Information Processing and Management 56 (2019) 102086

Contents lists available at ScienceDirect

Information Processing and Management

journal homepage: www.elsevier.com/locate/infoproman

On detecting business event from the headlines and leads of massive online news articles

Yu Qian[a], Xiongwen Deng[a], Qiongwei Ye[*,b], Baojun Ma[*,c], Hua Yuan[a]

[a] School of Management and Economics, University of Electronic Science and Technology of China, Chengdu 611731, China
[b] School of Business, Yunnan University of Finance and Economics, Kunming 650221, China
[c] School of Business and Management, Shanghai International Studies University, Shanghai 201620, China

ARTICLE INFO

Keywords:
Text mining
Business event detection
Word embedding
Online news article

ABSTRACT

Massive online news articles can be a good data resource for detecting the information of business events, which may be useful in many real-world applications. In this paper, we propose a three-step process of "clustering-annotation-classification strategy to extract high-quality information about business events from massive online news headlines and leads. To that end, we first introduce the word embeddings method to represent all the terms in a corpus into word vectors, based on which, we cluster the verbal terms into groups. Then, we introduce an expert to annotate each group of terms with a corresponding business events. Finally, we utilize the extracted information of business events as a classifier to detect the potential events from online news headlines and leads. By evaluating our approach with several state-of-the-art classification algorithms, the results show that our approach offers a competitive performance than the baselines in detecting business events from online news articles.

Findings indicate that the verbal terms in headlines of online news article have a significant effect on identifying business events by improving the performance of our method on *Recall* and $F-value$. On the contrary, the verbal terms in leads provide a more stable performance on *Precision*. As a result, the strategy of combining the headline of an online news article with its lead is a viable option for detecting event information from massive online texts.

1. Introduction

In recent years, online media systems and websites offer a different, and typically faster, source of information on inspecting current business events (Westerman, Spence, & Van Der Heide, 2014). In this paper, the term of business event refers to the activity performed by a firm at a specific time period, such as the organizational activities of investing, marketing, researching and developing. The detection of such a set of business events can be useful in the context of numerous real-world applications, such as industrial trend detection (Han, Hao, & Huang, 2018) and content recommendations for readers and subscribers (Karimi, Jannach, & Jugovac, 2018). Thus, it is a highly managerial research work to extract firms' event information from massive online documents.

There are several crucial challenges that prevent user from extracting event efficiently. Firstly, people discuss a wide variety of topics in an open domain and most of these information are not well tagged, making it unclear in advance what set of event types are appropriate for categorization. To address this problem, in the literature of natural language processing (NLP), there have been

* Corresponding authors.
E-mail addresses: qiany@uestc.edu.cn (Y. Qian), shawndxw@gmail.com (X. Deng), yqw@ynufe.edu.cn (Q. Ye), mabaojun@shisu.edu.cn (B. Ma), yuanhua@uestc.edu.cn (H. Yuan).

https://doi.org/10.1016/j.ipm.2019.102086
Received 19 December 2018; Received in revised form 11 June 2019; Accepted 13 July 2019

significant efforts to develop topic models for detecting events from online news articles (Sprugnoli & Tonelli, 2017). However, this stream of research still suffers from ambiguous definition for events (Li, Ritter, Cardie, & Hovy, 2014). In addition, variation and ambiguity language were commonly used in online texts to describe events (Stieglitz, Mirbabaie, Ross, & Neuberger, 2018), which also has a great impact on the results of such kind of research. Therefore, the following issue should be carefully addressed:

- Q1: How can we identify a business event from massive online textual sources efficiently?

Generally, online news article consists of many elements, such as headline, lead and body section (main text) (Dai, Taneja, & Huang, 2018), in which, the length of text in the body section is longer than that in the headline or lead, and thus the topics included in the body section is inherently noisy and heterogeneous (Liu, Morstatter, Tang, & Zafarani, 2016). In this paper, we notice that, in journalism, the headline can be an abstract of the full article for highlighting the main point of that article (Nir, 1993) and leads emphasize grabbing the attention of the reader by summarizing the key event in the story took place (Spark & Harris, 2011). To reduce the uncertainty of business event detection, this paper aims at extracting business events from massive headline and lead of online news rather than from the whole news article (León, 1997).

In traditional newspapers, the basic function of a headline was to give the reader a clear understanding of what the article was about. However, research results presented by Kuiken, Schuth, Spitters, and Marx (2017) show that, widespread adoption of Internet technologies has changed the way that news is created and consumed, and the function of the headline of an online news article has changed as well. On the Internet, the headline has changed to one of the primary ways to attract the readers attention and so as to lure the reader into opening the contents of online artefacts (e.g. news articles, videos and blogs) (Chen, Conroy, & Rubin, 2015). Therefore, it is important to have a good understanding of the characteristics of the headline and the lead of an online news article. This challenge can be summed up in the following research questions:

- Q2: What is the characteristic of the headline of an online news article in task of business event detection?

To address these problems, this paper presents a framework aimed at extracting business events from online data efficiently. To that end, a neural network based word embedding method is introduced firstly to train the terms in a corpus into word vectors. Based on such a representation of terms, a three-step process of "clustering-annotation-classification strategy is proposed in this paper. Finally, the detected events (as well as their associated triggers) are used as a classifier to identify the potential events from online text. Such a semi-supervised method can well solve the two problems of event definition and event detection as well.

The paper is organized as follows. Section 2 is the related work. Section 3 presents the research framework as a whole. Section 4 and Section 5 detail the proposed event clustering and event detecting technology respectively. Section 6 illustrates the experimental results of the proposed method on a real dataset. Section 7 concludes the paper.

2. Related work

Our work is mainly related to the various formation of events and their detection methods.

2.1. Formation of event

In the literature of NLP, the definition of event differs widely (Sprugnoli & Tonelli, 2017), which is mainly influenced by research areas and research methods (Morgeson, Mitchell, & Liu, 2015). However, the concept of event is also affected by the source of the data that records the event. For example, event was initially studied in broadcast news, thus it was mainly focused on the textual news document streams (e.g., newswire, news broadcast transcripts) (Yang, Pierce, & Carbonell, 1998). Recently, event has drawn notably attention of researchers since social media systems (e.g., blogs, Facebook, Twitter) have been the valuable sources for information about media events (Dong, Liang, & He, 2017). Although all studies in the literature claimed that their research were about "real-life events" (Ritter, Mausam, Etzioni, & Clark, 2012), currently, the "events" that are being studied can be summarized as follows:

- *Social event*: Social event refers to the activity performed by a group of people. Research on social events has gain much attention of researchers (Atefeh & Khreich, 2015). In general, the social event is conventionally represented by a number of keywords showing burst in appearance count (Paul, Peng, & Li, 2019; Yang et al., 1998), such as earth quake and election. Thus, such an event detection problem is closely related to that of topic detection and tracking (TDT) (Aggarwal & Subbian, 2012). Even TDT goes popular recently because it is suitable to perform tasks of social event detection on documents (Sprugnoli & Tonelli, 2017). Interestingly, Hu, Wang, Peng, Liang, and Du (2017) argued that changes in social networks, such as the addition of nodes and the reconnection of links, can also be seen as a series of events.
- *Individual event*: Individual event refers to the activity performed by an individual. Since the content published by users on online social media is often referred to as a User-Generated-Content (Moens, Li, & Chua, 2014), as a result, researchers are also starting to study "user events", such as marriage and traveling events (Ritter et al., 2012). These work focus on identifying "behavioral events" from online text (Li, Ritter et al., 2014; Ritter et al., 2012) and videos (Ke, Sukthankar, & Hebert, 2007; Tang, Li, & Koller, 2012). de Lira et al. (2019) investigated the novel problem of exploiting the content of posts on social media to infer the users' attendance of large events.

- *Business/organizational event*: Business event refers to the activity performed by an organization (or a firm). Recently, Morgeson et al. (2015) developed an organizationally focused event system theory to explain when and how events affect the behavior and features of organizational entities and trigger subsequent events. It is worth to note that, business event detection involved developing a list of events that could affect the ability of the enterprise to meet its strategic and operating objectives. Therefore, many scholars have used numerous terms to describe events (Morgeson et al., 2015), such as critical incidents (Flanagan, 1954), shocks (Lee, 1994), emergencies (Imran, Castillo, Diaz, & Vieweg, 2015), and risk management (O'Donnell, 2005).

We notice that little progress has been made towards the problem of automatically extracting business events from massive online documents. This study attempts to fill the research gap.

2.2. Event detection methods

As we can see, most existing event in social media were domain-related, such as disease outbreaks, civil unrest, and financial crises (Chen & Neill, 2014), thus the detection tasks in literature included identifying bursty public topics and topic evolution by observing the changes of word frequency (Becker, Naaman, & Gravano, 2011; Diao, Jiang, Zhu, & Lim, 2012). Along this line, two important research streams are focused on the technologies of event representation and event extraction.

To detect event information from various dataset, generally, unsupervised or heuristic methods are introduced in the research to obtain a set of verbal terms (phrases) as the representation of an event (Li, Zhu, & Zhou, 2014; Ritter et al., 2012). Jacobs, Lefever, and Hoste (2018) identified 10 types of company-specified economic events based on a subsample of random articles. To detect event-specific tweets that are likely to be beneficial for emergency response, Laylavi, Rajabifard, and Kalantari (2017) introduced a novel method, in which a sample dataset of tweets is manually labeled by three experts to obtain the ground truth of the event-related tweets.

There were mainly two types of methods for event extraction: the unsupervised and supervised approaches (Atefeh & Khreich, 2015). Most of the unsupervised methods are clustering based (Aggarwal & Subbian, 2012; Huang et al., 2016; Kuo & Chen, 2007), which attempt to find latent events by uncovering common patterns of texts that appear in the document set. For example, Hasan, Orgun, and Schwitter (2019) proposed an event detection system that incorporates specialized inverted indices and an incremental clustering approach to provide a low computational cost solution to detect newsworthy events from the Twitter data stream. Further, some approaches were associated with topic model to extract more detailed event information (Keane, Yee, & Zhou, 2015; Wei, Joseph, Lo, & Carley, 2015), such as event triggers (Li, Zhu et al., 2014; Ritter et al., 2012; Wei & Hachey, 2015), and some were occasionally associated with probabilistic model (Zhou, Gao, & He, 2016). The supervised approaches were mostly focused on text features in order to obtain more precision results of event detection (Tokarchuk, Wang, & Poslad, 2017). These efforts generally fall into two distinct types of approaches: feature based methods (with rich hand-designed feature sets) (Lefever & Hoste, 2016; Wei & Hachey, 2015) and neural networks based method (Nguyen & Grishman, 2016).

3. System overview

In this section, we first present the definition of the research problem and then show an overview of the proposed framework for event detection.

3.1. The problem definition

Assume that there are n types of business event

$$E = \{\boldsymbol{e}_1, \ldots, \boldsymbol{e}_i, \ldots, \boldsymbol{e}_n\}, \tag{1}$$

which may be recorded in various forms of online text.

In general, a business event was characterized by some so called event triggers. For example, in the text "Amazon launches restaurant delivery on prime now in Austin, the word "launches is a trigger for the event "restaurant delivery which can be categorized into a business event of *expand new business*. As we can see, an event trigger is often a single verb or nominalization (Li, Nguyen, Cao, & Grishman, 2015). Moreover, it is the main word which most clearly expresses an event occurrence in documents (Ji & Grishman, 2008). Without losing generality, we assume that each type of event, i.e., $\boldsymbol{e}_i$ can be represented by m_i triggers (Li, Zhu et al., 2014; Ritter et al., 2012):

$$\boldsymbol{e}_i = \left(v_{i1}, \ldots, v_{ij}, \ldots, v_{im_i}\right). \tag{2}$$

Given a set of online documents, $D = \{d_1, \ldots, d_i, \ldots d_{|D|}\}$, where $|D|$ denotes the total number of documents in D, the problem of event detection is equal to two subtasks as follows:

- How to explore the event types and their corresponding triggers from a set of given documents, and
- How to identify the events recorded in an online document.

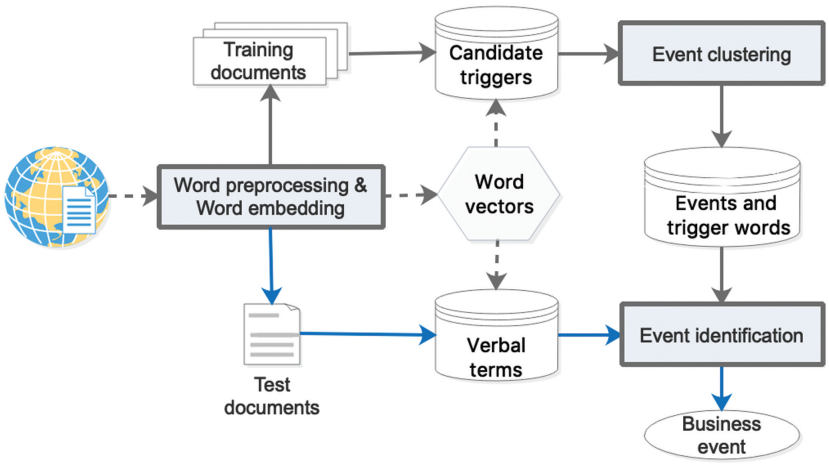


Fig. 1. The method framework.

The first subtask is about how to obtain the exact formation of E, and the second is about how to assign an appropriate event label to an unknown document d based on the structure of E. In this paper, we cast the first subtasks as a clustering problem in the learning process, and the second as a classification problem in the inferring process.

3.2. The pipeline architecture

An overview of our system for extracting business events from online news is presented in Fig. 1. The research framework consists mainly of three parts: *word preprocessing, event clustering and annotation*, and *event identification*. First of all, the *word preprocessing* module mainly performs necessary processing on the data, such as crawling data online, data cleaning, word segmentation and word vector embedding. In particular, in the *word embedding* step, we train a neural network to obtained high-quality vector representations for the words in corpus. The *event clustering and annotation* module obtains a list of business event types. The most important thing at this step is to annotate some major categories of business events from large quantities of unlabeled data. Finally, the *event identification* module is proposed to identify the event information from the test documents by making use of the learned event triggers as a classifier.

4. Business event clustering and annotation

In this section, we focus on identifying major categories of business events by leveraging large amount of unlabeled data. Here, the event clustering method is based on the representation of word vector, we fist introduce some necessary process for word preprocessing.

4.1. Word preprocessing and word embedding

The main tasks of *word preprocessing* include noise contents removing,[1] word segmentation, POS (part-of-speech) tagging, and word embedding. The first three tasks are routine data preprocessing in NLP, and the word embedding starts with word segmentation. After process of word segmentation, document $d_i \in D$ is then transformed into a word sequence as following:

$$d_i = \{w_{i1}, w_{i2}, ..., w_{ij}, ...\} \tag{3}$$

where w_{ij} means the jth words in document d_i. As a result, dataset

$$D = \bigcup_i \{d_i\} \tag{4}$$

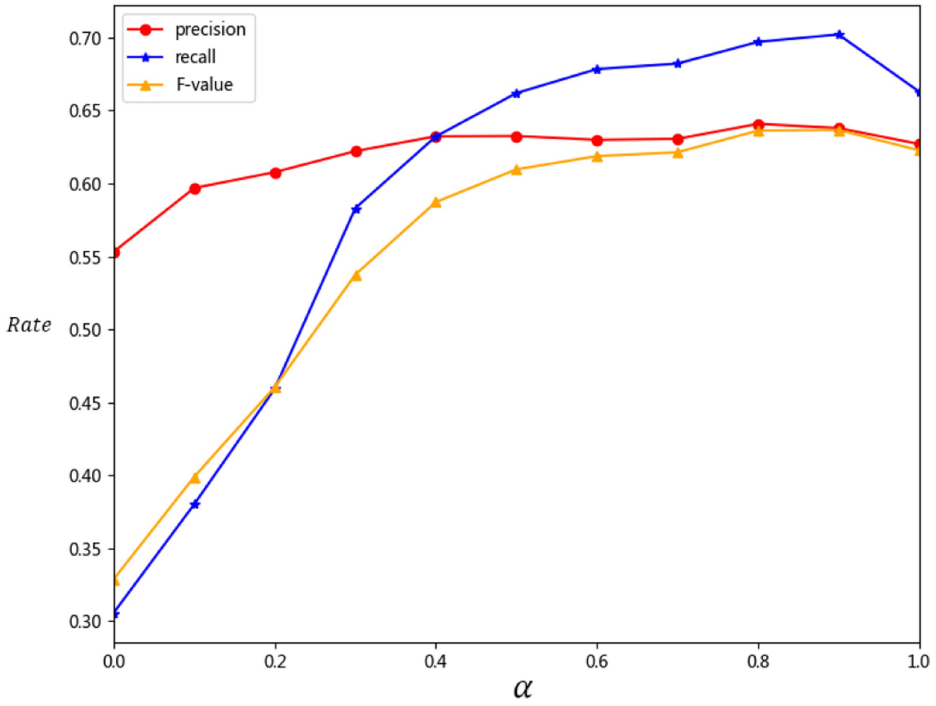
is formed. All the terms in D is denoted by Σ_D.

A word embedding,

$$WE: w_i \rightarrow \varpi_i \tag{5}$$

is a parameterized function that maps words $w_i \in \Sigma_D$ to a K-dimensional vectors $\varpi_i = (\omega_1, ..., \omega_K)$, where K is the dimension size of the embedded word vectors ranging from 50 to 500 (Mikolov, Chen, Corrado, & Dean, 2013). With the results of word embedding, we

[1] The data crawled online is not entirely related to the content of the article, there are often some noise, such as advertising contents and hyperlink text.



Algorithm 1. Exploring the information of event types from corpus.

can obtain a K-dimension matrix for all the words in D, which can be represented as follows:

$$WE: D \rightarrow \Omega^K, \tag{6}$$

where Ω is a K-dimension matrix consisting of $|\Sigma_D|$ rows. The ith row in Ω, i.e., $\boldsymbol{\varpi}_i$, means the new representation (vector) of word w_i.

4.2. Exploring the information of event types

The event detection can be treated as a separate problem of trigger tokens classifying, which involves how to categorize the extracted verbs into appropriate event types so that to obtain the formation of E. However, in a corpus, one of the major difficulties in effectively grouping trigger tokens is the problem caused by the ambiguity of potential trigger words (Liu, Liu, He, & Zhao, 2016). In addition, the occurrence frequency of these trigger tokens is unevenly distributed in corpus, and we believe that a few of the low-frequency verbs are meaningful for detecting business events. Therefore, how to make use of the low-frequency verbs in event detection is another problem.

To reduce the impact of noise terms (mixed with low-frequency verbs) on trigger classification, we first divide the verbs in D into two parts, i.e., one for high-frequency verbs and the other for low-frequency verbs. Then, we conduct the "clustering-annotation-classification strategy on the collection of high-frequency verbs. Finally, we compare the semantic similarity between each infrequent verb and each expert-labeled event, and then assign the low-frequency word to a most appropriate event class as a low-frequency trigger.

The flowchart of the proposed "clustering-annotation-classification strategy is given in Algorithm 1. The processes are specified as follows:

- *Process of clustering high-frequency terms*: firstly, the verb terms are extracted from D to a new dataset of D_v; Then, terms in D_v are sorted according to their frequency of occurrence in D, and the high frequency verbs are filtered out into a new set of D_v^H. At last, all the terms in D_v^H are clustered into k^* groups as $C_0 = \{\boldsymbol{c}_1, ..., \boldsymbol{c}_i, ..., \boldsymbol{c}_{k^*}\}$ (Line 4–6).
- *Process of human annotation*: one expert in business management inspects manually the terms in $\boldsymbol{c}_i (i = 1, ..., k^*)$ to infer the appropriate type of business event for it, such as "Investing, "Innovating, and "Merging.[2] At last, $n (n \leq k^*)$ valid clusters are selected as the initially detected events: $E = \{\boldsymbol{e}_1, \boldsymbol{e}_2, ..., \boldsymbol{e}_n\}$ (Line 7–12). Note that, the high-frequency terms with clear semantics in $\boldsymbol{c}_i$ can help expert to avoid bias in the annotation process. In addition, we allow a specific event $\boldsymbol{e}_i$ to belong to multiple business events. In the annotation process, the criteria for event $\boldsymbol{e}_i$ belonging to a certain business event depend on the judgment of the expert; however, in the process of event detection, the calculation of semantic similarity is the only criterion.

[2] Ambiguous or incoherent clusters are discarded in the annotation process.

- *Process of classifying low-frequency terms*: we first calculate the similarity between term $v_i \in (D_v - D_v^H)$ and term $v_j \in \bigcup_{\boldsymbol{e} \in E}\{\boldsymbol{e}\}$, and then adopt the KNN (k-nearest-neighbor) method to classify v_i into an appropriate high-frequency-verb cluster (Line 13–17).

The method presented in Algorithm 1 detects event types by grouping a set of terms with similar semantics into the same business event. Base on the vector representation of each word, the Cosine distance is a feasible method to measure the similarity between words (Mikolov et al., 2013). Given two words w_i, $w_j \in \Sigma_D$ and their word embedding results of *WE:* $w_i \rightarrow \boldsymbol{\varpi}_i \in \Omega^K$ and *WE:* $w_j \rightarrow \boldsymbol{\varpi}_j \in \Omega^K$, their Cosine similarity can be calculated as:

$$sim(w_i, w_j) = cosine(\boldsymbol{\varpi}_i, \boldsymbol{\varpi}_j) = \frac{\boldsymbol{\varpi}_i \cdot \boldsymbol{\varpi}_j}{\|\boldsymbol{\varpi}_i\| \|\boldsymbol{\varpi}_j\|}. \tag{7}$$

where, $\|\boldsymbol{\varpi}_i\|$ is the l_2-norm of the vector, and $\boldsymbol{\varpi}_i \cdot \boldsymbol{\varpi}_j$ is the dot product of the two vectors.

5. Business event identification

In this section, we propose a method to detect whether a given news article corresponds to one of the business event identified in Section 4.

5.1. Information fusion for verbs in document

Given an online document d, we are interested in what kind of event(s) are recorded in it. One rational way is to treat it as a classification problem by comparing the similarity between the verbs in d (denoted by V_d) and the trigger terms in $\boldsymbol{e}_i$ ($i = 1, ..., n$). However, there is generally more than one verb in V_d and $\boldsymbol{e}_i$ (i.e., $|V_d| \geq 1$ and $|\boldsymbol{e}_i| \geq 1$), thus the fusion of word vectors for both V_d and $\boldsymbol{e}_i$ must be done.

In addition, to explore the potential different effect between verbal terms in headlines and those in leads for event detection, we further split V_d into two parts as follows:

$$V_d = V_d^+ + V_d^-, \tag{8}$$

in which, $v \in V_d^+$ denotes the set of verbal terms appeared in article headlines, whereas, $v \in V_d^-$ account for the verbal terms in lead paragraphs. Accordingly, for all the terms in V_d^+, their mean vector (centroid) of $\omega_{V_d^+}$ is defined as:

$$\omega_{V_d^+} = \frac{\Sigma_{w_i \in V_d^+} \boldsymbol{\varpi}_i}{|V_d^+|}, \tag{9}$$

where $\boldsymbol{\varpi}_i$ is the word vector of w_i in Ω^K and Σ sums the vectors of all words in V_d^+ by dimension, $|V_d^+|$ denotes the number of words in V_d^+. Similarly, for all the terms in V_d^-, their mean vector (centroid) of $\omega_{V_d^-}$ can be defined as:

$$\omega_{V_d^-} = \frac{\Sigma_{w_j \in V_d^-} \boldsymbol{\varpi}_j}{|V_d^-|}, \tag{10}$$

where $\boldsymbol{\varpi}_j$ is the word vector of word w_j in Ω^K, $|V_d^-|$ denotes the number of words in V_d^-.

Finally, the vector of the centroid for all the verbs in V_d is calculated as:

$$\omega_{V_d} = \alpha\ \omega_{V_d^+} + (1 - \alpha)\omega_{V_d^-}, \tag{11}$$

where $\alpha \in [0, 1]$ is the parameter used to adjust the weight of $\omega_{V_d^+}$ and $\omega_{V_d^-}$. Without losing generality, we can expect that, when $\alpha = 1.0$, only the verbs in headlines play a role in event extraction. When $\alpha = 0$, the opposite is true. Moreover, when $\alpha = 0.5$, the verbs in headlines and the verbs in leads contribute equally to a task of event detection.

5.2. Information fusion for event triggers

Obviously, the trigger verbs in relationships (2) may not have equal contributions to event $\boldsymbol{e}_i$. We deal with this problem by assigning an appropriate weight for each trigger.

Assume the *term frequency* of verb v_{ij} is $TF_{v_{ij}}$, then the weight of v_{ij} is defined as:

$$\gamma_{v_{ij}} = \frac{TF_{v_{ij}}}{\Sigma_{v_{ij} \in \boldsymbol{e}_i} TF_{v_{ij}}}. \tag{12}$$

For the event of $\boldsymbol{e}_i = (v_{i1}, ...,v_{ij},...,v_{im_i})$, all the terms in it are weighted as $\boldsymbol{\gamma}_i = (\gamma_{v_{i1}},...,\gamma_{v_{ij}},...,\gamma_{v_{im_i}})$. Since term v_{ij} in $\boldsymbol{e}_i$ has been mapped on $\boldsymbol{\varpi}_{ij}$, then $\boldsymbol{e}_i$ can be represented in Ω^K as $\boldsymbol{\varpi}_{\boldsymbol{e}_i} = (\boldsymbol{\varpi}_{i1},...,\boldsymbol{\varpi}_{ij},...,\boldsymbol{\varpi}_{im_i})$. Finally, the centroid vector of $\boldsymbol{e}_i$ can be calculated as:

$$\omega_{\boldsymbol{e}_i} = \boldsymbol{\gamma}_i \boldsymbol{\varpi}_{\boldsymbol{e}_i}^T. \tag{13}$$

Input: Ω, d, sim_0 and α.
Output: Event type label for d.
Extract verbs $V_d = V_d^+ + V_d^-$ from d;
Extract $\Omega_{V_d^+}$ and $\Omega_{V_d^-}$ from Ω;
Calculate verbs centroid $\boldsymbol{\omega}_{V_d}$ for d with relation (11);
$i^* = NULL$;
for $i = 1$ to n **do**
    Calculate centroid $\boldsymbol{\omega}_{\boldsymbol{e}_i}$ for event $\boldsymbol{e}_i$ with relation (13);
    if $sim(\boldsymbol{\omega}_{V_d}, \boldsymbol{\omega}_{\boldsymbol{e}_i}) \geq sim_0$ **then**
        $i^* = i$;
        $sim_0 = sim(\boldsymbol{\omega}_{V_d}, \boldsymbol{\omega}_{\boldsymbol{e}_i})$;
    end if
end for
return i^*.

Algorithm 2. Event identification method.

5.3. Event identification algorithm

The event identification task tries to assign an appropriate event label for document d. To that end, we first compare the similarity between the centroids of V_d (i.e., $\boldsymbol{\omega}_{V_d}$) and $\boldsymbol{e}_i$ (i.e., $\boldsymbol{\omega}_{\boldsymbol{e}_i}$, $i = 1, \ldots, n$) by calculating the value of $sim(\boldsymbol{\omega}_{V_d}, \boldsymbol{\omega}_{\boldsymbol{e}_i})$. Then, we label document d with an optimal event type of $\boldsymbol{e}_{i^*}$, where i^* is resulted from

$$i^* = \arg \max_{i=1,\ldots,n} \{sim(\boldsymbol{\omega}_{V_d}, \boldsymbol{\omega}_{\boldsymbol{e}_i})\}, \tag{14}$$

which is subject to the condition $sim(\boldsymbol{\omega}_{V_d}, \boldsymbol{\omega}_{\boldsymbol{e}_i}) \geq sim_0$, and sim_0 is a predefined threshold. The event detection processes are illustrated as in following Algorithm 2.

6. Experimental results

In this section, we conduct a set of experiments over a real dataset crawled online to evaluate the performance of our method, namely the Word Representation based Clustering (denoted as WR_Clustering), in detecting business events from online text.

6.1. Dataset and evaluation metric

The data used in the experiments were all crawled from an online business news sharing website of http://www.investide.cn. Currently, the data on www.investide.cn has tracked across 7100 high-quality companies, and 900 investment institutions in China. The news articles published on it were previously classified into four category of business events by the editors, i.e., Financing Events (Finance), Merger and Acquisition Events (M&A), Initial Public Offerings Events (IPO), and Delisting Events (Exit). These categorizations can serve as a benchmark for the comparative experiments.

In the crawled dataset, documents duplicate contents, and documents without title or lead will be removed. Altogether, 14,556 documents were obtained, which were primarily published between 2011-10-01 and 2017-02-27. Some statistical characteristics of the crawled dataset are shown in Table 1. It can be seen that the distribution of the number of documents in different events is not balanced.

Consistent with the studies of event extraction in the NLP literature, we adopt the *Precision, Recall* and $F-value$ to evaluate the experimental results. The mathematical definitions of the three metrics are as follows (Han et al., 2018):

$$\begin{aligned} Precision &= \frac{\text{Number of correctly extracted events}}{\text{Number of all extracted events}} \\ Recall &= \frac{\text{Number of correctly extracted events}}{\text{Number of true events labeled by editor}} \end{aligned} \tag{15}$$

Table 1
Statistical characters of the dataset.

Category	From	To	# of documents	Total words
Financing	1988-1-1	2016-10-21	9,183	1,041,661
M&A	2000-1-1	2017-02-27	5,064	589,899
IPO	2006-09-24	2017-01-26	49	7,270
Delisting	2009-05-15	2016-12-26	260	24,186

and

$$F - value = \frac{2Precision \times Recall}{Precision + Recall}. \tag{16}$$

6.2. Data preprocessing and word vector training

In this work, we truncate the document text by the punctuations in the sentence and then adopt a Chinese NLP tool[3] to do the tasks of word segmentation and POS tagging simultaneously. Table 2 summarizes the main characteristics of the dataset we obtained after data preprocessing.

In total, there are approximately 1,368,034 terms in the dataset, of which the verbs are accounting for 286,376 (about 20.93%). Sorting all the words according to their rank of occurrence frequency in corpus, we will obtain the distribution of the top 200 frequent terms as shown in Fig. 2(a), and the distribution of the top 200 verbs as presented in Fig. 2(b).

As illustrated in Fig. 2, the word frequency is unbalanced distributed in the corpus, and showing a long-tail like distribution. This reminds us that if we use only the high-frequency buzzwords for event detection, we may lose some rare information represented by the low-frequency words. Fortunately, the strategy we proposed in Algorithm 1 can help us avoid this problem as much as possible.

Next, we embed all the data in the corpus into vectors by introducing the tool of Word2vec,[4] which is an efficient neural network implementation for learning distributed representations of words. The main parameters used in Word2vec are set as follows:

- Vector dimensionality = 100;
- The size of the context window = 5;
- Training algorithm: hierarchical softmax;
- Threshold for down sampling the frequent words = 5.

Finally, byinputting the word sequences (generated by word segmentation tool) into Word2vec, we could have the word vector representations for all the terms. The dimension of the word vector is $K = 100$.

6.3. Business event identification

The task of event identification in corpus is achieved by clustering verbs into groups according to their semantic similarity, annotating each group with an event label, and further using the grouped event triggers as a classifier.

6.3.1. Event trigger clustering

In order to obtain the initial set of triggers for business events, we first extract all the verbs into D_v from D. Then, we need to determine a suitable high-frequency verb set, i.e., D_v^H, which has a significant effect on the effectiveness of subsequent clustering experiments. To this end, we did 9 set of experiments to explore the impact of different set of D_v^H on the performance of WR_Clustering in detecting business event. In the first set of experiment, we introduce the top-100 high-frequency verbs as the elements of D_v^H; in the second set of experiment, the top-200 high-frequency verbs are introduced. Along this way, to the last set of experiment, the top-900 high-frequency verbs are introduced as D_v^H. Note that, in each set of experiment, we conduct five-fold cross validations on the dataset and report the average result. The results are shown in Fig. 3. The results in Fig. 3 indicate that WR_Clustering has an optimal average $F - value$ when D_v^H formed by top-500 high-frequency verbs. Therefore, in the subsequent experiments, we extract the top-500 verbs as D_v^H (the cumulative frequency of these words is corresponding to 17.7% of all the verbs used in the corpus) in WR_Clustering.

Next, the bisecting k-means algorithm is introduced to cluster the verbs in D_v^H into k groups. In the process of clustering, the cosine method is used to measure the similarity between verbs based on their word vectors trained by Word2vec. Assume that k clusters are initially obtained as $C_0 = \{\boldsymbol{c}_1, ..., \boldsymbol{c}_i, ..., \boldsymbol{c}_k\}$, and the mean vector of cluster $\boldsymbol{c}_i$ is $\boldsymbol{C}_{\boldsymbol{c}_i}$, then the Sum of Cosine Similarity (SCS) of $\boldsymbol{c}_i$ is calculated as:

$$SCS_i = \Sigma_j \cos(\boldsymbol{\varpi}_{ij}, \boldsymbol{C}_{\boldsymbol{c}_i}), \tag{17}$$

where $\boldsymbol{c}_i \in C_0$, $i = \{1, ..., k\}$ and $j = \{1, ..., |\boldsymbol{c}_i|\}$. The value of SCS_i is intuitively used to capture the coherence of a cluster, and the larger the SCS_i value is, the better the clustering results (Zhang, Zhai, & Han, 2013). To keep detail information as much as possible in the experiment, the cluster splitting process is stopped until the improvement of $\Sigma_{i=1}^{k} SCS_i$ is less than 0.5% by increasing the value of k.

6.3.2. Event type labeling and low-frequency terms classification

In this part of experiments, we try to explore the optimal clustering results and annotate each of these clusters with an appropriate event label based on the semantics of its triggers.

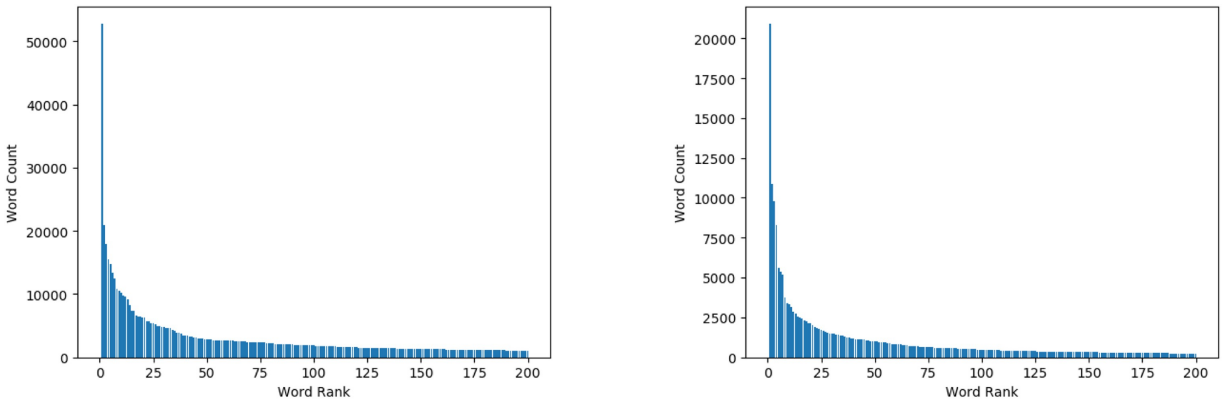
Out of the initially generated $k = 30$ clusters, the authors identify n=17 clusters as the detected events. At last, an additional type of OTHER is used to identify the exceptions (Ritter et al., 2012). Some sample "business events" as well as their triggers are displayed

[3] We use the Jieba (https://github.com/fxsjy/jieba) to do the initial NLP tasks.

[4] http://code.google.com/p/word2vec/.

Table 2
Statistical characters of the corpus.

Characters	Statistical description
Maximum length of text	1,504 words
Minimum length of text	2 words
Average length of text	93.98 words
Total words in corpus	1,368,034 words
Total verbs in corpus	286,376 words
Maximum number of verb in text	304 words
Minimum number of verb in text	0 word
Average number of verb in text	19.67 words



(a) All terms. (b) Verb terms.

Fig. 2. Frequency distribution of top-200 terms.

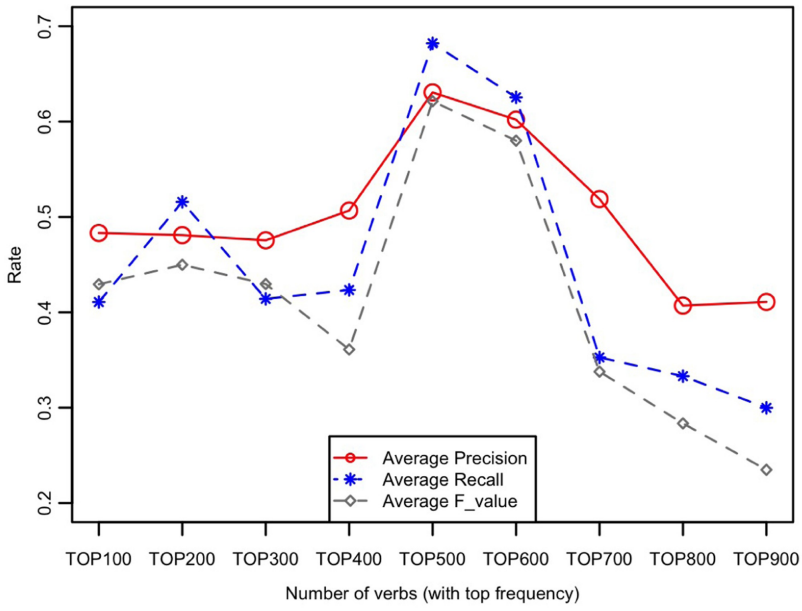


Fig. 3. Impacts of the volume of D_v^H on WR_Clustering.

Table 3
Some examples of events identified from the corpus (with the top-5 triggers).

e (Event type)	Example of the Top-5 trigger tokens
Innovation	变革(reform), 发展(develop), 迎来(receive), 创新(innovate), 具有(possess)
Business Upgrading	升级(upgrade), 整合(integrate), 综合(synthesis), 进军(advance), 加速(accelerate)
Investment & Financing	追加(append investment), 注资(capital injection), 获投(gain/get investment), 筹集(raise), 投资(invest), 融资(financing)
Delisting	达到(achieve), 下滑(descend), 增长(increase), 累计(accumulate), 亏损(loss)
Transformation	选择(choose), 变化(change), 带来改变(transform), 存在(exist), 解决(solute)
Mergers & Acquisitions	控股(holdings), 合并(merge), 分拆(split), 签署(sign), 谈判(negotiation)
Investigation	研究(research), 发布(publish), 调查(investigate), 统计(census), 报道(report)
Business expansion	加强(reinforce), 加快(accelerate), 拓展(launch), 推进(push), 扩张(expand)
Business Improvement	融合(fusion), 完善(improve and perfect), 优化(optimize), 强化(intensify), 转移(shift)
Conflict & Failure	拒绝(refuse), 攻击(attack), 取消(cancel), 失败(fail), 赔偿(compensate), 陷入(sink into)
Strategic planning	制定(formulate), 满足(fulfill), 需求(need), 吸引(attract), 采取(adopt)
Payment service	浏览(browse), 登录(log in), 保护(protect), 购买(purchase/buy), 付款(pay), 优惠(discount)
Legal event	诈骗(defraud), 投诉(complain), 违法(illegal), 欺诈(cheat), 传销(pyramid)
Marketing	创意(create), 瞄准(aim), 邀请(invite), 订购(place an order), 接触(contact)
Product design	连接(connect), 构建(construct), 设计(design), 定制(customize/personalize), 采用(adopt), 改版(revise)
IPO & Stock	招股(IPO), 认购(order to buy), 增持(overweight), 转让(transfer/assign), 套现(cash out)
Infringement	担心(worry), 侵权(tort), 引发(trigger), 质疑(doubt), 侵犯(invade)
OTHER	-

Table 4
Performance of the selected approaches on *Precision*.

Method	Delisting	Financing	IPO	M&A	Average
WR_clustering	0.5161	**0.9804**	**0.1563**	0.9104	**0.6408**
BOW + LR	**0.7143**	0.8801	0.0000	**0.9294**	0.6310
BOW + SVM	0.0000	0.8921	0.0000	0.9330	0.4563
BOW + Bayes	0.0710	0.8259	0.0045	0.8507	0.4380
Bigram + LR	0.0000	0.7654	0.0000	0.9155	0.4202
Bigram + SVM	0.0000	0.7626	0.0000	0.9210	0.4209
Bigram + Bayes	0.0223	0.9181	0.0040	0.8566	0.4503

Table 5
Performance of the selected approaches on *Recall*.

Method	Delisting	Financing	IPO	M&A	Average
WR_clustering	0.6154	0.7632	0.5556	**0.8538**	**0.6970**
BOW + LR	0.0893	0.9717	0.0000	0.8063	0.4668
BOW + SVM	0.0000	0.9733	0.0000	0.8389	0.4531
BOW + Bayes	**0.7857**	0.2845	0.6667	0.2816	0.5046
Bigram + LR	0.0000	0.9777	0.0000	0.5138	0.3729
Bigram + SVM	0.0000	**0.9787**	0.0000	0.5069	0.3714
Bigram + Bayes	0.0893	0.3422	**0.7778**	0.2184	0.3569

in Table 3 (only the top-5 event triggers are reported).[5]

In order to reduce the loss of rare information represented by low frequency words, we further calculate the similarity values between term $v_i \in (D_v - D_v^H)$ and all terms in $\bigcup_{c_j \in C_0}\{\boldsymbol{c}_j\}$, and introduce KNN method to classify v_i into a cluster if v_i has high similarity (equal or greater than a threshold) with the K terms in that cluster, or it will be abandoned.

In this way, we can effectively mine the events (as well as their triggers) information recorded in a set of online news articles.Event identification is always formalized as a multi-class classification problem (Nguyen & Grishman, 2016), as a result, the detected events as well as their triggers (See Table 3) can be used as an efficient classifier to identify the event information from online documents.

6.4. Performance evaluation

In this subsection, two feature representation methods combined with six popular document classification techniques, i.e., BOW + LR, BOW + SVM, BOW + Bayes, Bigram + LR, Bigram + SVM, and Bigram + Bayes, are introduced for the comparative purposes (Lai, Xu, Liu, & Zhao, 2015). Both of these methods can treat the problem of event detection as a classification task. Additionally, the criteria of feature selection for BOW are based on the TF-IDF values of the terms (Li, Sun, & Zhang, 2006).

We conduct a series of five-fold cross validations on the dataset for all the classification methods. Firstly, the dataset is broken into five non-overlapping and equal-sized (2,912 documents) subsets. Then, each of the classification method is trained with four of the subsets and tested with the fifth.

Consistent with the traditional method used in comparing the performance of data mining algorithms, we first focus on the accuracy of different methods in predicting events in the four categories of documents. The performances of all the selected methods on precision are reported in Table 4.

The experimental results of the comparison of precision show that the proposed method of WR_clustering has the best performance on detecting two type of business events, i.e., Finance and IPO, while the BOW + LR method has a very good performance in the detecting Delisting and M&A events. The overall results on all data show WR_clustering method has overcome other methods in terms of average precision.

Especialy, we see that WR_clustering method is far more effective than other methods in detecting IPO events, although there are only 48 news articles accounting for the IPO event. This indicates that WR_clustering have a good capability in detecting rare events (generally triggered by low-frequency terms) by overcoming the problem of biased data distribution.

The results of the comparison of *Recall* are reported in Table 5. It shows that, each of the selected method has their own advantages, and WR_clustering performs best in detecting the M&A events from the online documents. As a result, WR_clustering shows very good performance on the metric of average *Recall*. As we can see, the significant advantage of WR_clustering is very robust in detecting various events, i.e., both hot and unpopular events, from online documents. This makes it significantly different from the other six methods that can only perform well in detecting one or two types of business events. However, WR_clustering can't dominate the performance of *Recall* in all categories. One reason for it may be the distribution of verbal terms in headlines is different

[5] The English term for each verb is shown in brackets.

Table 6
Average performance of the selected approaches.

Method	Avg. Precision	Avg. Recall	Avg. F-value
WR_clustering	**0.6408**	**0.6970**	**0.6362**
BOW + LR	0.6310	0.4668	0.4865
BOW + SVM	0.4563	0.4531	0.4536
BOW + Bayes	0.4380	0.5046	0.2464
Bigram + LR	0.4202	0.3729	0.3792
Bigram + SVM	0.4209	0.3714	0.3778
Bigram + Bayes	0.4503	0.3569	0.2226

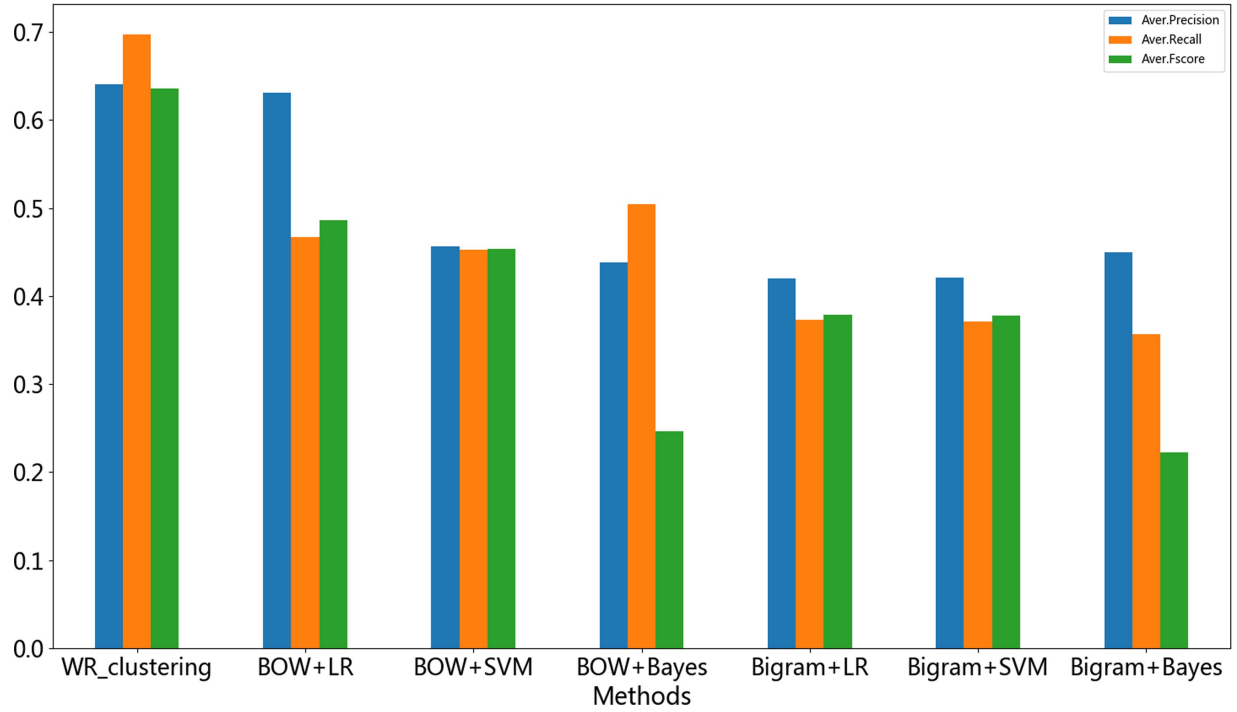


Fig. 4. The average performance of the methods.

from these in leads.

In summary, the averaged performance on *Precision, Recall* and $F-value$ are reported as the final results in Table 6. A comparative chart for the data is presented in Fig. 4. As we can see,the method proposed in this paper, i.e., WR_Clustering, has a best average performance for detecting business events from massive online news documents, which is followed by the BOW + LR. Especially, a series of methods based on LR and SVM perform very poorly for identifying the information of business events from an extremely biased dataset.

6.5. Effect of verbal terms in headline

As is discussed in Section 5.1, the verbal terms appeared in the headline of a news article are set as V_d^+, and the verbal terms appeared in the corresponding lead paragraph are set as V_d^-. According to relation (11), a parameter α is used to adjust the contribution of these two sets of verbs in event detection.

To answer the second research question, we further conduct a series of experiments to test the role of parameter α in detecting business event by ranging the value of α from 0 to 1.0. The regulation effects of α on the performance of WR_Clustering are shown in Fig. 5. The results show that the performance of WR_Clustering goes higher as the value of α increases until $\alpha \geq 0.8$, i.e., WR_Clustering performs best when α is about 0.8. Therefore, we set the value of α to 0.8 in the comparison experiments.

The results presented in Fig. 5 show some managerial implications addressed to the importance of headline and lead in detecting business event from online news articles. First of all, the performance of WR_Clustering is generally better at $\alpha \geq 0.5$ (compared to the cases of $\alpha < 0.5$), which means that the verbal terms in the headline of an article has a significant contribution in detecting event information. Further, as the value of α increases, the efficiency of WR_Clustering on *Prcision* ranges relative slightly from 0.55 ($\alpha = 0$, which means there is no verbal term from headline) to 0.61 ($\alpha = 0.8$). Simultaneously, we find that its efficiency on *Recall* increases significantly from 0.30 ($\alpha = 0$) to 0.68 ($\alpha = 0.8$). Such two results show that the verbal terms in the leads have provided a more stable

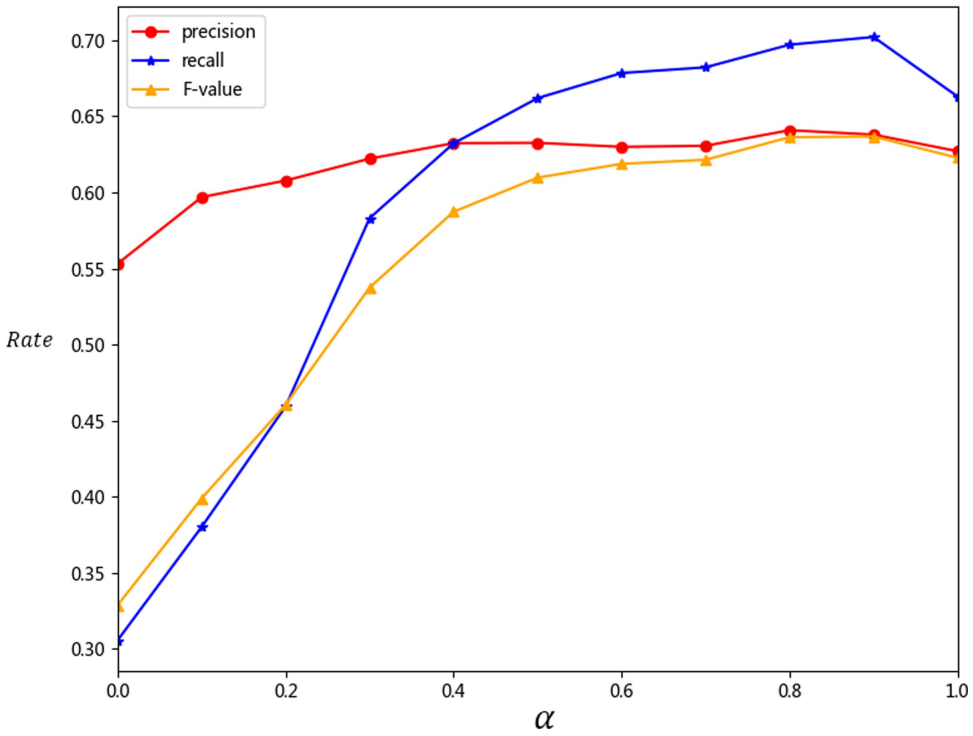


Fig. 5. Effect of α.

accuracy in identifying business events, indicating the lead of an online news plays a crucial role for understanding what event the article is about. On the contrary, the contribution of verbal terms in headlines are more diverse than that in leads, one reason for this is that there are some event-independent words introduced into the headlines of online news articles as "click bait" (Kuiken et al., 2017) to entice users to read the main content of the news. This also explains from another perspective, in an Internet news article, the headline plays a crucial role in drawing attention to the contents of online artefacts (e.g. news articles, videos and blogs) (Piotrkowicz, Dimitrova, & Markert, 2017), while the lead is more important for understanding what event(s) the article was about.

The results in Fig. 5 also show us some technical implications for event detecting. That is, in order to identify business events from a set of online news articles, if you rely solely on the triggers in headline, the precision of detecting result may be insufficient. Conversely, if you rely on all news content, the topics covered in the entire news may be too much and it causes confusion in understanding the detected events. Therefore, combining headline with lead contents is a viable option for the task of business event identification.

7. Conclusion

This research focuses on detecting business events from massive online news articles. To answer the research question of "How can we identify a business event from massive online textual sources efficiently?" we present a WR_Clustering method in this work to extract high-quality business event information. A series of experiments had been conducted on a real dataset crawled online to evaluate the performance of the proposed method. The results indicate that the proposed method is efficient in extracting high-quality event information from massive online documents. To answer the second research question of "What is the characteristic of the headline of an online news article in business event detection?" we collected the verbal terms in headlines and leads separately, and introduced a parameter α to adjust their contribution in the task of business event detection. The experimental results show that, the verbs in leads have provided a more stable accuracy in identifying business events, while the verbs in headlines contribute more from the perspective of increasing the values of *Recall* and $F-value$. This is partly because online news tends to use various words in headlines to attract readers.

7.1. Theoretical implications

We adopt a Skip-gram neural network architecture to train a word embedding model on large amounts of online news data. The subsequent learning and classification process is thus based on the result of word embedding, that is, a set of word vectors. Such a method can be used to deal with the challenge in detecting business events from massive noisy unstructured textual data and showcased its implementation in the context of online news documents. In comparison, most of the previous studies of event extracting have their intrinsic limitation of using BOW and Bigram models for feature representation. From a technical point of view,

our modeling approach provides at least a better alternative way of feature representation for extracting business information.

As mentioned earlier, the uneven distribution of different categories of data has always been a major obstacle to the efficiency of traditional classification methods. However, the occurrence of different types of business events is naturally uneven. The semi-supervised classification method proposed in this paper can greatly improve the efficiency of performing classification task on an unbalanced dataset. Such an approach, as shown in Table 6, performs particularly well in terms of *Recall* and $F-value$.

7.2. Practical implications

Online news articles exhibit valuable messages related to business events. In this paper, a business event is an incident or occurrence that emanates from online data sources during a particular interval of time.

From a practitioner perspective, on the basis of the proposed framework, people can effectively extract the events of firms from massive online data source, which is very useful for marketers. In particular, it enables marketers to keep track of business events of their competitors through competitors' online data such as news articles and blog texts. Along this line, the foremost application of our work is to monitor the behavior of firms through online documents. Moreover, our method is also meaningful to use online data to detect various business information, such as industrial trends, and to explore how the business events come to impact organizations across space and time.

7.3. Limitation and future work

To answer the second research question, we present only the experimental results of the different roles of verbal terms in headlines and leads for detecting business events. A worthy research direction in the future is to use a theoretical approach to measure the different roles of headline and lead in online news. Moreover, considering the formation of an online news the lead is not necessarily essential. Therefore, the comparative study of the readers' response in the different situations, i.e., with and without lead, must be very interesting, and the results of the study should have an impact on the publisher's strategy of information display in a limited area, for example, the screen of smart phone.

Acknowledgment

The work was partly supported by the National Natural Science Foundation of China (71572029/71671027/91846105/71490723/71772017), the Beijing Municipal Social Science Foundation (No. 17GLB009). Thanks also to the supports of Yunnan Science and Technology Fund (2017FA034, 2017DS012) and Kunming Key Laboratory of E-Business and Alternative Finance (KGF [2018]18).

References

Aggarwal, C. C., & Subbian, K. (2012). *Event detection in social streams. Proceedings of the 2012 SIAM international conference on data mining, Anaheim, CA, USA*624–635.
Atefeh, F., & Khreich, W. (2015). A survey of techniques for event detection in twitter. *Computational Intelligence, 31*(1), 132–164.
Becker, H., Naaman, M., & Gravano, L. (2011). *Beyond trending topics: Real-world event identification on twitter. Fifth international AAAI conference on weblogs and social media. Barcelona, Spain*438–441.
Chen, F., & Neill, D. B. (2014). *Non-parametric scan statistics for event detection and forecasting in heterogeneous social media graphs. Proceedings of the 20th ACM SIGKDD international conference on knowledge discovery and data mining.* New York, NY, USA: ACM1166–1175 KDD '14
Chen, Y., Conroy, N. J., & Rubin, V. L. (2015). *News in an online world: The need for an "automatic crap detector". Proceedings of the 78th ASIS&T annual meeting: Information science with impact: Research in and for the community*. Silver Springs, MD, USA: American Society for Information Science81:1–81:4 ASIST '15
Dai, Z., Taneja, H., & Huang, R. (2018). *Fine-grained structure-based news genre categorization. Proceedings of the workshop events and stories in the news 2018*61–67.
Diao, Q., Jiang, J., Zhu, F., & Lim, E.-P. (2012). *Finding bursty topics from microblogs. Proceedings of the 50th annual meeting of the association for computational linguistics*. ACL536–544.
Dong, T., Liang, C., & He, X. (2017). Social media and internet public events. *Telematics and Informatics, 34*(3), 726–739.
Flanagan, J. C. (1954). The critical incident technique. *Psychological Bulletin, 51*, 327â358.
Han, S., Hao, X., & Huang, H. (2018). An event-extraction approach for business analysis from online chinese news. *Electronic Commerce Research and Applications, 28*, 244–260.
Hasan, M., Orgun, M. A., & Schwitter, R. (2019). Real-time event detection from the twitter data stream using the twitternews+ framework. *Information Processing & Management, 56*(3), 1146–1165.
Hu, W., Wang, H., Peng, C., Liang, H., & Du, B. (2017). An event detection method for social networks based on link prediction. *Information Systems, 71*, 16–26.
Huang, L., Cassidy, T., Feng, X., Ji, H., Voss, C. R., Han, J., & Sil, A. (2016). *Liberal event extraction and event schema induction. Proceedings of the 54th annual meeting of the association for computational linguistics*258–268.
Imran, M., Castillo, C., Diaz, F., & Vieweg, S. (2015). Processing social media messages in mass emergency: A survey. *ACM Computing Surveys, 47*(4), 67:1–38.
Jacobs, G., Lefever, E., & Hoste, V. (2018). Economic event detection in company-specific news text. In U. Hahn, V. Hoste, & M.-F. Tsai (Eds.). *Proceedings of the first workshop on economics and natural language processing*. ACL.
Ji, H., & Grishman, R. (2008). *Refining event extraction through cross-document inference. Proceedings of ACL-08: Hlt.* Columbus, Ohio: ACL254–262.
Karimi, M., Jannach, D., & Jugovac, M. (2018). News recommender systems - survey and roads ahead. *Information Processing & Management, 54*(6), 1203–1227.
Ke, Y., Sukthankar, R., & Hebert, M. (2007). *Event detection in crowded videos. Ieee 11th international conference on computer vision*1–8.
Keane, N., Yee, C., & Zhou, L. (2015). *Using topic modeling and similarity thresholds to detect events. Proceedings of the 3rd workshop on EVENTS: Definition, detection, coreference, and representation*34–42.
Kuiken, J., Schuth, A., Spitters, M., & Marx, M. (2017). Effective headlines of newspaper articles in a digital environment. *Digital Journalism, 5*(10), 1300–1314.
Kuo, J.-J., & Chen, H.-H. (2007). Cross-document event clustering using knowledge mining from co-reference chains. *Information Processing & Management, 43*(2), 327–343.
Lai, S., Xu, L., Liu, K., & Zhao, J. (2015). *Recurrent convolutional neural networks for text classification. Proceedings of the twenty-ninth AAAI conference on artificial intelligence*2267–2273 AAAI'15

Laylavi, F., Rajabifard, A., & Kalantari, M. (2017). Event relatedness assessment of twitter messages for emergency response. *Information Processing & Management, 53*(1), 266–280.
Lee, T. R., & Mitchell, T. W. (1994). An alternative approach: The unfolding model of voluntary employee turnover. *Academy of Management Review, 19*, 51–89.
Lefever, E., & Hoste, V. (2016). *A classification-based approach to economic event detection in dutch news text. Proceedings of the tenth international conference on language resources and evaluation LREC 2016*330–335.
León, J. A. (1997). The effects of headlines and summaries on news comprehension and recall. *Reading and Writing, 9*(2), 85–106.
Li, J., Ritter, A., Cardie, C., & Hovy, E. (2014). *Major life event extraction from twitter based on congratulations/condolences speech acts. Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*. ACL1997–2007.
Li, J., Sun, M., & Zhang, X. (2006). *A comparison and semi-quantitative analysis of words and character-bigrams as features in chinese text categorization. Proceedings of the 21st international conference on computational linguistics and the 44th annual meeting of the association for computational linguistics*. ACL545–552.
Li, P.-F., Zhu, Q.-M., & Zhou, G.-D. (2014). Using compositional semantics and discourse consistency to improve chinese trigger identification. *Information Processing & Management, 50*(2), 399–415.
Li, X., Nguyen, T. H., Cao, K., & Grishman, R. (2015). *Improving event detection with abstract meaning representation. Proceedings of the first workshop on computing news storylines*. ACL11–15.
de Lira, V. M., Macdonald, C., Ounis, I., Perego, R., Renso, C., & Times, V. C. (2019). Event attendance classification in social media. *Information Processing & Management, 56*(3), 687–703.
Liu, H., Morstatter, F., Tang, J., & Zafarani, R. (2016). The good, the bad, and the ugly: Uncovering novel research opportunities in social media mining. *International Journal of Data Science and Analytics, 1*(3), 137–143.
Liu, S., Liu, K., He, S., & Zhao, J. (2016). *A probabilistic soft logic based approach to exploiting latent and global information in event classification. Proceedings of the thirtieth AAAI conference on artificial intelligence*2993–2999 AAAI'16
Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). *Efficient estimation of word representations in vector space. Proceeding of the international conference on learning representations workshop track, Arizona, USA*1301–3781 ICLR 2013
Moens, M.-F., Li, J., & Chua, T.-S. (2014). *Mining user generated content.* Chapman & Hall/CRC.
Morgeson, F. P., Mitchell, T. R., & Liu, D. (2015). Event system theory: An event-oriented approach to the organizational sciences. *Academy of Management Review, 40*(4), 515–537.
Nguyen, T. H., & Grishman, R. (2016). *Modeling skip-grams for event detection with convolutional neural networks. Proceedings of the 2016 conference on empirical methods in natural language processing*886–891.
Nir, R. (1993). A discourse analysis of news headlines. *Hebrew Linguistics, 37*, 23–31.
O'Donnell, E. (2005). Enterprise risk management: A systems-thinking framework for the event identification phase. *International Journal of Accounting Information Systems, 6*(3), 177–195.
Paul, D., Peng, Y., & Li, F. (2019). *Bursty event detection throughout histories. Proceedings of the 35th IEEE international conference on data engineering. Macau, China* ICDE 2019
Piotrkowicz, A., Dimitrova, V., & Markert, K. (2017). *Automatic extraction of news values from headline text. Proceedings of the student research workshop at the 15th conference of the european chapter of the association for computational linguistics*. Valencia, Spain: Association for Computational Linguistics64–74.
Ritter, A., Mausam, Etzioni, O., & Clark, S. (2012). *Open domain event extraction from twitter. Proceedings of the 18th ACM SIGKDD international conference on knowledge discovery and data mining*1104–1112 KDD '12
Spark, D., & Harris, G. (2011). *Practical newspaper reporting.* SAGE Publications.
Sprugnoli, R., & Tonelli, S. (2017). One, no one and one hundred thousand events: Defining and processing events in an inter-disciplinary perspective. *Natural Language Engineering, 23*(4), 485–506.
Stieglitz, S., Mirbabaie, M., Ross, B., & Neuberger, C. (2018). Social media analytics - challenges in topic discovery, data collection, and data preparation. *International Journal of Information Management, 39*, 156–168.
Tang, K. D., Li, F., & Koller, D. (2012). *Learning latent temporal structure for complex event detection. 2012 IEEE conference on computer vision and pattern recognition, Providence, RI, USA*1250–1257.
Tokarchuk, L., Wang, X., & Poslad, S. (2017). Piecing together the puzzle: Improving event content coverage for real-time sub-event detection using adaptive microblog crawling. *PLOS ONE, 12*, 1–18.
Wei, S. S. C., & Hachey, B. (2015). *A comparison and analysis of models for event trigger detection. Proceedings of the australasian language technology association workshop 2015, Parramatta, Australia*128–132.
Wei, W., Joseph, K., Lo, W., & Carley, K. M. (2015). *A bayesian graphical model to discover latent events from twitter. Proceedings of the ninth international conference on web and social media, ICWSM 2015*503–512.
Westerman, D., Spence, P. R., & Van Der Heide, B. (2014). Social media as information source: Recency of updates and credibility of information. *Journal of Computer-Mediated Communication, 19*(2), 171–183.
Yang, Y., Pierce, T., & Carbonell, J. (1998). *A study of retrospective and on-line event detection. Proceedings of the 21st annual international ACM SIGIR conference on research and development in information retrieval.* ACM28–36 SIGIR '98
Zhang, D., Zhai, C., & Han, J. (2013). Mitexcube: Microtextcluster cube for online analysis of text cells and its applications. *Statistical Analysis and Data Mining, 6*(3), 243–259.
Zhou, D., Gao, T., & He, Y. (2016). *Jointly event extraction and visualization on twitter via probabilistic modelling. Proceedings of the 54th annual meeting of the association for computational linguistics (volume 1: Long papers)*269–278.