-Gram-ARM is used to recheck the result. Specifically, association rule mining performed on sub-words of the named entity recognition model results in different lengths. The range of lengths n in the n-Gram method is:

$$len(Dic) < n < len(NER)$$

where $len(Dic)$ is the length of a word labeled by the dictionary and $len(NER)$ is the length of a word labeled by the named entity recognition model.

In the example above, association rule mining was performed after applying the n-Gram method. Table 7 shows the n-Gram results.

The phrases to be retained as locations were based on the association rule mining results. A higher n-Gram threshold was used to further filter the words. Since there is no dictionary labeling in Situation 3, the

Table 7. n-Gram results.

n-Gram Phrases for ARM with Sample Data: # 旺角地鐵站		
n = 5	n = 4	n = 3
# 旺角地鐵/ 旺角地鐵站	# 旺角地/ 旺角地鐵/ 角地鐵站	# 旺角/ 旺角地/ 角地鐵/ 地鐵站

confidence cannot be computed. Therefore, the solution for Situation 3 is named Simple n-Gram-ARM.

4. Experiments and Results

Comparative experiments were conducted with several commonly-used Chinese named entity recognition models to determine the model with the best performance.

Table 8. Chinese named entity recognition model results.

Model	**Percentage**	**Recall**	**F1-Score**	**F1-Score (Location Only)**
HMM	48.69%	46.03%	47.32%	19.40%
CRF	56.68%	58.08%	57.37%	22.57%
LSTM	48.56%	46.30%	47.41%	16.34%
CAN-NER	63.89%	54.79%	59.00%	25.81%
Lattice-LSTM	59.61%	59.45%	59.53%	24.69%
BiLSTM-CRF	62.87%	58.90%	60.82%	25.28%

Table 8 shows the experimental results. The BiLSTM-CRF model was the best performer on the dataset. The experiments revealed that locations were the main reason for the low accuracy. Table 8 shows that the best F1-Score with only the locations labeled is just 25.28%.

As mentioned above, the n-Gram-ARM algorithm was applied to improve the location extraction performance. The ARM thresholds were set to 7 for Support and 0.01 for Confidence based on experience.

Table 9 shows sample outputs obtained after validation using associated rule mining. None of the locations appeared in the dictionary. The locations in the last row of the table were filtered using the ARM threshold due to the low Support and Confidence values. The results in Table 9 also show that associated rule mining can expand the locations in the dictionary by identifying new location vocabulary with the

Table 9. Association rule mining validation outputs.

New Location	English Translation	Support	Confidence
金鐘站	Admiralty Station	63	0.078553616
葵涌警署	Kwai Chung Police Station	32	0.198757764
元朗西鐵站	Yuen Long West Rail Station	55	0.037826685
深圳羅湖	Shenzhen Luohu	35	0.224358974
維園	Vi-Park	560	—
元朗站# 罵村民	Yuen Long Station#Curse villagers	2	0.001375516

base locations. This also complies with the rules for naming and using locations.

Table 10. n-Gram-ARM algorithm result.

N-Gram Phrases	Support	Confidence
元朗站# 罵村民	2	0.0013755
元朗站# 罵村	2	0.0013755
朗站# 罵村民	2	0.0013755
元朗站# 罵	2	0.0013755
朗站# 罵村	2	0.0013755
站# 罵村民	2	0.0013755
元朗站	19	0.0130674
朗站# 罵	2	0.0013755
站# 罵村	2	0.0013755
# 罵村民	2	0.0013755
元朗站	**148**	**0.1017881**
朗站	19	0.0130674
站# 罵	2	0.0013755
# 罵村	2	0.0013755
罵村民	2	0.0013755

Table 10 shows the n-Gram-ARM algorithm result (row in bold font). The n-Gram thresholds were set to 20 for Support and 0.03 for Confidence based on experience. The application of the n-Gram method enables the recognized entities to be further decomposed and analyzed.

Table 11 shows that applying the n-Gram-ARM algorithm effectively reduced the number of false negatives. Additionally, the accuracy of location recognition is increased along with the overall experimental accuracy.

Table 11. Evaluation of dictionary n-Gram-ARM verification.

Model	**Precision**	**Recall**	**F1-Score**	**F1-Score Overall**
BiLSTM-CRF	17.24%	51.40%	25.28%	60.82%
BiLSTM-CRF + n-Gram-ARM	48.75%	36.44%	41.71%	78.08%

Table 12. Iterative training method evaluation.

Iteration	**Precision**	**Recall**	**F1-Score**	**F1-Score Overall**
0	17.24%	51.40%	25.28%	60.82%
1	54.39%	28.97%	37.80%	68.09%
2	60.00%	36.45%	45.35%	72.77%
3	58.28%	37.29%	45.48%	72.32%
4	56.24%	36.37%	44.17%	70.13%
5	54.84%	37.28%	44.39%	71.02%

Table 12 shows the effectiveness of the iterative training method. The method addresses problems posed by insufficient labeled data, low quality labeled data and irregular data. The results show that during the first three iterations, as the amount of data in the training dataset increased, the prediction results became increasingly accurate, until a steady state was reached. At the same time, the dictionary was also updated with more than 100 new locations. Not only is the proposed BiLSTM-CRF model more robust and accurate than the other named entity recognition models, but it also yields a comprehensive and richer dictionary of locations.

5. Conclusions

The location extraction methodology described in this chapter employs dictionary-based, semi-supervised learning to obtain location information from web forum data (specifically, from LIHKG, one of the most active platforms for discussing the Anti-ELAB Movement in Hong Kong). The extracted location information can assist law enforcement in planning and mobilizing resources to ensure that protests are peaceful and orderly. Experimental results demonstrate that the iteratively-trained model has much better performance than other models described in the research literature.

Future research will attempt to extract other information such as time and user behavior that would help identify relationships between posts and political movements. Research will also analyze web forum user behavior to discover user relationships and identify user roles in forums and events, such as organizers and active users.

References

[1] A. Breuer, The Role of Social Media in Mobilizing Political Protest: Evidence from the Tunisian Revolution, Discussion Paper No. 10/2012, German Development Institute, Bonn, Germany, 2012.

[2] R. Collobert, J. Weston, L. Bottou, M. Karlen, K. Kavukcuoglu and P. Kuksa, Natural language processing (almost) from scratch, *Journal of Machine Learning Research*, vol. 12(2011), pp. 2493–2537, 2011.

[3] S. Eddy, Hidden Markov models, *Current Opinion in Structural Biology*, vol. 6(3), pp. 361–365, 1996.

[4] Government of Hong Kong, Government Response to Public Meeting in Victoria Park, Hong Kong, China (`www.info.gov.hk/gia/general/201908/18/P2019081800818.htm`), August 18, 2019.

[5] J. Hammerton, Named entity recognition with long short-term memory, *Proceedings of the Seventh Conference on Natural Language Learning*, pp. 172–175, 2003.

[6] Z. Huang, X. Wei and K. Yu, Bidirectional LSTM-CRF Models for Sequence Tagging, arXiv: 1508.01991 (`arxiv.org/abs/1508.01991`), 2015.

[7] J. Lafferty, A. McCallum and F. Pereira, Conditional random fields: Probabilistic models for segmenting and labeling sequence data, *Proceedings of the Eighteenth International Conference on Machine Learning*, pp. 282–289, 2001.

[8] F. Lee, H. Liang, E. Cheng, G. Tang and S. Yuen, Affordances, movement dynamics and a centralized digital communication platform in a networked movement, to appear in *Information, Communication and Society*, 2021.

[9] J. Li, A. Sun, J. Han and C. Li, A survey of deep learning for named entity recognition, to appear in *IEEE Transactions on Knowledge and Data Engineering*, 2021.

[10] LIHKG, LIHKG Online Forum, Hong Kong, China (`lihkg.com`), 2021.

[11] W. Liu, T. Xu, Q. Xu, J. Song and Y. Zu, An encoding strategy based word-character LSTM for Chinese NER, *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pp. 2379–2389, 2019.

[12] A. McCallum, D. Freitag and F. Pereira, Maximum entropy Markov models for information extraction and segmentation, *Proceedings of the Seventeenth International Conference on Machine Learning*, pp. 591–598, 2000.

[13] M. Purbrick, A report on the 2019 Hong Kong protests, *Asian Affairs*, vol. 50(4), pp. 465–487, 2019.

[14] X. Tang and C. Song, Microblog public opinion analysis based on complex networks, *Journal of the China Society for Scientific and Technical Information*, vol. 31(11), pp. 1153–1163, 2012.

[15] C. Yang and T. Ng, Analyzing and visualizing web opinion development and social interactions with density-based clustering, *IEEE Transactions on Systems Man and Cybernetics – Part A: Systems and Humans*, vol. 41(6), pp. 1144–1155, 2011.

[16] T. Zeitzoff, How social media is changing conflict, *Journal of Conflict Resolution*, vol. 61(9), pp. 1970–1991, 2017.

[17] Y. Zhang and J. Yang, Chinese NER Using Lattice LSTM, arXiv: 1805.02023 (`arxiv.org/abs/1805.02023`), 2018.

[18] Y. Zhu, G. Wang and B. Karlsson, CAN-NER: Convolutional Attention Network for Chinese Named Entity Recognition, arXiv: 1904.02141 (`arxiv.org/abs/1904.02141`), 2020.