symmetry

MDPI

Article

# Food Safety Event Detection Based on Multi-Feature Fusion

**Kejing Xiao [1,2], Chenmeng Wang [1,*], Qingchuan Zhang [1] and Zhaopeng Qian [3]**

[1] National Engineering Laboratory for Agri-Product Quality Traceability, Beijing Technology and Business University, Beijing 100048, China; xiaokejing0501@163.com (K.X.); zhangqingchuan@btbu.edu.cn (Q.Z.)
[2] School of Information, Renmin University of China, Beijing 100872, China
[3] School of Biological Science and Medical Engineering, Beihang University, Beijing 100191, China; qianzhaopeng@buaa.edu.cn
* Correspondence: jane0712@163.com; Tel.: +86-010-68985674

Received: 22 August 2019; Accepted: 24 September 2019; Published: 1 October 2019


**Abstract:** Food safety event detection is a technique used to discover food safety events by monitoring online news. In general, a set of keywords are extracted as features to represent news, and then the news is clustered to generate events. The most popular method for news feature extraction is Term Frequency-Inverse Document Frequency (TF-IDF), however, it has some defects such as being prone to the "dimension disaster", low computational efficiency, and a lack of semantic information. In addition, Latent Dirichlet Allocation (LDA) is also widely used in news representation. Despite its low dimension, it still suffers from some drawbacks such as the need to set a predefined number of clusters and has difficulty recognizing new events. In this paper, a method based on multi-feature fusion is proposed, which combines the TF-IDF features, the named entity features, and the headline features to represent the news. Based on the representations, the incremental clustering method is used to cluster the news documents and to detect food safety events. Compared with the traditional methods, the proposed method achieved higher Precision, Recall, and F1 scores. The proposed method can help regulatory authorities to make decisions and improve the reputation of the government, whilst reducing social anxiety and economic losses.

**Keywords:** food safety; multi-feature fusion; event detection; TF-IDF

## 1. Introduction

Topic Detection and Tracking (TDT) is an information processing technique for the information flow on news media [1], which can detect the appearance of new topics and track their reappearance and evolution [2], whilst helping people deal with the problem of the internet information explosion [3]. Topic detection is a sub-task of TDT, which can help decision makers find meaningful topics or events in a timely manner [4] and has attracted a great deal of attention in many application areas, such as public opinion monitoring, emergency management, decision-making support systems, and online reputation monitoring [5–8]. In the context of news, topic detection and event detection can be viewed as the same concept [9]. Food safety event detection is very important for governments and for society. In recent years, food safety events have occurred frequently, making rapid food safety event detection an urgent problem to be solved. Food safety events include food poisoning, food-borne diseases, food contamination, etc. Examples include the horsemeat scandal that occurred in Europe [10], rat meat that was found in famous snacks in Korea [11], and the melamine, Sudan red egg, the gutter oil scandals that occurred in China [12–14]. These events not only caused huge economic losses and brought anxiety to the public, but also seriously undermined the reputation of the relevant governments.

Several approaches have been proposed for events detection, including: (1) Document clustering based on news feature extraction and representation [15–17], wherein most researchers use Term Frequency-Inverse Document Frequency (TF-IDF) [18] to extract keywords and Vector Space Model (VSM) [19] to represent news, then the clustering algorithms such as single-pass [20] or k-means [21] are used to cluster news (news describing the same event are clustered to generate events); (2) the method based on a topic model [22], where Latent Dirichlet Allocation (LDA) [23], Probabilistic Latent Semantic Analysis (PLSA) [24], and various extension versions are used to explore the latent semantic knowledge of documents, i.e., treating each document as a probability distribution over topics, then representing news based on this distribution and clustering the news accordingly; (3) The method based on neural networks [25], which uses deep neural network models such as Doc2vec [26] and Sentece2vec [27] to obtain document vectors, and then clustering document vectors to generate events; (4) The community partitioning method based on a complex network [28], which takes co-occurrence words as nodes in the network to establish a topic graph and detect topics by using community partitioning. The community partitioning methods include the Kernighan–Lin algorithm [29], Blondel algorithm [30], etc. There are a few studies currently available about food safety event detection [31,32]. These studies are based on LDA and have lower data dimensions, achieving better results than methods based on TF-IDF [31,32].

Nevertheless, the document representation ability of the above research is still limited by the low semantic information, and the inference algorithm used in the model can be too complex [4]. In addition, such methods need manual labeling of events and setting a predefined number of clusters, and have difficulty in detecting new events [33], which is not conducive to large-scale data modeling and affects the precision of event detection.

In this paper, TF-IDF is used to calculate the weights of all words in the news, and a fixed number of words are selected as the feature of news. McMinn et al. [34] proposed a real-time entity-based event detection method for Twitter, which proved that named entities play a crucial role in describing an event. Their entity-based event detection method is able to detect more events than previous approaches whilst also providing improved precision and retaining low computational complexity. Therefore, we use the named entity as a part of the features to represent news in this paper, and combine it with the feature words obtained by TF-IDF to form the joint feature of documents. This entity can significantly reduce the data dimension and computation overhead. In addition, it can retain the important news information effectively. Furthermore, the news headlines of food safety event can effectively summarize the news content, therefore, this paper uses the semantics of headlines to update the weight of the joint features, so that words with high similarity to the headlines have greater weights. In this paper, we proposed the concept of news "fusion feature", which fuses multiple features together, including the TF-IDF features, the named entity features, and the headline features. In this way, key information can be more prominent and document semantics can be highly covered, meaning more accurate representation of the news can be obtained to improve the event detection results.

The main contributions of this paper lie in: (1) the combination of TF-IDF features and named entity features used to form the joint feature of news; (2) a method for updating the weight of feature words based on semantics of headlines, which highlights the key information and allows the fusion feature of news to be obtained. The multi-feature fusion method proposed in this paper is used to document a representation of food safety news, which has enhanced the detection results of food safety events, and can help regulatory authorities to more accurately detect food safety events. In order to verify the effectiveness of the method proposed in this paper, experiments were carried out on real food news data, and the experimental results of TF-IDF, LDA, and multi-feature fusion are compared.

## 2. Methods

This paper proposes a food safety event detection method based on multi-feature fusion, and the process was as follows: (1) preprocessing the news data; (2) TF-IDF is used to calculate the weight of each word in the news document, then the first M words with the largest weight of each news document are selected to form a feature words set W; (3) the named entities in the news document are

recognized by using the Bi-LSTM-CNN-CRF framework [35] to form the set E, then the joint feature set K is obtained by combining E and feature set W; (4) the word2vec is used to obtain the vector of all words in the news dataset, then establish the dictionary D and corresponding word vector set $V_D$; (5) establish the headline vector $V_h$ of the headlines in the dataset which has been preprocessed by (1), calculate the similarity between each word in the feature set K and headline vector, and update the weight of the feature words according to the similarity value, then the VSM is used to represent the news document; and (6) the single-pass algorithm is used to cluster the news documents and generate events. The process is outlined in Figure 1 and described in detail in the next subsections.

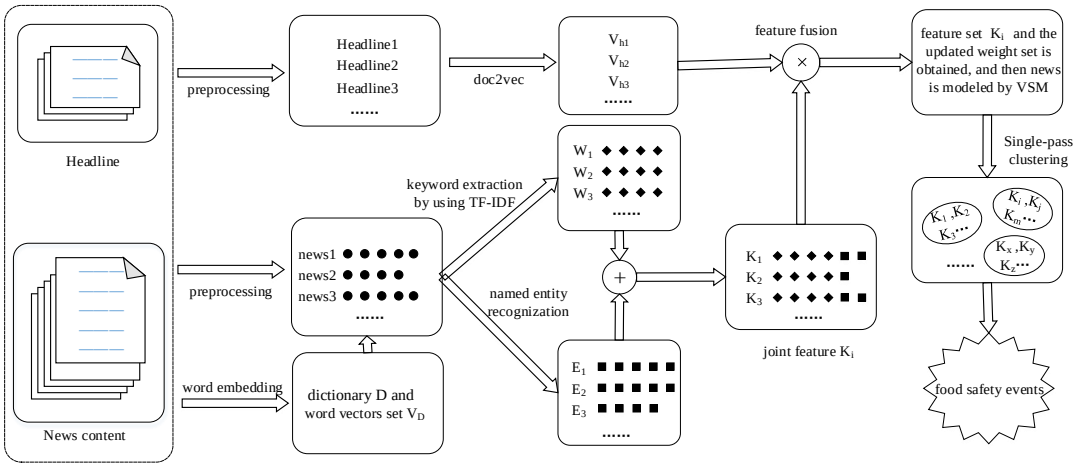


**Figure 1.** Overview of food safety event detection based on multi-feature fusion, where D denotes the dictionary constituted by all words in news data and $V_D$ denotes the word vectors corresponding to D. $\oplus$ denotes the feature joint calculator, $\otimes$ denotes feature fusion calculator, $W_i$ denotes the keywords set of each news document extracted by Term Frequency-Inverse Document Frequency (TF-IDF), and $E_i$ denotes the named entities set in a news document. $K_i$ denotes the set of joint feature words and $V_{hi}$ denotes the headline vector. The black dots denote the words in bag of words after preprocessing, the black diamonds denote the keywords feature extracted by TF-IDF, and the black squares denote the named entities in the news content.

## 2.1. Preprocessing

The data preprocessing includes filtering noise, removing meaningless symbols such as space and links, word segmentation, and stop words. The news dataset $S$ contains plenty of news documents, each news document is represented as a word bag and recorded by a set $news_i$ after preprocessing, and as the input of the subsequent components, as shown as Formula (1)

$$\begin{aligned} S &= \{news_1, news_2, news_3, \ldots news_i \ldots, news_m\} \\ news_i &= \left\{word_1, word_2, \cdots word_j \cdots word_n\right\} \end{aligned} \tag{1}$$

where $m$ is the number of news documents in the news dataset $S$, and $n$ is the number of words in each news document.

## 2.2. TF-IDF Feature Extraction

TF-IDF is a feature extraction algorithm, where TF denotes word frequency, that is, the frequency of a word appearing in the document, and IDF denotes the inverse document frequency. The main idea is that if a word or phrase appears more frequently in one document and less frequently in other documents, it is considered to have good representation ability for the document. Generally, the words

or phrases with higher TF-IDF values are more important in the documents. The $tf$ of the word $t_i$ appearing in document $d_j$ is calculated by Formula (2):

$$tf_{i,j} = \frac{n_{i,j}}{\sum_k n_{k,j}} \tag{2}$$

where $n_{i,j}$ is the number of occurrences of word i in document $d_j$, $\sum_k n_{k,j}$ is the sum of the TF of all the words in document $d_j$, which is the normalization process. $idf$ is the reverse document frequency and is calculated by Formula (3):

$$idf_i = \log\left(\frac{|D|}{1 + \left|j : t_i \in d_j\right|}\right) \tag{3}$$

where $|D|$ is the total number of documents in the dataset $S$, $\left|\left\{j : t_i \in d_j\right\}\right|$ denotes the number of documents containing the word $t_i$. In general, $1 + \left|\left\{j : t_i \in d_j\right\}\right|$ is used as denominator to avoid it being zero. TF-IDF is calculated by Formula (4).

$$\text{TF} - \text{IDF} = tf_{i,j} * idf_i \tag{4}$$

Then the weights of all words are calculated by TF-IDF, for each news document, the first M words with the largest weights are selected to form the feature set $\text{W} = \{w_1, w_2 \cdots w_M\}$, the corresponding weights set is $\theta = \{\theta_1, \theta_2, \cdots \theta_M\}$.

### *2.3. Named Entity Feature Extraction*

Named entities include person names, place names, organization names, and proper nouns. In this paper, named entities are regarded as one of the features of food safety news. The Bi-LSTM-CNN-CRF framework is used to recognize named entities in a food safety news dataset, the framework is based on Bi-directional Long Short-Term Memory (Bi-LSTM) [36], Convolutional Neural Networks (CNN) [37], and Conditional Random Field (CRF) [38]. The steps are as follows: firstly, word embedding is used to obtain the vectors of words; then CNN is used to encode character-level information of a word into its character-level representation, then the character and word-level representations are fed into Bi-LSTM to the model context information of each word. Finally, a sequential CRF is used to jointly decode labels for the whole sentence. For each news, the extracted named entities set can be expressed as $\text{E} = \{e_1, e_2, \cdots e_N\}$.

Combine the TF-IDF feature set W with the named entity feature set E to obtain the joint feature set of the news, shown as Formula (5).

$$\text{K} = \{w_1, w_2 \cdots w_M\} \cup \{e_1, e_2, \cdots e_N\} = \{t_1, t_2 \cdots t_i \cdots t_T\} \tag{5}$$

where $\text{T} \leq \text{M} + \text{N}$, the weight set of the joint feature set K is $\theta' = \left\{\theta'_1, \theta'_2, \cdots \theta'_i \cdots \theta'_T\right\}$.

### *2.4. Feature Fusion Based on the Semantic of Headline*

In general, the headline for food safety news is a summarization of the news content, as it contains the keywords of a certain food safety event. Figure 2 shows a news document about a food safety event. In this Figure, (a) shows the original news in Chinese and (b) shows the English translation of the news in (a). Through the keywords “苏州”(Suzhou), “喜茶”(Heytea), “苍蝇”(flies) in the headline, we can understand what happened in this food safety event.

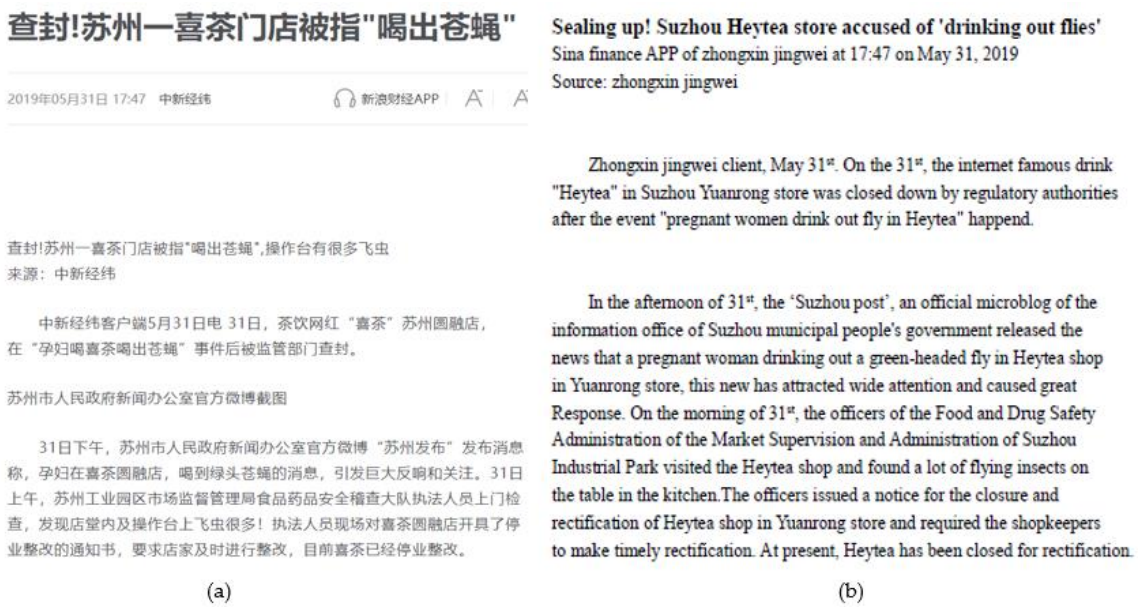


**Figure 2.** Example of food safety news, (**a**) is the Chinese version; (**b**) is the English translated version.

In this paper, a dictionary D of food news dataset is constructed, the vectors of all words in D are obtained by word2vec [39] and form a set $V_D = \{v_1, v_2 \cdots v_i \cdots v_z\}$, where $z$ is the size of D and each vector contains 256 dimensions. The preprocessing of the headlines involves removing punctuation marks, spaces, Chinese word segmentation, and stop words. Then Doc2vec was used to map headlines into vectors with fixed dimensions, thus the headline vector $v_d$ of each news $d$ was obtained and its dimension is also 256. The vectors of words and sentences contain its semantic meaning, while the relationships between words and sentences can be calculated by the vectors.

Each news document is represented by joint feature set K, words with high similarity to headlines can better represent the key information of a news and should be given greater weight. For each word $t_i$ in K, the distance between $t_i$ and the headline vector $v_d$ is calculated by the cosine similarity $\mathrm{s}(v_i, v_d)$, and it is calculated by Formula (6):

$$\mathrm{s}(v_i, v_d) = \frac{v_i \cdot v_d}{||v_i|| ||v_d||} \tag{6}$$

Thus, the similarity of every word in the joint feature set K and headline is obtained and the similarity set is expressed as $S = \{s_1, s_2, \cdots, s_i, \cdots s_T\}$. The updated weights $\delta$ can be obtained by Formula (7).

$$\delta_i = \theta_i' + 2s_i \tag{7}$$

where $\theta_i'$ is the original weight, $s_i$ is the similarity of the word $i$ with the headline, $\delta_i$ is the updated weight of word $i$. Please note the coefficient of the similarity value $s_i$ is determined by our preliminary experiment.

### *2.5. News Representation Based on Multi-Feature Fusion Using the Vector Space Model*

In this paper, the news representation based on multi-feature fusion is modeled by VSM (vector space model). VSM is one of the most popular methods for text modeling, as it regards the news document as a set of unordered words. The joint feature of VSM is shown as set K in Section 2.3, while the calculation process of the words' weights is shown in Sections 2.1–2.4 Thus, the vector space model of a news can be expressed by the weights of the unordered words, as shown in Formula (8):

$$v_{news} = \{\delta_1, \delta_2, \cdots \delta_i, \cdots \delta_D\} \tag{8}$$

where $\delta_i$ is the weight of the $i$-th word, and D is the number of words in the dataset after preprocessing, TF-IDF feature extraction, and named entity recognition. Then the representation of news in the dataset can be obtained as $\left\{v_{news_1}, v_{news2}, \cdots v_{news_i}, \cdots v_{news_N}\right\}$, in which $N$ is the number of news in the dataset. The news representation combined the TF-IDF features, named the entity features, and fused the headline information, thereby developing the news representation model based on multi-feature fusion. In this way, the similarity between different types of news can be calculated and cluster analysis can be performed.

### *2.6. Experiment*

#### 2.6.1. Data Preparation

The food news data in our experiment were gathered from several popular news websites such as Headlines Today (https://www.toutiao.com/), Sina News (https://news.sina.com.cn/), and Sohu News (http://news.sohu.com/) in China, these websites provides valuable and real-time information for people. The news data were used to evaluate the performance and robustness of our approach. The dataset was named "Food Safety News" and contains the human-annotated facts. It contains 1255 Chinese news documents and corresponding headlines from 10 events, where each event contains a variable number of news items ranging from 68 to 180. The vocabulary contains 84,198 unique terms after preprocessing. The total time span for the 10 events is from 1 January 2017 to 30 June 2019 (Table 1).

**Table 1.** Safety news data summary.

| Event Number | Year | Event Name | Number of News | Average Words/News |
|---|---|---|---|---|
| 1 | 2017 | Mouse in the kitchen of Haidilao hot pot | 125 | 1520.2 |
| 2 | 2017 | Molds exceeded in Three Squirrels | 89 | 1230.0 |
| 3 | 2017 | Tianjin Duliu fake seasoning | 93 | 1134.0 |
| 4 | 2017 | Rice in Jiujiang contaminates by cadmium | 68 | 985.5 |
| 5 | 2018 | "Wumao" snacks are harmful | 135 | 1031.4 |
| 6 | 2018 | Take out food using cooking bags | 175 | 1348.2 |
| 7 | 2018 | African swine fever dumplings | 130 | 940.6 |
| 8 | 2018 | Formaldehyde cabbage | 85 | 1025.8 |
| 9 | 2018 | Debate over salmon standards | 175 | 1457.0 |
| 10 | 2019 | Flies found in Heytea | 180 | 1150 |

Therein the People's Daily [40] annotated corpus, which contains a large number of annotated person names, place names, organizations, and other proper nouns, is combined with the "Food Safety News" dataset to train the named entity recognition model.

#### 2.6.2. Evaluation Metrics

TDT [41] proposed several evaluation metrics for topic detection, including Precision P, Recall R and F1 score. In addition, the Miss Rate ($P_{miss}$) and False Alarm Rate ($P_{fa}$) are also important evaluation metrics of system performance. The evaluation status of event detection is shown in Table 2.

**Table 2.** Evaluation status of event detection.

| Category | Event Related | Event Irrelevant |
|---|---|---|
| detected news | $A$ | $b$ |
| undetected news | $C$ | $d$ |

Where $a$ is the number of detected news stories related to an event, $b$ is the number of detected news stories irrelevant to the event, $c$ is the number of undetected news stories related to the event, and

$d$ is the number of undetected news stories irrelevant to the event. Following the notation in Table 2, the evaluation metrics of TDT are shown in Formula (9):

$$\begin{array}{c} \mathrm{P} = \frac{a}{a+b} \\ \mathrm{R} = \frac{a}{a+c} \\ \mathrm{F} = \frac{2PR}{P+R} \\ P_{miss} = \frac{c}{a+c} \\ P_{fa} = \frac{b}{b+d} \end{array} . \tag{9}$$

In this paper, the detection cost function $C_{det}$ is used to evaluate the system performance [41], which is a metric that combines the miss rate $P_{miss}$ and the false alarm rate $P_{fa}$ proposed in TDT2004 [42] and is calculated by Formula (10):

$$C_{det} = C_{miss} \cdot P_{miss} \cdot P_{target} + C_{FA} \cdot P_{FA} \cdot P_{non-target} \tag{10}$$

where $C_{miss}$, $P_{target}$ , $C_{fa}$ and $P_{non-target}$ are predefined parameters, (TDT2004 set these parameters as 1.0, 0.02, 0.1, and 0.98, respectively). $C_{det}$ is usually normalized by Formula (11):

$$\mathrm{Norm}(C_{det}) = \frac{C_{det}}{\min\left(C_{miss} \cdot P_{target},\ C_{fa} \cdot P_{non-target}\right)}. \tag{11}$$

The lower the $\mathrm{Norm}(C_{det})$ value, the better the system performed [42]. In the experiment, the evaluation metrics of each event were firstly calculated, and then the average value is calculated to determine the system performance.

## 3. Results

In this paper, experiments were designed to compare different methods and verify the advantage of the proposed method. The experiments are consisted of two parts: (1) explore the influence of the TF-IDF feature number M and cluster thresholds T on system performance, then determine the optimal value of M and T; (2) compare the P, R, and F1 score of the proposed method and other methods under the same feature number M and threshold T.

### *3.1. Parameter Selection*

For the first experiment, we set M = 5, 10, 15, 20, . . . , 50, and cluster threshold T = 0.10, 0.15, 0.20, 0.25, 0.30, 0.35, and 0.4 to test the system performance. The $\mathrm{Norm}(C_{det})$ values of the system under different M and T are shown in Figure 3.

From Figure 3, we can see that the $\mathrm{Norm}(C_{det})$ value is different under different M and T. When M < 30, the $\mathrm{Norm}(C_{det})$ value decreases gradually with the increase of M; when M = 30, the $\mathrm{Norm}(C_{det})$ value can reach the minimum value under a different threshold; when M > 30, the $\mathrm{Norm}(C_{det})$ value increases with the increase of M. When the number of features M = 30, the clustering result and the performance of the system achieved the best results, i.e., a lower $\mathrm{Norm}(C_{det})$.

Under different T in Figure 3, we found that the dotted line (threshold T = 0.25) was lower no matter what the M is. So in the following experiments, we used M = 30 as the number of news features and T = 0.25 as the clustering threshold.

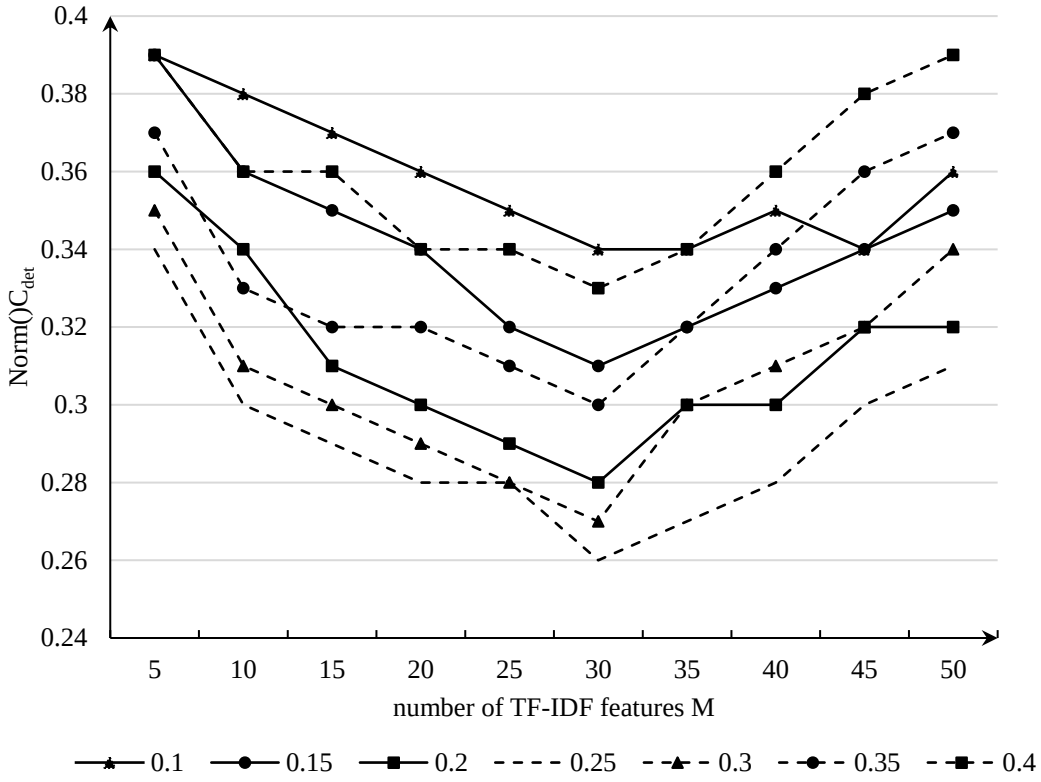


**Figure 3.** Norm($C_{det}$) under different number of features and threshold, the abscissa denotes the number of TF-IDF features M and the ordinate denotes the value of Norm($C_{det}$). The solid line with triangle, solid line with circle, solid line with square, dotted line, dotted line with triangle, dotted line with circle and dotted line with square denote the value of Norm($C_{det}$) when the clustering threshold T = 0.1, 0.15, 0.2, 0.25, 0.30, 0.35, and 0.4, respectively.

### *3.2. Food Safety Event Detection Results*

The food safety event detection method based on multi-feature fusion combines the TF-IDF features with the named entity features and forms the joint features, then fuses the headline features to make the news representation more accurate. In this paper, a single-pass clustering algorithm was used to cluster news and generate food safety events. The experiment compared the results of event detection under the following news representations methods: (1) TF-IDF, (2) LDA, (3) TF-IDF and named entity (TI-NE), and (4) multi-feature fusion.

The Precision P, Recall R, and F1 score of food event detection when the number of TF-IDF features M = 30 and clustering threshold T = 0.25, under different news representation methods are shown in Table 3.

**Table 3.** Comparison of Precision P, Recall R, and F1 score of different methods.

| Method | Precision P | Recall R | F1 Score |
|---|---|---|---|
| TF-IDF | 0.76 | 0.75 | 0.75 |
| Latent Dirichlet Allocation (LDA) | 0.79 | 0.81 | 0.79 |
| TF-IDF and named entity (TI-NE) | 0.86 | 0.84 | 0.84 |
| Multi-feature fusion | 0.94 | 0.93 | 0.93 |

The experimental results show that the Precision P, Recall R, and F1 score of event detection based on LDA are 0.79, 0.81, and 0.79, which are 3%, 6%, and 4% higher than the values of the method based on TF-IDF, which means that the method based on LDA is better than the method based on TF-IDF. After being combined with named entity features (TI-NE), the Precision P, Recall R, and F1 score are

0.86, 0.84, and 0.84, which are higher than the method based on LDA by 7%, 3%, and 5%, which means that the named entity features are important in representing the news documents. Compared with the method based on TI-NE, our method based on multi-feature fusion is 8%, 9%, and 9% higher on the three metrics. Compared with the method based LDA, our method is 15%, 12%, and 14% higher than TI-NE, which only fused the named entities features with TF-IDF. Compared with TF-IDF, our method is 18%, 18%, and 18% higher on the Precision P, Recall R, and F1 score, which proves that our proposed multi-feature fusion method is effective and better than the baseline method based on TF-IDF and LDA.

When using the method based on TF-IDF, the data dimension is equal to the size of the dictionary that constituted by all words in the news dataset, while the dictionary size is 84,198 and provides a high dimensional and sparse matrix, thus leading to low computational efficiency. In the multi-feature fusion method proposed in this paper, the dictionary size is reduced to 4562 after preprocessing, TF-IDF feature extraction, and named entity recognition, which means that the dimension of the news representation based on the multi-feature fusion method is only 4562. Compared to the traditional TF-IDF method, the dimension of news representation is greatly reduced, so the computational efficiency is greatly improved.

## 4. Discussion

A food safety event detection method based on multi-feature fusion is proposed in this paper. The method combines TF-IDF features and named entity features of food news, then the headline features are fused and more accurate news representation is obtained. Finally, the news is clustered based on the representation and events are obtained. The method proposed in this paper solves the problems of a too large data dimension and low computational efficiency of traditional TF-IDF [43], as well as the problems of manual data annotation, which are an inability to identify new events that occurs when using the LDA method [33].

In this paper we designed experiments on the real food safety news dataset. The experimental results show that the value of normalization of detection cost function (Norm($C_{det}$)) varies with the number of TF-IDF features M. When M = 30, the Norm($C_{det}$) can reach the smallest value, this is because when the number of TF-IDF features is less than 30, the smaller the number of features available, the smaller the amount of semantic information contained, thus the content of a news report cannot be represented well; when the number of features is too large, some unimportant information is introduced, which makes the clustering results worse. Therefore, the number of features affects the performance of event detection, meaning an appropriate number of features is important for the system performance. Experimental results show that the system performance at its best when M = 30. The clustering results show that the threshold also affects the clustering results; when the threshold T=0.25, the Norm($C_{det}$) value is the smallest possible value, which indicates that the system performance is the best when the threshold T = 0.25.

Compared the results of event detection based on different news representation methods, when the TF-IDF features are combined with the named entity features, the Precision P, Recall R, and F1 score are better than LDA, which is because the named entities is a part of the important information of the news report, while the joint feature has richer information and the news representation is better. When fusing the headline semantic information, the results are higher than those obtained from other methods, which is because headline is a summarization of the news content. By calculating the similarity between feature words and the headline vector, the weights of the feature words are updated and the key information of the food safety news representation is more prominent, which improves the results of the event detection compared with the method based on TF-IDF and LDA.

The method proposed in this paper also has some limitations, since the data used in this paper is derived from the events that has occurred, so it cannot guarantee the real-time performance of events detection. Nevertheless, the method proposed in this paper still reduced the data dimension, enhanced the results, and more effectively solved the problem of food safety event detection.

In the future, we will focus on combining a variety of data sources and constructing a versatile event detection method, solving the computation overhead, and addressing problems dealing with real-time news feed.

## 5. Conclusions

In this paper, we designed a food safety event detection method based on multi-feature fusion, the method integrates TF-IDF features, named entity features, and used headline features for news representation and food safety event detection, which solves the shortcomings of traditional methods, such as the high dimensionality of data, a lack of semantic information, the need to be labeled in advance, etc., which therefore enhanced the results of event detection. Our proposed methods is of great significance in improving the reputation of governments and reducing social anxiety and economic losses.

**Author Contributions:** Conceptualization, K.X. and C.W.; methodology, K.X.; software, Z.Q.; validation, K.X. and Q.Z.; formal analysis, Q.Z. and Z.Q.; investigation, K.X.; data curation, C.W. and Z.Q.; writing—original draft preparation, K.X. and Z.Q.; writing—review and editing, C.W.

**Funding:** This research was funded by National Key Technology R&D Program of China, grant number 2016YFD0401205 and Open Project Program of National Engineering Laboratory for Agri-product Quality Traceability, grant number AQT-2018-YB4.

**Conflicts of Interest:** The authors declare no conflict of interest.

## References

1. Hong, Y.; Zhang, Y.; Ting, L.; Li, S. Topic detection and tracking review. *J. Chin. Inf. Process.* **2007**, *6*, 77–79.
2. Allan, J.; Carbonell, J.-G.; Doddington, G.; Yamron, J.; Yang, Y. *Topic Detection and Tracking Pilot Study Final Report*; Carnegie Mellon University: Pittsburgh, PA, USA, 2003. [CrossRef]
3. Allan, J. *Topic Detection and Tracking: Event-Based Information Organization*; Springer Science & Business Media, LLC Press: New York, NY, USA, 2012. [CrossRef]
4. Chen, Q.; Guo, X.; Bai, H.-X. Semantic-based topic detection using Markov decision processes. *Neurocomputing* **2017**, *242*, 40–50. [CrossRef]
5. Sakaki, T.; Okazaki, M.; Matsuo, Y. Tweet analysis for real-time event detection and earthquake reporting system development. *IEEE Trans. Knowl. Data Eng.* **2013**, *25*, 919–931. [CrossRef]
6. Yang, Z.-L.; Liu, Y.; Hou, D.-B.; Feng, T.-H.; Wei, Y.; Zhang, J.; Huang, P.; Zhang, G. Water quality event detection based on Multivariate empirical mode decomposition. In Proceedings of the 2014 IEEE International Conference on Systems, Man and Cybernetics (SMC), San Diego, CA, USA, 5–8 October 2014; pp. 2663–2668.
7. Sabbah, T.; Selamat, A.; Selamat, M.-H.; Ibrahim, R.; Fujita, H. Hybridized term-weighting method for dark web classification. *Neurocomputing* **2016**, *173*, 1908–1926. [CrossRef]
8. Spina, D.; Gonzalo, J.; Amigo, E. Learning similarity functions for topic detection in online reputation monitoring. In Proceedings of the 37th International ACM SIGIR Conference on Research & Development in Information Retrieval, Gold Coast, Australia, 6–11 July 2014; pp. 527–536. [CrossRef]
9. Chen, K.-Y.; Luesukprasert, L.; Seng-cho, T.-C. Hot topic extraction based on timeline analysis and multidimensional sentence modeling. *IEEE Trans. Knowl. Data Eng.* **2007**, *19*, 1016–1025. [CrossRef]
10. Falkheimer, J.; Heide, M. Trust and brand recovery campaigns in crisis: Findus Nordic and the horsemeat scandal. *Int. J. Strateg. Commun.* **2015**, *9*, 134–147. [CrossRef]
11. Park, M.-S.; Kim, H.-N.; Bahk, G.-J. The analysis of food safety incidents in South Korea, 1998–2016. *Food Control* **2017**, *81*, 196–199. [CrossRef]
12. Wu, X.-L.; Lu, Y.-Q.; Xu, H.-X.; Lv, M.-Y.; Hu, D.-S.; He, Z.-D.; Liu, L.-Z.; Wang, Z.-M.; Feng, Y. Challenges to improve the safety of dairy products in China. *Trends Food Sci. Technol.* **2018**, *76*, 6–14. [CrossRef]
13. Liu, R.-Y.; Hei, W.-J.; He, P.-L.; Li, Z. Simultaneous determination of fifteen illegal dyes in animal feeds and poultry products by ultra-high performance liquid chromatography tandem mass spectrometry. *J. Chromatogr. B Anal. Technol. Biomed. Life Sci.* **2011**, *879*, 2416–2422. [CrossRef]
14. Lu, F.; Wu, X.-L. China food safety hits the "gutter". *Food Control* **2014**, *41*, 134–138. [CrossRef]

15. Brants, T.; Chen, F.; Farahat, A. A system for new event detection. In Proceedings of the 26th Annual International ACM SIGIR Conference on Research and Development in Informaion Retrieval, Toronto, ON, Canada, 28 July–1 August 2003; pp. 330–337.
16. Kumaran, G.; Allan, J. Text classification and named entities for new event detection. In Proceedings of the 27th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Sheffield, UK, 25–29 July 2004; pp. 297–304.
17. Zhang, K.; Zi, J.; Wu, L.G. New event detection based on indexing-tree and named entity. In Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Amsterdam, The Netherlands, 23–27 July 2007; pp. 215–222.
18. Salton, G.; Buckley, C. Term-weighting approaches in automatic text retrieval. *Inf. Process. Manag.* **1988**, *24*, 513–523. [CrossRef]
19. Salton, G.; Wong, A.; Yang, C.-S. A vector space model for automatic indexing. *Commun. ACM* **1975**, *18*, 613–620. [CrossRef]
20. Allan, J.; Papka, R.; Lavrenko, V. On-line new event detection and tracking. In Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Melbourne, Australia, 24–28 August 1998; pp. 37–45.
21. Li, M.-J.; Ng, M.-K.; Cheung, Y.-M.; Huang, J.-Z. Agglomerative fuzzy k-means clustering algorithm with selection of number of clusters. *IEEE Trans. Knowl. Data Eng.* **2008**, *20*, 1519–1534. [CrossRef]
22. AlSumait, L.; Barbará, D.; Domeniconi, C. On-Line LDA: Adaptive Topic Models for Mining Text Streams with Applications to Topic Detection and Tracking. In Proceedings of the 2008 Eighth IEEE International Conference on Data Mining, Pisa, Italy, 15–19 December 2008; pp. 3–12. [CrossRef]
23. Blei, D.-M.; Ng, A.-Y.; Jordan, M.-I. Latent dirichlet allocation. *J. Mach. Learn. Res.* **2003**, *3*, 993–1022.
24. Hofmann, T. Probabilistic latent semantic analysis. In Proceedings of the 15th Conference on Uncertainty in Artificial Intelligence (UAI), Stockholm, Sweden, 30 July–1 August 1999; pp. 289–296.
25. Hashimoto, K.; Kontonatsios, G.; Miwa, M.; Ananiadou, S. Topic detection using paragraph vectors to support active learning in systematic reviews. *J. Biomed. Inform.* **2016**, *62*, 59–65. [CrossRef] [PubMed]
26. Le, Q.; Mikolov, T. Distributed representations of sentences and documents. In Proceedings of the 31st International Conference on Machine Learning, Beijing, China, 21–26 June 2014; pp. 1188–1196.
27. Arora, S.; Liang, Y.; Ma, T. A simple but tough-to-beat baseline for sentence embeddings. In Proceedings of the ICLR 2017 Conference, Toulon, France, 24–26 April 2017.
28. Zhang, C.; Wang, H.; Cao, L.-L.; Wang, W.; Xu, F.-J. A hybrid term–term relations analysis approach for topic detection. *Knowl. Based Syst.* **2016**, *93*, 109–120. [CrossRef]
29. Kernighan, B.-W.; Lin, S. An efficient heuristic procedure for partitioning graphs. *Bell Syst. Tech. J.* **1970**, *49*, 291–307. [CrossRef]
30. Blondel, V.-D.; Guillaume, J.-L.; Lambiotte, R.; Lefebvre, E. Fast unfolding of communities in large networks. *J. Stat. Mech. Theory Exp.* **2008**, *2008*, P10008. [CrossRef]
31. Liang, M.; Du, J.; Hue, J.; Yang, Y. Study on food safety emergency topic detection model based on semantics. In Proceedings of the International Conference on Advanced Intelligence and Awareness Internet (AIAI 2011), Shenzhen, China, 28–30 October 2011.
32. Liu, J.-S.; Li, Y.-B.; Peng, Y.-Y.; Deng, J.; Chen, X. Detection of Food Safety Topics Based on SPLDAs. In Proceedings of the International Conference on Security and Privacy in Communication Networks, Beijing, China, 24–26 September 2014; pp. 551–555. [CrossRef]
33. Castellanos, A.; Cigarrán, J.; García-Serrano, A. Formal concept analysis for topic detection: A clustering quality experimental analysis. *Inf. Syst.* **2017**, *66*, 24–42. [CrossRef]
34. McMinn, A.-J.; Jose, J.-M. Real-time entity-based event detection for twitter. In Proceedings of the 6th International Conference of the CLEF-Association (CLEF), Toulouse, France, 8–11 September 2015; pp. 65–77. [CrossRef]
35. Ma, X.; Hovy, E. End-to-end sequence labeling via bi-directional lstm-cnns-crf. *arXiv* **2016**, arXiv:1603.01354.
36. Dyer, C.; Ballesteros, M.; Ling, W.; Matthews, A. Transition-based dependency parsing with stack long short-term memory. *arXiv* **2015**, arXiv:1505.08075.
37. LeCun, Y.; Boser, B.; Denker, J.-S.; Henderson, D. Backpropagation applied to handwritten zip code recognition. *Neural Comput.* **1989**, *1*, 541–551. [CrossRef]

38. Ratinov, L.; Roth, D. Design challenges and misconceptions in named entity recognition. In Proceedings of the Thirteenth Conference on Computational Natural Language Learning. Association for Computational Linguistics, Boulder, CO, USA, 4–5 June 2009; pp. 147–155.
39. Mikolov, T.; Sutskever, I.; Chen, K.; Corrado, G.-S.; Dean, J. Distributed representations of words and phrases and their compositionality. In Proceedings of the Advances in Neural Information Processing Systems, Lake Tahoe, NV, USA, 5–10 December 2013; pp. 3111–3119.
40. China Core Newspapers Full-text Database (CCND). People's Daily (1–23 January 2014). Available online: http://gb.oversea.cnki.net/kns55/brief/result.aspx?dbPrefix=CCND (accessed on 1 May 2019).
41. Fiscus, J.-G.; Doddington, G.-R. *Topic Detection and Tracking Evaluation Overview*; Springer: Berlin/Heidelberg, Germany, 2002; pp. 17–31.
42. Trieschnigg, D.; Kraaij, W. TNO Hierarchical topic detection report at TDT 2004. Available online: http://www.cs.ru.nl/~{}kraaijw/pubs/Biblio/papers/TNO-HTD-paper.pdf (accessed on 1 May 2019).
43. Zhu, Z.-L.; Liang, J.; Li, D.-Y.; Yu, H.; Liu, G.-Q. Hot Topic Detection Based on a Refined TF-IDF Algorithm. *IEEE Access* **2019**, *7*, 26996–27007. [CrossRef]