Public Opinion Monitoring for Proactive Crime Detection Using Named Entity Recognition

Wencan Wu, Kam-Pui Chow, Yonghao Mai, Jun Zhang

Chapter 11

PUBLIC OPINION MONITORING FOR PROACTIVE CRIME DETECTION USING NAMED ENTITY RECOGNITION

Wencan Wu, Kam-Pui Chow, Yonghao Mai and Jun Zhang

Abstract Public opinion monitoring has been well studied in sociology and informatics. Considerable amounts of crime-related information are available on social media platforms every day. Current methods for monitoring public opinion are typically based on rule matching and manual searching instead of automated processing and analysis. However, the extraction of useful information from large volumes of social media data is a major challenge in public opinion monitoring.

This chapter describes a methodology for extracting key information from a large volume of Chinese text using named entity recognition based on the LSTM-CRF model. Since traditional named entity recognition datasets are small and only contain a few types, a custom crime-related corpus was created for training. The results demonstrate that the methodology can automatically extract key attributes such as person, location, organization and crime type with a precision of 87.58%, recall of 83.22% and F1 score of 85.24%.

Keywords: Public opinion monitoring, named entity recognition, crime alerts

1. Introduction

Public opinion monitoring – or social listening – is a promising approach for alerting law enforcement about crimes before they occur, because some crimes are planned using social media [8]. Several such cases were encountered during the protests against the Hong Kong extradition bill of 2019. The demonstrations against the bill began in March and April 2019 and escalated significantly in June 2019 [1]. A significant number of criminal activities occurred during the protests, including intimidation, beatings and looting that seriously impacted public safety

and social order. Many of these activities were planned and coordinated using social media platforms and online discussion groups.

Unfortunately, discovering potential crimes is difficult because of the need to sift through large volumes of data and interpret the slang terms used by criminal entities. Current approaches for recognizing criminal activities, which employ simple rule matching or manual processing, are inefficient and error-prone.

Named entity recognition is a fundamental component of many natural language processing applications such as relation extraction, event extraction, knowledge graphs and question-answering systems. It can classify specific and useful entities into appropriate semantic classes such as persons, locations, organizations, dates and times [5].

This chapter describes a named entity recognition methodology for monitoring public opinion in Chinese language posts and extracting crime-related features. Specifically, the LSTM-CRF model, an artificial recurrent neural network [3], is employed to extract key information from a large volume of Chinese text. Since traditional named entity recognition datasets are small and contain few types, a custom crime-related corpus was created for training. Experiments reveal that the trained LSTM-CRF model was able to recognize special features that did not exist in the training dataset. The methodology automatically extracted key attributes such as person, location, organization and crime type with a precision of 87.58%, recall of 83.22% and F1 score of 85.24%.

2. Named Entity Recognition

Named entity recognition, also known as sequence labeling, is used to identify special entities in structured or unstructured text. Conventional named entity recognition methods fall in two categories, one based on rules or dictionaries and the other based on statistics [2].

Named entity recognition methods based on rules typically employ finite-state machines to match specific language models. However, the rule maker needs to have sufficient knowledge of the language to construct the finite-state machine. Methods based on dictionaries rely on previously-created dictionaries of persons, locations and organizations. Thus, the methods based on rules and dictionaries require large amounts of time and resources to prepare the supporting materials. Additionally, the methods have high error rates.

Statistics-based named entity recognition methods were developed to address the disadvantages of rule and dictionary based methods. These methods employ n-gram, hidden Markov, maximum entropy, conditional random field, support vector machine or decision tree models. All these

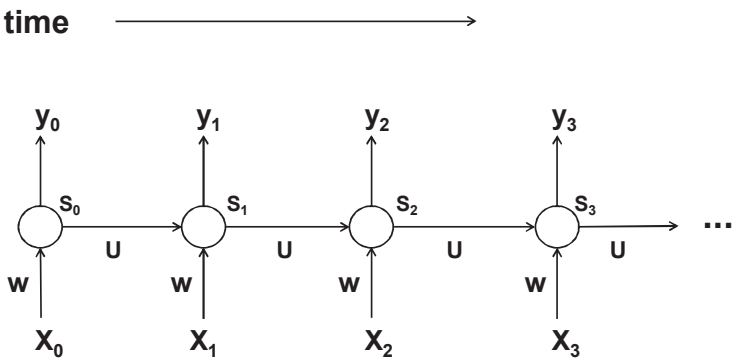

Figure 1. Recurrent neural network structure.

models require training datasets. While creating the datasets is not difficult, model performance depends significantly on the quality and quantity of the datasets [15].

The proposed methodology processes large amounts of text using deep learning and transfer learning. The first step is to create the training and testing datasets. Since no public corpora containing crime-related words exist, a custom criminal corpus was created. BIO labels were added to each word and the resulting corpus was divided into a training dataset (90% of the data) and a testing dataset (10% of the data).

3. LSTM-CRF Model

A long short-term memory and conditional random field (LSTM-CRF) model combines a long short-term memory (LSTM) model and a conditional random field (CRF) model. The LSTM model is a special type of recurrent neural network that processes long-term dependence better than conventional recurrent neural networks. The CRF model is effective at labeling and segmenting serialized data.

A traditional neural network has an input layer, a hidden layer and an output layer, where all the nodes in each layer are fully connected to nodes in the next layer. The output values of each layer, which are computed from the input values of the layer, are passed as input values to the next layer. Each input value is processed independently and the process has no memory. For input data that is sequential, such as a sentence, it is necessary to process the data in sequence, one element at a time. A recurrent neural network is a special type of neural network that is geared for processing sequential data. Specifically, it iterates through the data in sequence and maintains state information while processing. Figure 1 shows the structure of a recurrent neural network.

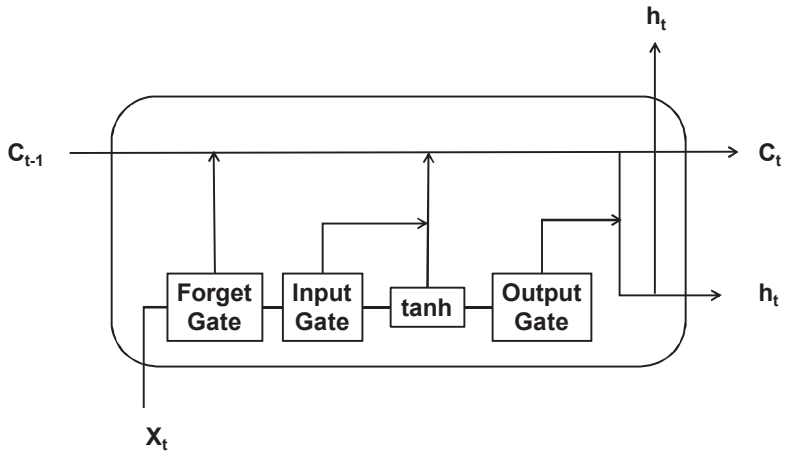

Figure 2. LSTM memory cell.

The LSTM model is a special type of recurrent neural network where the neurons are replaced by memory cells, each with input gates, forget gates and output gates. This special structure makes the LSTM model better at processing long-term dependence than a recurrent neural network; also, it avoids gradient disappearance and gradient expansion problems [13].

Figure 2 shows an LSTM memory cell. Note that C_{t-1} and X_t are the input values at time t, tanh is a neural layer, h_t is the state at time t and C_t is the output value at time t.

The CRF model is a statistical name entity recognition technique. A conditional random field defines when a random variable Y, conditioned on a set of observations X, Prob($Y \mid X$), obeys the Markov property. In this work, X is a set of words and Y is the corresponding label. The CRF model can then be used to learn the relationship between labels. For example, when a word is labeled as B-PER, the label of the next word is strongly believed to be I-PER. Compared with a conventional labeling model, the CRF model is better at using sentence-level label (tag) information and is able to model the transition behavior of different tags. Also, with CRF, the labeling of one character considers the labels of neighboring characters to determine the final label [4, 7].

The proposed LSTM-CRF model combines the LSTM and CRF models. Figure 3 shows the structural graph of the LSTM-CRF model. It comprises three layers. The first layer is the word embedding layer, which transforms each word into a corresponding vector so that the entire sentence can be represented as an embedding matrix. The second layer is the LSTM layer, which uses forward propagation and backward

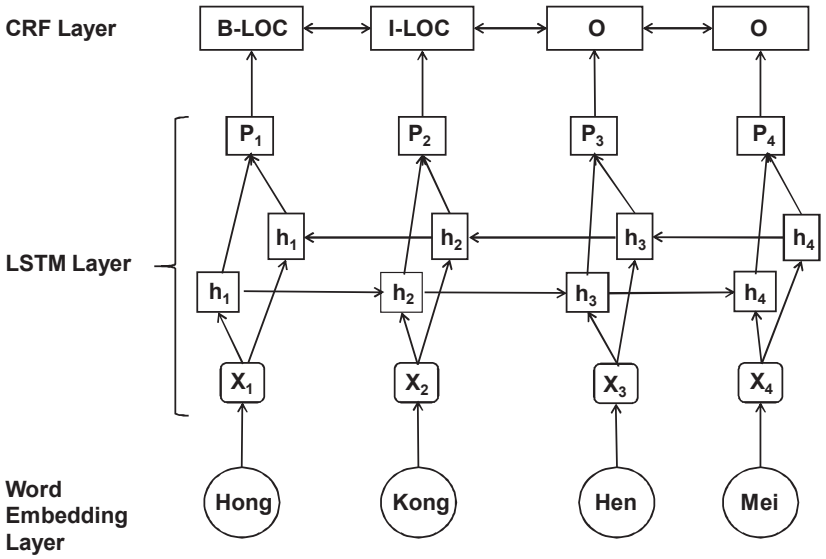

Figure 3. LSTM-CRF model.

propagation to extract features automatically. The third layer is the CRF layer, which uses the output of the second layer to label words with maximum probability.

4. Related Work

As mentioned above, a significant number of criminal activities occurred during the protests against the Hong Kong extradition bill of 2019, including intimidation, beatings and looting that impacted public safety and social order. Since many of these activities were planned and coordinated using social media platforms and online discussion groups, Hong Kong government authorities were interested in monitoring public opinion to identify potential crimes and work proactively to mitigate the hazards. However, very limited research has been done on applying deep learning techniques to detect potential criminal events by analyzing Chinese text in social media and online discussion groups.

Wang et al. [12] have employed machine learning for sentimental entity recognition with a precision of 89%. Kleinberg et al. [6] have developed an automated verbal deception detection system that employs the spaCy and Stanford NER tools. Motivated by these efforts, the research described in this chapter extracts information from large volumes of Chinese text using named entity recognition based on the LSTM-CRF model.

Table 1. Named entity recognition corpora.

Corpus	Training Data		Testing Data	
	Sentences	**Tokens**	**Sentences**	**Tokens**
Normal Part	36,657	1,595,064	4,360	177,231
Crime Part	4,726	11,670	224	1,295

5. Experiments

This section discusses the experimental setup and the classification of named entities.

5.1 Experimental Setup

The corpora and LSTM-CRF model are the two key components in the experiments. The corpora comprise a normal part and a crime part. The normal part is a portion of the MSRA corpus [9] whereas the crime part comprising three Chinese dictionaries specializing in crime was downloaded from the Sougou platform [11].

Table 1 shows the distribution of data in the named entity recognition corpora. One corpus is the normal part, which contains four types of entities, i.e., person, location, organization and not a named entity (non-named entity). The other corpus is the crime part, which only contains the criminal entity type.

Each entry (sentence) in the corpora was processed to extract a set of tokens (Chinese characters). The BIO tagging style employed for labeling uses O, B-PER, I-PER, B-LOC, I-LOC, B-ORG, I-ORG, B-CRM and I-CRM, where (i) O means that the word is not a named entity; (ii) B-X means that the word is the beginning of X (e.g., B-PER means that the word is the beginning of person); and (iii) I-X means that the word is inside word X. Figure 4 shows examples that use the BIO tagging style [10].

5.2 Classification of Named Entities

The classification of named entities involves two steps:

- **Step 1: Data Processing:** Based on the corpus, each word is labeled with a corresponding tag. Next, the corpus is serialized and a dictionary containing non-replicative words is constructed. The dictionary has the form: {"first word": [id1, counts], "second word": [id2, counts], ... }, where a raw word is in quotes and the square brackets contain the identification number of the word and

调 O 查 O 范 O 围 O 涉 O 及 O 故 B-LOC 宫 I-LOC 、O 历 B-LOC 博 I-LOC 、O 古 B-ORG 研 I-ORG 所 I-ORG 、O 北 B-LOC 大 I-LOC 清 I-LOC 华 I-LOC 图 B-LOC 书 I-LOC 馆 I-LOC 、O 北 B-LOC 图 I-LOC 、O 日 B-LOC 伪 O 资 O 料 O 库 O 等 O 二 O 十 O 几 O 家 O , O 言 O 及 O 文 O 物 O 二 O 十 O 万 O 件 O 以 O 上 O , O 洋 O 洋 O 三 O 万 O 余 O 言 O , O 是 O 珍 O 贵 O 的 O 北 B-LOC 京 I-LOC 史 O 料 O 。	The survey covers the Forbidden City, Libo, Institute of antiquity, Tsinghua University Library, beitu, Japanese and puppet databases and more than 200,000 cultural relics and 30,000 foreign words, which are precious historical materials of Beijing.
微 B-CRM 信 I-CRM 催 B-CRM 眠 I-CRM 水 I-CRM 催 B-CRM 情 I-CRM 粉 I-CRM 催 B-CRM 情 I-CRM 药 I-CRM 催 B-CRM 情 I-CRM 藥 I-CRM 挫 B-CRM 仑 I-CRM 毕 B-CRM 业 I-CRM 证 I-CRM 答 B-CRM 案 I-CRM 包 I-CRM 答 B-CRM 案 I-CRM 提 I-CRM 供 I-CRM 发 B-CRM 票 I-CRM 出 I-CRM 发 B-CRM 票 I-CRM 代 I-CRM 发 B-CRM 票 I-CRM 销 I-CRM 發 B-CRM 票 I-CRM 蒙 B-CRM 汗 I-CRM 药 I-CRM 迷 B-CRM 幻 I-CRM 型 I-CRM 迷 B-CRM 幻 I-CRM 药 I-CRM 迷 B-CRM 幻 I-CRM 藥 I-CRM 迷 B-CRM 昏 I-CRM 口 I-CRM 迷 B-CRM 昏 I-CRM 药 I-CRM	Wechat hypnotic water, aphrodisiac powder, aphrodisiac drug, Tulun graduation certificate, answer package, provide invoice, replace invoice, sell invoice, Mongolian sweat drug, psychedelic drug, psychedelic drug, aphrodisiac drug

Figure 4. BIO tagging style.

the number of occurrences. Based on the counts, words with low frequencies are eliminated from the dictionary. Figure 5 shows the data processing step.

- **Step 2: Model Setup:** The LSTM-CRF model was employed in the experiments – the LSTM layer is located at the bottom whereas the CRF layer is located on top. The softmax function was used to compute the probabilities of each target class over all possible target classes.

The hyper-parameters used in the experiments are shown in Table 2. The batch size was set to 64, meaning that 64 samples were trained in each epoch. The epoch value was set to 10, meaning that each sample was trained ten times over the experiment. The dimension of the hidden

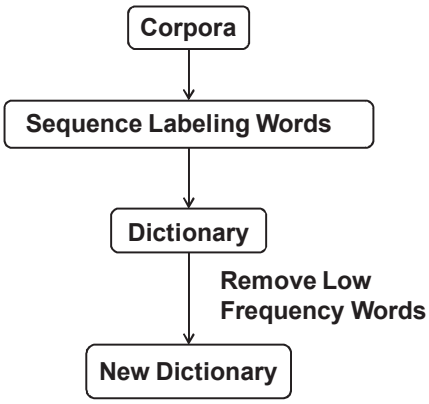

Figure 5. Data processing.

Table 2. Hyper-parameter values.

Parameter	Value	Parameter	Value
batch_size	64	clip	5
epoch	10	dropout	0.5
hidden_dim	300	update_embedding	TRUE
optimizer	Adam	embedding_dim	300
lr	0.001	pretrain_embedding	random

state (hidden_dim) was 300, the optimizer was Adam, the learning rate (lr) was 0.001 and the gradient clipping (clip) was 5.

Table 3. Dataset proportions.

Corpus	Training Dataset	Testing Dataset
Normal Part	1,595,064	177,231
Crime Part	11,670	1,295
Total Number	1,606,734	178,526
Proportion	90%	10%

Table 3 shows the proportions of the training dataset and testing dataset. The amount of training data was set to 1,606,734 and the amount of testing data was set to 178,526, corresponding to proportions of 90% and 10%, respectively. In the experiment, the extracted tokens were used as the basic unit of processing. After each epoch, the loss function value, global step, precision, recall and F1 score were recorded.

Example 1	据警方報導示威者準備去中環站打警察	The police said demonstrators will set fire and beat cops at the central station
	PERSON: ["示威者"] LOCATION: ['中環站'] ORGANIZATION: ['警方'] CRIM: ['打警察']	PERSON: ['demonstrators'] LOCATION: ['central station'] ORGANIZATION: ['The police'] CRIM: ['beat cops']
Example 2	如果警方想/引你們暴亂，令(全球)怪罪你們，你們便全敗！他們脫罪！明未？😡官方中有秦會和高球，死的只會是岳家軍和水浒军！ 😡😡😡聰明些吧！必須(立即)停暴！你們才能勝！你們影响民生時，人民只會放棄你們而(不會)幫你們迫政府的！要勝必須停！！！	If the police want to / lead you to riot and blame you all over the world, you will all be defeated! They get away with it! Do you understand? There are Qin Hui and Gao Qiu in the official, only Yue Jiajun and Shui hujun will die! Be smart! The violence must be stopped (immediately)! You can win! When you influence people's livelihood, the people will only give up on you and (will not) help you to force the government! To win must stop!!!
	PERSON: ['秦會'，'高球'] LOCATION: ['全球'] ORGANIZATION: ['警方'，'岳家軍'，'水浒军'，'人民'，'政府'] CRIM: ['暴亂'，'脱罪'，'停暴']	PERSON: ['Qin Hui'，'Gao Qiu'] LOCATION: ['all over the world'] ORGANIZATION: ['the police'，'Yue Jiajun'，'Shui Hujun'，'the people'，'the government'] CRIM: ['riot'，'get away with it'，'The violence must be stopped']

Figure 6. Two experimental examples.

6. Experimental Results and Discussion

After being trained, the LSTM-CRF model was able to identify the four entities in a sentence. Figure 6 shows the input sentence: "The police said demonstrators will set fire and beat cops at the central station." The model was able to recognize person as demonstrators, location as the central station, organization as the police and the crime as set fire and beat cops.

In the second example in Figure 6, the model was able to identify person as Qin Hui and Gao Qiu, location as the world, organization as the police, Yue Jiajun, Shui Hujun, the people and the government, and crime as riot, get away with it and the violence must be stopped.

Table 4 compares the LSTM-CRF and LSTM models. The LSTM-CRF model produces better results. The LSTM model is good at learning the sequential relationships of entities (i.e., words in this study) automatically, but it ignores the sequential relationships of labels. On the other hand, the CRF model is good at learning the sequential relationships of labels. Since the CRF model addresses the LSTM model deficiencies, the LSTM-CRF model performs better than the basic LSTM model.

As shown in Table 4, when the crime part is eliminated, higher values for precision of 90.37%, recall of 86.27% and F1 score of 88.27% are

Table 4. LSTM and LSTM-CRF model evaluation.

Models	Precision	Recall	F1 Score
LSTM	84.16%	82.07%	83.11%
LSTM-CRF	87.58%	83.22%	85.24%
LSTM-CRF (Without Crime Part)	90.37%	86.27%	88.27%

obtained. This is because the other three parts (person, location and organization) have been studied in public datasets by other researchers, but only this research contains the crime part. Also, since the corpus was created for crimes, it is more difficult to train the model to recognize crime-related entities.

As expected, the models produce different results. The named entity recognition technique is limited in that it only extracts key entities from text, but does not analyze the entities. As a consequence, the investigator has to analyze the extracted entities and make manual decisions.

7. Conclusions

Automated monitoring of social media platforms and online discussion groups can provide insights into potential criminal events, enabling law enforcement to work proactively to mitigate the hazards. The combined LSTM-CRF model described in this chapter is able to extract key information from large volumes of Chinese text using named entity recognition. Experiments indicate that the automated extraction of key attributes such as person, location, organization and crime is accomplished with a maximum precision of 87.58%, recall of 83.22% and F1 score of 85.24%. These results demonstrate that the methodology is effective at discovering potential criminal events.

Due to the absence of crime-related corpora, custom corpora had to be created for training and testing. Future research will focus on developing richer and larger corpora with criminal events. Training the model using these corpora would improve the overall performance.

A limitation of the methodology is that, while it identifies key entities, it cannot analyze them. Yang and Chow [14] have employed statistical methods to create relationships between entities. Future research will pursue this line of inquiry and also focus on relation extraction and emotional analysis using deep learning techniques.

References

[1] H. Chan, In pictures: 12,000 Hongkongers march in protest against "evil" China extradition law, organizers say, *Hong Kong Free Press*, March 31, 2019.

[2] N. Greenberg, T. Bansal, P. Verga and A. McCallum, Marginal likelihood training of BiLSTM-CRF for biomedical named entity recognition from disjoint label sets, *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pp. 2824–2829, 2018.

[3] S. Hochreiter and J. Schmidhuber, Long short-term memory, *Neural Computation*, vol. 9(8), pp. 1735–1780, 1997.

[4] Z. Huang, W. Xu and K. Yu, Bidirectional LSTM-CRF Models for Sequence Tagging, `arXiv:` 1508.01991v1, 2015.

[5] A. Katiyar and C. Cardie, Nested named entity recognition revisited, *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, vol. 1 (Long Papers), pp. 861–871, 2018.

[6] B. Kleinberg, M. Mozes and A. Arntz, Using named entities for computer-automated verbal deception detection, *Journal of Forensic Sciences*, vol. 63(3), pp. 714–723, 2018.

[7] G. Lample, M. Ballesteros, S. Subramanian, K. Kawakami and C. Dyer, Neural architectures for named entity recognition, *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 260–270, 2016.

[8] C. Marcum, *Cyber Crime*, Wolters Kluwer, Frederick, Maryland, 2014.

[9] Pudn, MSRA (`www.pudn.com/Download/item/id/2435241.html`), 2020.

[10] C. Santos and V. Guimaraes, Boosting Named Entity Recognition with Neural Character Embeddings, *arXiv:* 1505.05008v2, 2015.

[11] Sougou, Sougou Corpus (`pinyin.sougou.com`), 2020.

[12] Z. Wang, X. Cui, L. Gao, Q. Yin, L. Ke and S. Zhang, A hybrid model of sentimental entity recognition on mobile social media, *EURASIP Journal on Wireless Communications and Networking*, vol. 2016, article no. 253, 2016.

[13] D. Xu, R. Ge and Z Niu, Forward-looking element recognition based on the LSTM-CRF model with the integrity algorithm, *Future Internet*, vol. 11(1), article no. 17, 2019.

[14] M. Yang and K. Chow, An information extraction framework for digital forensic investigations, in *Advances in Digital Forensics XI*, G. Peterson and S. Shenoi (Eds.), Springer, Cham, Switzerland, pp. 61–76, 2015.

[15] J. Zhang and X. Liu, Research on Chinese named entity recognition based on deep learning, *Proceedings of the Fourth IEEE International Conference on Computer and Communications*, pp. 2142–2147, 2018.