HRCE: Detecting Food Security Events in Social Media

Wang Gao 1*, Yuan Fang 2, Yuwei Wang 1 and Fan Zhang 3

1 School of Mathematics and Computer Science, Jianghan University, Wuhan, 430070, China
2 School of Computer Science and Technology, Wuhan University of Technology, Wuhan, 430070, China
3 College of Computer Science, Wuhan Donghu University, Wuhan, 430070, China
* Corresponding author’s e-mail: gaow@jhun.edu.cn

Abstract. Analyzing food security events shared on social networks not only helps people deepen their understanding of food security events, but also helps managers cope with these events. In this paper, we propose a model that utilizes task-specific features and a deep learning model to detect food security events from tweets, called HRCE. Specifically, the proposed model leverages a hierarchical Recurrent Neural Network (RNN) with Long Short-Term Memory (LSTM) that takes word embeddings as inputs, and combines contextual embeddings to identify food security events from social media. We collected a novel food security related dataset from Twitter, and manually annotated 2,418 tweets. We conducted experiments on this dataset and concluded that HRCE outperforms baseline methods in terms of precision, recall and F1-score.

1. Introduction

Social media, such as Microblog and Twitter, have become a new kind of medium in which people can share timely information about anything. Individuals share news, facts, opinions, and discuss them through this information. Due to the convenience of social media, people also share food security events such as food poisoning, food contaminations, hogwash oil and so forth in their tweets. As shown in Figure 1, automatically extracting these events has several practical applications, such as timely control and reduction of the hazards of food security incidents.

At present, food security related text mining has attracted great interest from researchers in food security and Natural Language Processing (NLP) [1-2]. However, extracting food security events from social media is a challenging task, mainly because tweets are very noisy and often lack sufficient...
background to identify food security incidents. In recent years, classification methods based on deep learning have achieved better results than traditional methods in various NLP tasks [3-4]. Inspired by this, we propose a Hierarchical RNN with Contextual Embeddings (HRCE) model to classify social media posts related to food security events. The key contributions of this work are summarized as follows: (1) HRCE attempts to apply a hierarchical RNN model with attention mechanisms to text classification tasks, which are used to detect food security events in social media. To the best of our knowledge, this is the first work of this kind. (2) We learn task-specific features to capture the contextual information, thereby reducing the impact of noise and improving classification performance. (3) We conducted experiments on a real-world social media dataset. Experimental results show that our model can significantly outperform baseline methods.

2. Related Work

There have been many studies on event detection for Twitter datasets. Atefeh et al. developed a method of extracting relational tuples from a Twitter corpus without manual tagging of data [5]. Harris et al. proposed a supervised machine learning approach to capture tweets, including “food poisoning”, “foodpoisoning” or both, and classified them as relevant, unclear or irrelevant [6]. Chang et al. proposed an LSTM based method that automatically captures tweet-level features to detect events from tweets [7]. Tonon et al. proposed an approach to identify important events in Twitter datasets through semantic association analysis [8]. Meyer et al. used Twitter and Wikipedia as data sources to report some of the initial insights into social media analysis and web data mining to monitor food safety events that may develop into potential crises [9]. Jiang et al. proposed a computational processing pipeline to study the role of social media (especially Twitter) for monitoring dietary supplement safety [10]. Our model is different from previous studies because we define the detection of food security events as a binary classification task, and employ a hierarchical RNN architecture to learn tweet level representations, while combining task-specific characteristics to improve classification performance.

3. Methodology

An overview of the proposed model is shown in Figure 2. In the next subsections, we will address the challenges of food security event detection.

### 3.1. Task-specific Features

The proposed model captures task-specific features from tweets to understand contextual information.

**CRFTM:** Conditional Random Field regularized Topic Model (CRFTM) is a probabilistic graphical model that leverages word embeddings to discover topics from short texts [4]. CRFTM works in an unsupervised manner and represents short text as a mixture of hidden topics, where each hidden topic is a probability distribution of words. After training the CRFTM model, we summarize each short text by using the topic label with the maximum likelihood. For example, the topic of the tweet in Figure 1 is "food hygiene"
NER: Named Entity Recognition (NER) is a basic task in NLP that identifies specified named entities from texts, and pave the way for tasks such as relationship classification. Named entities generally refer to entities with specific meaning or strong reference in the text, usually including person names, place names, organization names, date, time and so forth. We use an annotated corpus containing entities from food safety domain to train a conditional random field model using word-level features such as uni-gram, bi-gram and part-of-speech. The corpus contains 135K tokens including named entities such as "restaurant", "powdered milk", "formaldehyde", "additive", "unclean food", in the standard inside-outside-beginning format. Our NER model labels "Yoshino's" as "restaurant" and "cockroach" as "unclean food" in the tweet of Figure 1.

Contextual Encoder: HRCE employs the word embeddings of the output tokens of CRFTM and NER algorithms to extract a combined vector representation. Therefore, contextual embeddings can be calculated as follows.

\[ CE(s) = \frac{CT(s) + \sum_{i=1}^{N} NR(s)_i}{1 + N}, \]  

Where \( s \) represents a sentence, \( CE \) function obtains contextual embeddings and \( N \) is the number of the outputs of our NER model. \( CT \) and \( NR \) denote the word embeddings of the output tokens of CRFTM and NER, respectively.

3.2. Hierarchical RNN Model

In this section, we propose a hierarchical RNN model to learn tweet level representations. The proposed model gradually constructs a sentence representation through the representation of the word it contains, and then the sentence representations constitute a tweet representation. These tweet representations are used as the input to a softmax classifier and output classification decisions.

To this end, the model uses a series of bidirectional RNN encoders with LSTM. A sequence of tokens \( X_s = \{w_1, w_2, ..., w_s\} \) and contextual embedding \( CE(s) \) of sentence \( s \) are input to the encoder, which computes a backward sequence of hidden states \( (\bar{h}_1, \bar{h}_2, ..., \bar{h}_s) \) and a forward sequence of hidden states \( (\bar{h}_1, \bar{h}_2, ..., \bar{h}_s) \). The sentence representation \( \bar{h}_s \) is obtained by concatenating two hidden states \( \bar{h}_s \) and \( \bar{h}_s \).

\[ v_s = \bar{h}_s \oplus \bar{h}_s, \]  

where \( \oplus \) denotes the concatenation operator.

More precisely, the proposed model contains a bidirectional RNN model with LSTM to learn the sentence representations of a tweet \( t = \{x_1, x_2, ..., x_t\} \). In the traditional models, each sentence embedding can be calculated using the mean, maximum, sum, etc. of word embeddings. In HRCE, we exploit an attention mechanism to calculate the weights and input the weighted sum of these representations to a tweet encoder. The tweet level representation is learned from the sentence representations and model the hierarchical RNN. Finally, the food security event label can be obtained by inputting it into a softmax layer, as shown in Figure 2.

The attention mechanism is employed to weight the specific encoder output of the input sequence. As a result, the attention function refers to mapping a set of key-value pairs and a query to the output, where the key, value, query and output are vectors. The output can be computed by a weighted sum of values, where the weight of each value is computed by the corresponding key and the query's compatibility function.

HRCE employs a hierarchical attention mechanism to reinforce the weight of words and sentences that are important for food security events. The attention weight is calculated as follows:

\[ a_j = \frac{\exp(r^j \cdot \tau_j)}{\sum_{i=1}^{t} \exp(r^j \cdot \tau_i)}, \]  

where \( a_j \) is the weight of the j-th sentence,
where
\[ \tau_s = \text{ReLU}(v_s), \]
where ReLU is a non-linear activation function that stands for rectified linear unit. Therefore, the importance of a unit is measured as the similarity of \( \tau_s \) to the contextual vector \( \tau_c \) that is learned together during the training process. Through a softmax function, importance weight is standardized. Finally, tweet representation \( tr \) is calculated as follows:
\[ tr = \sum_{s=1}^{T} a_s v_s. \]

4. Experiment

In this section, we compare our model with three baseline methods to verify the effectiveness of HRCE on the task of classifying tweets related to food security.

4.1. Dataset

We employ the Twitter's streaming API to collect 1.27M English tweets using a set of seed keywords from August 17, 2017 to January 23, 2019. These seed words are selected by five food security experts, including "additive", "biotoxin", "hogwash oil", "food contamination" etc. to extract relevant tweets. Considering the sparsity problem of food security related tweets in the Twitter datastream, using seed keywords is a practical method to filter noise.

In the pretreatment process, we reserve the keyword RT that represents retweets and delete all emoticons. $mention$ and $url$ are used to replace user handles and hyperlinks, respectively. We exploit the Twitter tokenizer in the NLTK library\(^1\) to tokenize tweets and convert uppercase letters to lowercase.

Furthermore, we hire five experts in the food security field to manually tag the dataset, and require them to provide a binary label for food security events in given tweets. If they are not sure of their decision, they can skip the tweet. When at least 4 of the 5 annotators are consistent, we will accept the annotation. As a result, the quality of the labels we obtain is reliable. In the end, 2,418 tweets are tagged. The tagged Twitter dataset is balanced because there are 1216 non-event samples and 1202 event-related samples. The training and testing datasets have 1814 and 604 tweets, respectively.

4.2. Baselines

We compare our model with three baseline methods: (1) BOW-SVM is a model for training a Support Vector Machine (SVM) classifier using a bag-of-words (BOW) that treats tweets as a collection of words. (2) W2V-SVM is a model that uses word embeddings to train an SVM classifier. (3) W2V-CNN proposed by Kim employs convolutional neural networks to classify tweet [11].

For W2V-CNN, we select the parameters based on the original paper. For SVM, we use a linear kernel SVM classifier in sklearn\(^2\) according to default parameter settings. For word embeddings, we use open source pre-trained 300-dimensional word vectors\(^3\) from Google News datasets.

We implemented our hierarchical RNN model using the Keras toolkit with LSTM. Follow [12], the size of the hidden units of the encoder is set to 50, and the size of a mini-batch is set to 70. Additionally, we utilize stochastic gradient descent to train the hierarchical RNN model with momentum of 0.9. We use grid search to select the best learning rate and train the model to minimize the classification cross entropy loss.

---

\(^1\) http://www.nltk.com/.
\(^2\) http://scikit-learn.org/.
\(^3\) https://code.google.com/p/word2vec.
4.3. Evaluation

To compare different models, we use three commonly metrics: precision (PR), recall (RC) and F1-score (F1), which can be calculated as follows:

\[
PR = \frac{TP}{TP + FP},
\]

\[
RC = \frac{TP}{TP + FN},
\]

\[
F1 = \frac{2 \times PR \times RC}{PR + RC},
\]

where \(TP\) represents the number of true samples classified as positive, \(FP\) represents the number of false samples classified as positive, and \(FN\) represents the number of false samples classified as negative.

![Figure 3. The classification results of food security events.](image)

4.4. Results and Discussions

Figure 3 shows the classification results of food security events for each method. The result of each report is the average of a five-fold cross-validation experiment. Obviously, our method is superior to various simple and neural baselines. Therefore, we believe that task-specific features can capture the contextual information, which helps to understand the semantics of tweets. In addition, considering the sequential characteristics of the texts made by the hierarchical RNN model, it can be useful to analyze short texts related to food security. In addition, the classification results are improved because of our hierarchical attention mechanism. These findings are consistent with the results presented by [13].

5. Conclusion

In this work, we propose a novel model for food security events detection, namely Hierarchical RNN with Contextual Embeddings (HRCE) model. HRCE first employs task-specific features to capture the contextual information. Furthermore, the proposed model exploits a hierarchical RNN model with attention mechanisms to text classification tasks. We provide a unique dataset of food security related tweets collected from Twitter, and conduct experiments on this short text dataset. The experiment results show the effectiveness of HRCE compared with baselines. In the future, we will consider hyperlinks and user handles that may provide additional information to improve detection accuracy.

Acknowledgments

We would like to thank Yan Xiao for her efforts on data processing.
References

[1] Monica L.N., Noel P.G. (2011) Rule-based complex event processing for food safety and public health. In: the international conference on rule-based reasoning, programming, and applications. Barcelona. pp. 376-383.

[2] Fang L., Yawei L., Qunxiong Z., Xiaoyong L. (2015) Research of food safety event detection based on multiple data sources. In: the international conference on cloud computing and big Data (CCBD). Shanghai. pp. 213-216.

[3] Issa A., Keng H. (2018) Term weighting scheme for short-text classification: Twitter corpuses. Neural Computing and Applications. (1): 1-13.

[4] Wang G., Min P., Hua W., Yanchun Z., Qianqian X., Gang T. (2018) Incorporating word embeddings into topic modeling of short text. Knowledge and Information Systems. (12): 1-23.

[5] Farzindar A., Wael K. (2015) A survey of techniques for event detection in twitter. Computational Intelligence. 31(1): 132–164.

[6] Jenine K. H., Jared B. H., Leila N., Elaine O. N., Gaurav T., Raed M., John S. B. (2017) Using twitter to identify and respond to food poisoning: the food safety STL project. Journal of Public Health Management and Practice. 23(6): 577-580.

[7] Chingyun C., Zhiyang T., Yue Z. (2016) Expectation-regulated neural model for event mention extraction. In: the North American chapter of the association for computational linguistics (NAACL). San Diego. pp. 400-410.

[8] Alberto T., Philippe C., Albert B., Vincent L., Boris M. (2017) Armatweet: detecting events by semantic tweet analysis. In: European semantic web conference (ESWC). Portorož. pp. 138–153.

[9] Christian H. M., Martin H., Wiltrud T., Johannes R., Patrick P. (2015) Web data mining and social media analysis for better communication in food safety crises. International Journal on Food System Dynamics. 6(3): 129-138.

[10] Keyuan J., Yongbing T., G. Elliott C., Michael M. M. (2017) Discovering potential effects of dietary supplements from twitter data. In: the international conference on digital health. London. pp. 119-126.

[11] Yoon K. (2014) Convolutional neural networks for sentence classification. In: the conference on empirical methods in natural language processing (EMNLP). Doha. pp. 1746-1751.

[12] Kyunghyun C., Dzmitry B., Fethi B., Holger S., Yoshua B. (2014) Learning phrase representations using RNN encoder–decoder for statistical machine translation. In: the conference on empirical methods in natural language processing (EMNLP). Doha. pp. 1724-1734.

[13] Zichao Y., Diyi Y., Chris D., Xiaodong H., Alex S., Eduard H. (2016) Hierarchical attention networks for document classification. In: the North American chapter of the association for computational linguistics (NAACL). San Diego. pp. 1480–1489.