Natural Language Engineering (2023), **29**, pp. 181–200
doi:10.1017/S1351324921000462

CAMBRIDGE UNIVERSITY PRESS

ARTICLE

Enhancement of Twitter event detection using news streams

Samaneh Karimi[1,2], Azadeh Shakery[1,3,*] and Rakesh M. Verma[2]

[1]School of Electrical and Computer Engineering, College of Engineering, University of Tehran, P.O. Box: 14395-515, Tehran, Iran, [2]Computer Science Department, University of Houston, Houston, TX 77204-5008, USA and [3]Institute for Research in Fundamental Sciences (IPM), P.O. Box 19395-5746, Tehran, Iran
*Corresponding author. E-mail: shakery@ut.ac.ir

(Received 8 May 2019; revised 21 November 2021; accepted 22 November 2021; first published online 24 January 2022)

Abstract

A new framework for improving event detection is proposed that employs joint information in news media content and social networks, such as Twitter, to leverage detailed coverage of news media and the timeliness of social media. Specifically, a short text clustering method is employed to detect events from tweets, then the language model representations of the detected events are expanded using another set of events obtained from news articles published simultaneously. The expanded representations of events are employed as a new initialization of the clustering method to run another iteration and consequently enhance the event detection results. The proposed framework is evaluated using two datasets: a tweet dataset with event labels and a news dataset containing news articles published during the same time interval as the tweets. Experimental results show that the proposed framework improves the event detection results in terms of $F1$ measure compared to the results obtained from tweets only.

Keywords: News event detection; Language model; Bipartite graph matching; Short text clustering

1. Introduction

Today, Twitter is known as a primary source of news information (Kwak *et al*. 2010) that is widely used by members of the public, organizations, and researchers for different aims. These aims range from getting informed of recent events to analyzing the trends of the events for business improvement and complementing news articles with rich and timely event-related information by journalists. Detecting events from social media, such as Twitter, is an important research area with potential benefits to several applications including responding to natural disasters (Verma *et al*. 2019), tracking epidemics, or following trending topics such as political elections (Farzindar and Inkpen 2020). Hence, we see increasing interest in analyzing tweets for event detection. A study by Pew research center[a] shows that 52% of American Twitter users employed it as a source of news in 2013 and this share increased to 63% in 2015 (Barthel *et al*. 2015). The popularity of Twitter news content stems from its unique characteristics:

- Timeliness: Most real-world events are narrated by Twitter users all over the world and propagated instantly on Twitter. The brevity of tweets due to Twitter's character length limit facilitates this process.

[a]http://www.pewresearch.org/.

- Unbiasedness: Tweets are generated by a huge number of ordinary users (not professional content producers), so Twitter content is less prone to be biased (by a particular party or group) in reporting events.
- Low cost: In Twitter, information on different events is freely available to all news consumers including ordinary people and journalists.

Despite these advantages of Twitter content over other types of online news content, there are a few weaknesses:

- Informality: The informal nature of a tweet's language reduces the accuracy of text mining methods that process its event-related content due to the increased amount of noise in the text.
- Missing details in each tweet: The short length of tweets restricts the details included in the body of each tweet describing an event.

In contrast to Twitter content, news articles published by news websites typically use more formal language and also provide more detailed reports of events. The formal language used in news websites' publications is more effectively processed by text mining methods. Furthermore, the larger size of news articles in comparison with tweets yields more complete descriptions with richer event-related vocabulary. Therefore, each of these two types of online content can benefit event detection methods differently. In this article, a framework for improving the event detection performance is proposed which takes advantage of both types of event-related content, that is news articles and tweets. Therefore, the detailed description of the events in a formal language along with the earliest event-related content would be available for the event detection task. In the proposed framework, tweets are processed by a basic event detection method, then the detected events are enhanced using the supplementary information available in news articles published during the same time interval as the tweets, and finally, the event detection process is repeated using the enhanced representations of events as a new initialization.

The use of tweets and news articles has been studied by some of the previous works for different aims such as user modeling for news recommender systems, analyzing the use of tweets as quotes in news articles, tweet recommendation for news articles, and connecting news articles to Twitter conversations (Abel *et al*. 2011a; Broersma and Graham 2013; Shi, Ifrim, and Hurley 2014; Krestel *et al*. 2015). However, the simultaneous use of both of these event-related contents to achieve higher performance in the event detection task is particularly targeted in this study.

The main aim of this framework is to benefit from both types of event-related online content in the event detection task (i.e., timely and generally-unbiased tweets along with detailed reports from news websites). This framework enhances the event detection methods that provide a distribution of words for each detected event. In this study, a short text clustering method called GSDMM, which is a collapsed Gibbs Sampling algorithm for the Dirichlet Multinomial Mixture model (Yin and Wang 2014), is employed for event detection. The proposed framework consists of the following steps: In the first step, the event detection method is applied to tweets and also news articles published during the same time interval as the tweets. After obtaining the two sets of events, the most similar event from news articles to each event from tweets is identified using a bipartite graph matching approach. In the next step, each event from tweets is expanded using its most similar event from news articles employing a language modeling approach. In the final step of the framework, the event detection method is reapplied on the expanded representations of the events.

Each of the steps mentioned above is explained in more detail in the next sections. Overall, the main contributions of this work can be summarized as follows:

https://doi.org/10.1017/S1351324921000462 Published online by Cambridge University Press

- Proposing a novel framework for having the advantages of user-generated content in the event detection task along with enriching the vocabulary of events by news news articles' contents.
- Proposing a graph-based approach for discovering the tweet–news relations.
- Proposing a new language model-based expansion method for enhancing the textual content of the detected events.
- Performing a set of detailed experiments to evaluate the performance of the proposed framework.

We evaluated the performance of the proposed framework through detailed experiments on a labeled dataset of tweets. The results indicate that the proposed framework improves the event detection method performance. We also evaluated the performance of our graph-based approach for discovering tweet–news relations and showed that the proposed approach outperforms the baseline method used for this aim. We also provided some examples from the dataset to illustrate how the proposed framework can help the employed event detection method find events more precisely.

The rest of this study is organized as follows. Section 2 reviews some related works. The proposed framework is described in Section 3. Section 4 contains the evaluation of the proposed framework and experimental results. Conclusions are given in Section 5.

2. Related work

There have been several definitions for *event* in the literature, especially TDT (Topic Detection and Tracking) research. In Allan *et al.* (2018), *event* is defined as "some unique thing that happens at some point in time." In another definition proposed in Yang *et al.* (1999), the spatial aspect is also considered in *event* definition along with the temporal aspect and is defined as "something (non-trivial) happening in a certain place at a certain time" and in comparison with *topic*, *events* are defined as "instances of topics, associated with certain actions."

In Kalyanam *et al.* (2016), an *event* is defined as "a conglomerate of information that encompasses all of the social media content related to a real-world news occurrence." This definition emphasizes the use of social media content in event detection that is also used in the present work. The event definition introduced in Liang, Caverlee, and Cao (2015), which is the most consistent with the present work, uses three constraints for a group of terms to represent an event: (1) semantic consistency, (2) simultaneity, and (3) the terms should refer to similar geographic locations. Considering the mentioned event definitions, in the present work, social network content that is generated at the same time and is semantically consistent is referred to as the textual representation of an event.

Event detection has been widely studied in previous works using different approaches along with Topic Detection and Tracking research (Atefeh and Khreich 2015). "TDT is a study of the organization and utilization of information flows based on topical events." (Xu *et al.* 2019) A major group of studies in this area are probabilistic approaches such as Probabilistic latent semantic analysis (PLSA) (Hofmann 2001) and Latent Dirichlet Allocation (LDA) (Blei *et al.* 2003) that have shown to be effective for topic detection (Chen *et al.* 2017). In Hoffman, Blei, and Bach (2010), the authors propose Online–LDA to detect topics in reviews. Incorporating the temporal information of textual data into topic models has been studied in another group of papers in this area. As an example, in a probabilistic approach proposed for event detection (Li *et al.* 2005), both content and time information of news articles are incorporated to detect retrospective news events. In Dubey *et al.* (2013), two main shortcomings of traditional topic models, for example LDA, are addressed by introducing a non-parametric topic modeling approach that (1) is not limited to a fixed prespecified number of topics and (2) allows the temporal variations of topics.

https://doi.org/10.1017/S1351324921000462 Published online by Cambridge University Press

Various categorizations have been proposed by research papers for event detection studies. One of the categorizations is based on the type of input data to the method. According to this categorization, the methods are classified as event detection methods on news articles' content (Filatova and Hatzivassiloglou 2003; Ahn 2006; Ji and Grishman 2008) or social networks content (Paltoglou 2016; Xie *et al*. 2016; Balalau, Castillo, and Sozio 2018; Rudra *et al*. 2018). The third category is devoted to methods that use both types of data in event and news-related tasks (Krestel *et al*. 2015; Lourentzou *et al*. 2015; Mele, Bahrainian, and Crestani 2019).

Since the main aim of this study is to enhance the event detection performance by utilizing the joint information in news media and Twitter, it belongs to the third group of studies. Hence, the main focus of this section is on reviewing the previous works of this category that addresses different tasks in events and news area.

2.1 Comparing Twitter and newswire as news sources

Two similar works to this study are Petrovic *et al*. (2013), Verma *et al*. (2019), which compare Twitter and newswire, as two main sources of events information, from different viewpoints. In these papers, the authors compare the events covered in Twitter and newswire and the timeliness of these two sources in news reporting. The type of events that are reported on Twitter before newswire is also investigated. The main similarity between these papers and our work is extracting events from two different sources published contemporaneously.

The main difference is that in Petrovic *et al*. (2013), Verma *et al*. (2019), the authors compare the event-related content on Twitter and newswire while in this work, the joint information in both resources is utilized in a framework to improve the event detection performance.

Two other differences between Petrovic *et al*. (2013), Verma *et al*. (2019) and the present work are (1) how the events are represented and (2) how the similarity between tweet events and newswire events is calculated. In Petrovic *et al*. (2013), tweets and newswire events are represented by the sum of all document vectors that correspond to the cluster (event). The similarity values between tweet events and newswire events are computed using cosine similarity between their vectors. Also, the nearest neighbor method is used to find the events (clusters) that are reported on both Twitter and newswire. In Verma *et al*. (2019), each pair of tweet and news article is represented by a set of features and for finding similar pairs, a classification model is learned using a training set with matching labels.

In McCreadie, Macdonald, and Ounis (2013), a user study is conducted on the use of news articles and user-generated content as two sources of event-related information in news retrieval and aggregating the two sources in the retrieval results to satisfy end-users.

Another study comparing news stream and tweets in reporting events is in Mele *et al*. (2019). The authors propose a discrete dynamic topic modeling and Hidden Markov Model for event detection. In their next step, the detected events are used for clustering news documents, and finally, the tweets' and news articles' timeliness in reporting events is analyzed.

Next, we review the relevant previous work on discovering the linkage between these two sources of online event-related content, that is tweets and news articles.

2.2 Tweet–news linkage discovery

One method that addresses the task of exploiting correlations between tweets and news is proposed in Guo *et al*. (2013). The authors propose a graph-based latent variable model to model text-to-text correlations. For this goal, they employ the named entities of news articles, as news specific features, hashtags of tweets, as tweet specific features, and the temporal information in both genres to exploit correlations between texts.

In Shi *et al*. (2014), the authors introduce a framework for connecting news articles to Twitter conversations. The method uses keyword similarity to separate tweets per article. Local cosine

https://doi.org/10.1017/S1351324921000462 Published online by Cambridge University Press

similarity, global cosine similarity, local frequency of hashtags, and global frequency of hashtags are the features extracted for each article–hashtag pair for classification. These features are used in a series of Weka classifiers including MultilayerPerceptron, regularised logistic regression, and K* to retrieve all the hashtags relevant to each article.

In the study, Abel *et al.* (2011b), the task of linking tweets with news articles is investigated for constructing user profiles. Here, two sets of strategies to find relevant news articles to each tweet are proposed: URL-based strategies and content-based strategies. In content-based strategies, the similarities between hashtag-based, entity-based, and bag-of-word-based representations of tweets and news articles are computed to discover their relationships.

Next, we discuss work on employing tweets and news articles for two news-related tasks including controversy detection and news recommendation.

2.3 Controversy detection in news and events

Another use of newswire in event-related tasks is detecting controversial events from Twitter. In Popescu and Pennacchiotti (2010), authors propose regression models that use a rich feature set from Twitter content and news articles' content. They show that the features they define and use from news and the Web are useful in events' controversy detection and mention that "coupling Twitter information with traditional media helps validate and explain social media reactions."

In Lourentzou *et al.* (2015), another example of using joint information in news articles and social media is proposed. Here, the controversial sentences in news articles are found by leveraging relevant comments on Twitter as well as comments on news websites to score the controversy of opinions about an issue mentioned in the news article. In this method, each sentence of news articles is used as a query and the relevant comments on Twitter and news websites are retrieved and investigated for controversy detection. This study employs tweets for finding controversial sentences in news articles, while in our paper, the news articles' content is employed for event detection enhancement on tweets.

2.4 News recommendation systems

News recommendation has benefited from Twitter content. In Abel *et al.* (2011a), a new user modeling framework is introduced in the context of a personalized news recommendation system. In this framework, hashtags, named entities, and topics mentioned in tweets, as well as their related news articles, are employed to construct hashtag-based, entity-based, and topic-based profiles for a user of news recommender systems. They show that semantic enrichment using Twitter content improves the quality of the generated user profiles.

In Krestel *et al.* (2015), the authors propose a method for recommending tweets for any given news article. For this, they use information retrieval, classification, and text similarity computation methods to find tweets similar for each news article, including language models, topic models, logistic regression, and boosting. In addition to tweets and news articles' textual content, a few Twitter-specific features such as publication time, length, and follower count are also used to find the most relevant tweets to each news article.

The use of Twitter content for recommendation systems is not restricted to news recommenders. In Chakraborty (2018), tweets are employed to supplement the recommendation process in a contextual point of interest recommendation system. "Contextual Point of Interest (POI) recommendation is of particular interest for travel and tourism, to suggest places to visit for a user traveling to a new city." In that paper, people's opinions about a POI on Twitter are used to enrich the document representation for POIs. Then, the enriched representations of POIs are used as documents in a retrieval model.

Now, we present the proposed methodology for event detection enhancement using news stream.

https://doi.org/10.1017/S1351324921000462 Published online by Cambridge University Press

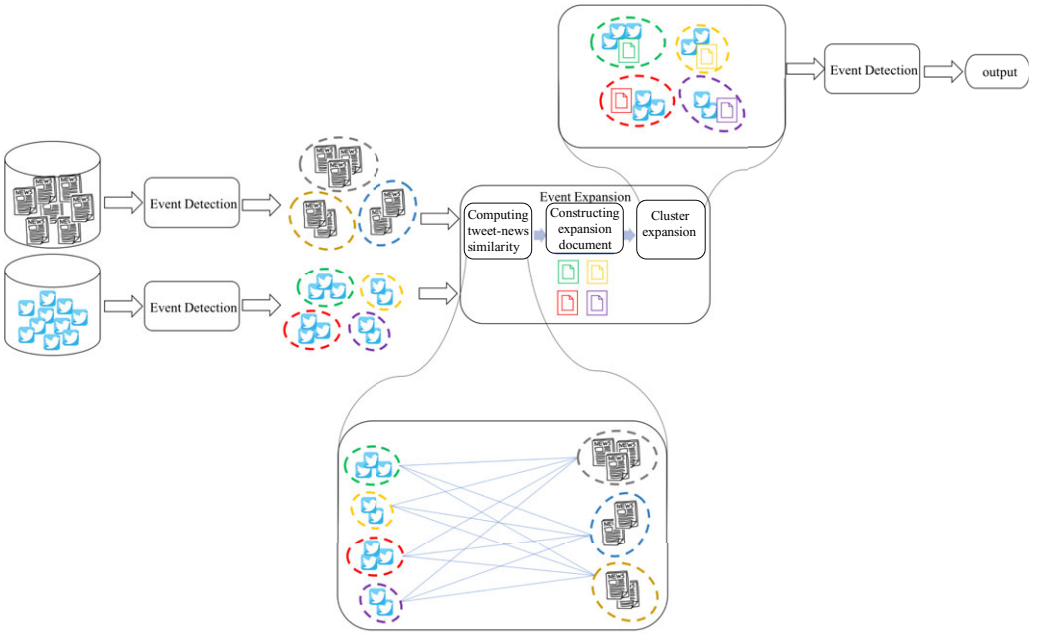

Figure 1. The illustrations of the proposed framework.

3. Proposed methodology

Event detection in Twitter has recently gained much attention among researchers. All challenges of processing user-generated content also apply to event detection from Twitter data. In this study, to tackle these challenges, an event detection improvement framework is proposed that involves the content available in relevant online news articles published by news websites in the task. The news articles' content is employed in an event expansion phase to enrich the events' representations by including more event-related words. Then, the event detection method is reapplied on the expanded versions of events' representations. Figure 1 shows the overall schema of the proposed framework. In the rest of this section, each step of the proposed framework is explained in detail.

3.1 Step 1: Event detection

First, we need to define the term event in our paper. There are various definitions for event in the literature (Yang *et al.* 1999; Chua, Razikin, and Goh 2011; Allan 2012; Kalyanam *et al.* 2016; Allan *et al.* 2018). According to Chua *et al.* (2011), "An event, in the context of social media, is an occurrence of interest in the real world which instigates a discussion on the event-associated topic by various users of social media, either soon after the occurrence or, sometimes, in anticipation of it."

Based on the framework presented in Chua *et al.* (2011), event detection approaches in the literature include techniques that are based on term interestingness, topic modeling, or incremental clustering. The clustering approach is employed to "group related tweets so that each group/cluster of tweets corresponds to a candidate event."

In our paper, we leverage the clustering approach, a source of event-related simultaneously published textual data, and language modeling to detect the group of words that represent events that occurred in the studied time frame. In the first step of the proposed framework, an event

https://doi.org/10.1017/S1351324921000462 Published online by Cambridge University Press

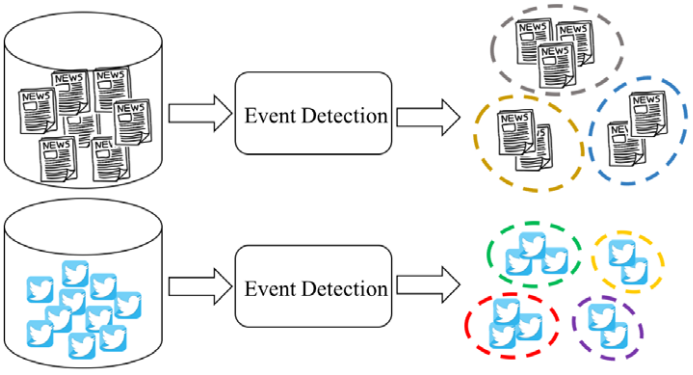

Figure 2. Event detection step.

detection method is applied on two sets of data: tweets and news articles that are both generated during the same time interval and thereby cover similar sets of real-world events (Figure 2).

In this study, a short text clustering method called GSDMM which is a collapsed Gibbs Sampling algorithm for the Dirichlet Multinomial Mixture model (Yin and Wang 2014) is employed as a tool for event detection. So, in the rest of this study, the terms "event" and "cluster" are used interchangeably.

The main idea of the method in Yin and Wang (2014) is to repeatedly compute the probability of document d being generated by cluster z based on the similarity between the document's content and the cluster's content through several iterations until the probabilities converge. With this idea, each document is assigned to a random cluster in the initialization step, then in each iteration, a cluster is re-assigned to each document d according to the following conditional probability distribution and the iterations are repeated until convergence. This probability is based on two rules: the completeness of the clusters which prioritize larger clusters (the first part of the formula) and homogeneity which assigns higher probabilities to clusters that have more similar content to the document in comparison with other clusters (the second part of the formula).

$$p(z_d = z|\vec{z}_{\neg d}, \vec{d}) \propto \frac{m_{z,\neg d} + \alpha}{D - 1 + K\alpha} \frac{\prod_{w \in d} \prod_{j=1}^{N_d^w} (n_{z,\neg d}^w + \beta + j - 1)}{\prod_{i=1}^{N_d} (n_{z,\neg d} + V\beta + i - 1)} \tag{1}$$

In this formula (Yin and Wang 2014), $\vec{z}_{\neg d}$ represents the vector of clusters excluding the cluster label of document d, $m_{z,\neg d}$ represents the number of documents in cluster z without considering document d, D is the total number of documents in the corpus, K is the number of clusters, α and β are Dirichlet priors for each of the two parts of the formulas, n_z is the number of words in cluster z, N_d is the number of words in document d, N_d^w is the frequency of word w in document d, $n_{z,\neg d}^w$ represents the frequency of word w in cluster z without considering document d, $n_{z,\neg d}$ is the number of words in cluster z without considering document d and $|V|$ is the vocabulary size. GSDMM is used as the event detection method in this framework since it has been proposed for short text clustering and also provides language model representations of events, that is clusters, so it comports with our language model based framework.

In the first step, the events are detected from tweets and news articles using the GSDMM method. In the proposed framework, the language model representations of events are needed for the following steps. According to Yin and Wang (2014), the probability of word w in each cluster z, $\phi_{z,w}$, can be computed using the following formula.

$$\phi_{z,w} = \frac{n_z^w + \beta}{\sum_{w=1}^{V} n_z^w + |V|\beta} \tag{2}$$

https://doi.org/10.1017/S1351324921000462 Published online by Cambridge University Press

where n_z^w represents the number of occurrences of word w in cluster z. Using this probability which also represents the importance of word w to cluster z, the language model of cluster z, θ_z, can be estimated as:

$$\theta_z{:}\{\phi_{z,w_i}\}_{i=1}^{|V|} \tag{3}$$

Therefore, at the end of the first step, the language models of all clusters of tweets, θ^T, and news articles, θ^N, are calculated.

$$\theta^T = \left\{\theta_z^T | 0 \leq z \leq k\right\} \tag{4}$$

$$\theta^N = \left\{\theta_z^N | 0 \leq z \leq k'\right\} \tag{5}$$

where θ_z^N represents the language model of cluster z from news articles that has k' clusters in total.

3.2 Step 2: Expansion data computation

A simple way to employ news articles for expansion is to use their textual content directly. However, due to different characteristics of tweets and news articles including their different sizes, using the news articles' content directly for expansion might not lead to good results. To tackle this challenge, we use the language model representations of the events detected from tweets and news articles in the expansion step. Therefore, in the second step of the framework, the events from tweets are expanded using their most similar events from news articles to include more relevant words to each event in their language model. The addition of more event-related words to the language model of each event modifies the words' distribution such that more event-related tweets would be inclined to their corresponding clusters and also non-relevant tweets would become less similar to their wrongly assigned clusters. As shown in Figure 3, this step consists of three consecutive sub-steps:

Tweet–News similarity computation: First, all clusters whether from tweets or news articles, are organized in a weighted undirected bipartite graph. In this graph, the nodes are clusters and the weight of each edge is the similarity value between its connecting clusters. The edge weight between cluster i from tweets part and cluster j from news articles part is computed using Jensen-Shannon (JS) divergence (Fuglede and Topsoe 2004) between the language models of cluster i, θ_i^T, and cluster j, θ_j^N, as follows:

$$D\left(\theta_i^T || \theta_j^N\right) = \sum_{w \in V} p\left(w|\theta_i^T\right) \log \frac{p\left(w|\theta_i^T\right)}{p\left(w|\theta_j^N\right)} \tag{6}$$

$$M = \frac{\theta_i^T + \theta_j^N}{2} \tag{7}$$

$$JS\_divergence\left(\theta_i^T, \theta_j^N\right) = \frac{D\left(\theta_i^T || M\right) + D\left(\theta_j^N || M\right)}{2} \tag{8}$$

$$JS\_SimilarityScore\left(\theta_i^T, \theta_j^N\right) = -JS\_divergence\left(\theta_i^T, \theta_j^N\right) \tag{9}$$

Next, the problem of finding the most similar news article cluster to each tweet cluster in the mentioned bipartite graph is solved as an assignment problem using the Hungarian method (Kuhn 1955). The reason for choosing this method is that in this problem, we need to match each news cluster to at most one tweet cluster (i.e. a matching is needed) and also make the matching

https://doi.org/10.1017/S1351324921000462 Published online by Cambridge University Press

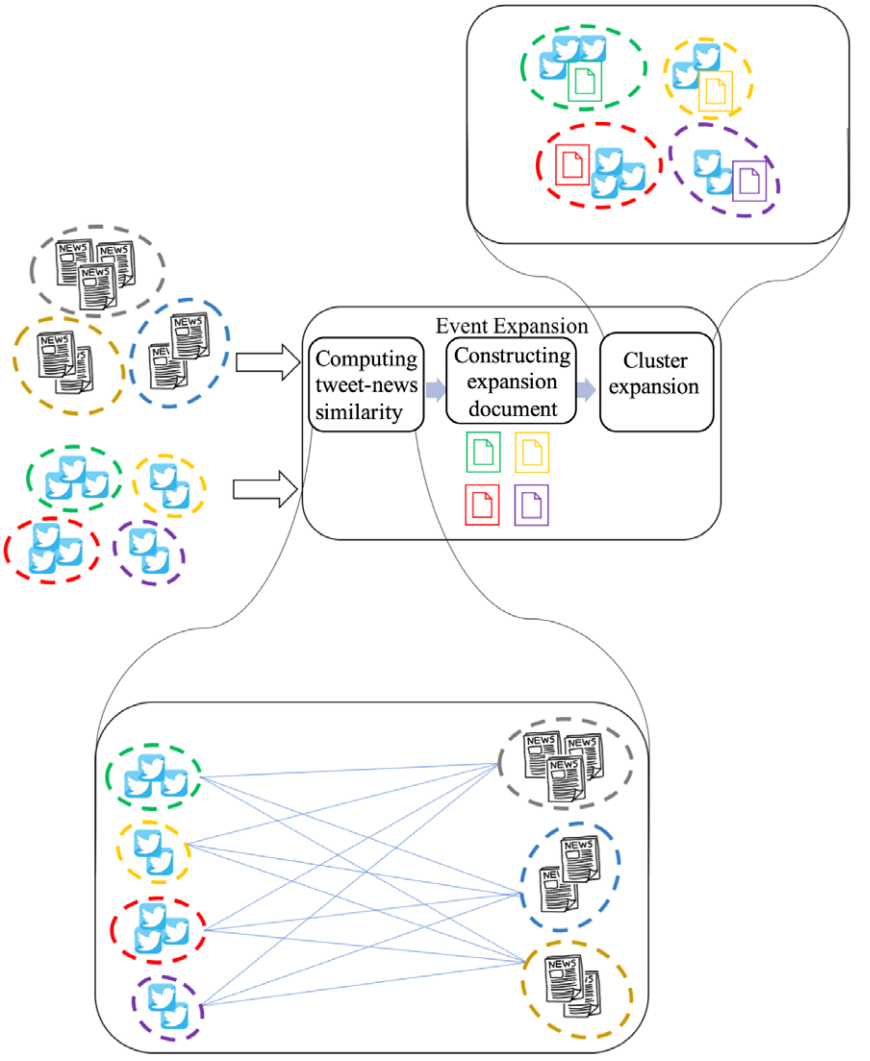

Figure 3. Event expansion step.

affected by the similarity values between clusters (i.e. the graph is weighted). Therefore, the maximum matching in a weighted bipartite graph is required which can be found using the Hungarian method.

In this method, the maximum weighted matching problem is mapped to finding a cover of minimum cost in the graph. Therefore, in this study, the weights of the bipartite graph that are JS similarity values are flipped to turn them into costs and also fit the problem setting. The input of the Hungarian algorithm is a square matrix of the assignment costs. The final assignments are found through a four-step procedure: (1) subtracting the minimum of each row from all row elements, (2) subtracting the minimum of each column from all column elements, (3) covering all zeros in the resulting matrix using a minimum number of horizontal and vertical lines, (4) if the minimum number of lines equals the number of rows in the matrix, an optimal assignment exists among the zeros of the matrix. Otherwise, the smallest element that is not covered by a line in the previous step should be subtracted from all uncovered elements and also added to all elements that are covered and the process is repeated from step 3.

https://doi.org/10.1017/S1351324921000462 Published online by Cambridge University Press

Expansion documents construction: After applying the Hungarian algorithm on the graph, the language model of the matched news cluster for each tweet cluster is employed to compute the expansion data for enriching the corresponding tweet cluster's language model. The idea of using the matched news cluster's language model as useful data for expanding each tweet cluster lies in its similar distribution of words to the tweet cluster's distribution, which facilitates finding relevant words suitable for constructing the expansion document. Let us assume the matched news cluster for tweet cluster i is $M(i)$ and we want to find the best p expansion words. We assume that the best word candidates for expanding a language model are the representative words chosen from its most similar language model. The words having higher probabilities in the language model are considered as the representative ones, due to their substantial role in representing the words' distribution. To this aim, top p words with the highest probabilities in the language model of the matched news cluster $M(i)$ are employed as expansion words $Exp\_Words(i)$ to be added to tweet cluster i as follows.

$$Exp\_Words\ (i) = \left\{ \left(w_t, P\left(w_t | \theta^N_{M(i)} \right) \right) | P\left(w_1 | \theta^N_{M(i)} \right) \right. \\ \left. \geq P\left(w_2 | \theta^N_{M(i)} \right) \geq .. \geq P\left(w_{|V|} | \theta^N_{M(i)} \right), t = 1, .., p\ , p < |V| \right\} \tag{10}$$

In our method, the expansion content is employed by the clustering method through building an artificial document called expansion document and adding it to its corresponding tweet cluster. In other words, the set of expansion words obtained for each tweet cluster is employed to build an expansion document that is added to its tweet cluster. To this aim, two issues about the expansion document should be addressed:

1. The frequency of expansion terms in the expansion document.
2. The number of expansion documents to be added to the corresponding tweet cluster.

Our approach to resolving the first issue is considering the expansion words with higher probabilities in the news cluster language model to be more important than other words and consequently, should be more frequent than other words in the artificially built expansion document. Therefore, the goal is to estimate the frequency of expansion words in the expansion document using their probabilities in the news cluster language model. To this aim, we employed the normalized probabilities of p expansion words in the news cluster's language model and the average documents' size in the target cluster, $Avg\_TSize(i)$, that is the average tweet size in the corresponding tweet cluster i. These two factors are used for computing the expansion words frequencies such that their relative frequencies would be proportional to their importance in the source news cluster language model while the expansion documents' length would be comparable to other documents in the target cluster. Therefore, the frequency of each expansion word w in the expansion document $Exp\_Doc(i)$ which expands the tweet cluster i is computed as follows.

$$Avg\_TSize(i) = avg_{t \in i} |t| \tag{11}$$

$$Term\_Freq\left(w, Exp\_Doc(i)\right) = \frac{P\left(w | \theta^N_{M(i)} \right)}{\sum_{w \in Exp\_Words(i)} P\left(w | \theta^N_{M(i)} \right)} \cdot Avg\_TSize(i), \qquad w \in Exp\_Words(i) \tag{12}$$

The number of expansion words, p is one of the method's parameters determined when tuning the parameters using the validation set. For the second issue, we heuristically make the number of expansion documents for each tweet cluster proportionate to the similarity between the corresponding tweet and news clusters such that more similar tweet clusters to the news cluster obtain more expansion documents. The idea is to have more expansion documents added to the tweet cluster when the tweet cluster and its matched news clusters are more similar compared to

https://doi.org/10.1017/S1351324921000462 Published online by Cambridge University Press

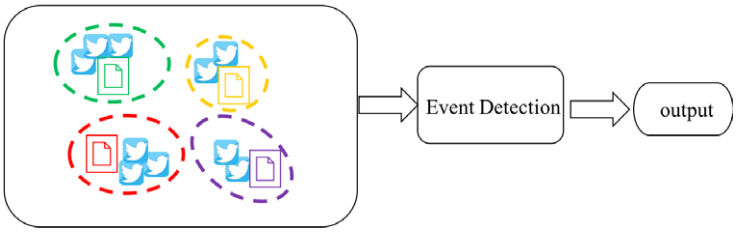

Figure 4. Enhanced event detection step.

other tweet–news cluster pairs. In other words, higher tweet–news clusters' similarity makes the obtained expansion content more reliable. To this aim, we used the JS similarity value between two corresponding tweet–news clusters, θ_i^T and $\theta_{M(i)}^N$, normalized by the sum of similarity values between the same news cluster and all tweet clusters, calculated in the denominator. The JS similarity values are calculated using Equation (9). We also used the average number of tweets in tweet clusters, $Avg\_TNum(i)$, to estimate the number of times that an expansion document $Exp\_Doc(i)$ should be added to the tweet cluster i according to the following formula.

$$Avg\_TNum(i) = avg_{k \in i} |\{t | t \in k\}| \tag{13}$$

$$Doc\_Freq\left(Exp\_Doc(i)\right) = \gamma \cdot \frac{JS\_SimilarityScore\left(\theta_i^T, \theta_{M(i)}^N\right)}{\sum_{j \in I} JS\_SimilarityScore\left(\theta_j^T, \theta_{M(i)}^N\right)} \cdot Avg\_TNum(I) \tag{14}$$

where I represents the set of tweet clusters, t represents tweet and γ is the parameter (with a value between 0 and 1) defined to control the proportion of expansion documents to existing tweets of the cluster in the expansion step.

Cluster expansion: After constructing the expansion documents for each tweet cluster, we add them to their corresponding tweet clusters by updating the clusters' statistics used by the GSDMM method (mentioned in Section 3.1) such as the number of documents in each cluster, the frequency of words in each cluster, the vocabulary size, and the total number of documents in the corpus. Using this expansion method, the assignments of all documents, including the original and expansion documents, to the clusters are recomputed using the new set of documents, which means the cluster-assignment probabilities based on formula (1) are recomputed considering the expanded set of documents. The impact of the original documents words on updating the cluster assignment probabilities is similar to the impact of expansion documents words, even if the words are common. In the next section, the enhanced event detection step is explained in more detail.

3.3 Step 3: Enhanced event detection

In the third step of the framework, the event detection method is applied to the expanded versions of tweet clusters computed in the previous step as shown in Figure 4. The rationale behind this step is to repeat the event detection method using an enhanced initialization of clusters rather than a random initialization to achieve better event detection results. In other words, the proposed framework through the mentioned steps tries to pause the event detection process in order to involve the content of another rich source of event-related information, that is news articles, to enhance the tweet clusters' content and then resume the event detection process with the updated clusters.

https://doi.org/10.1017/S1351324921000462 Published online by Cambridge University Press

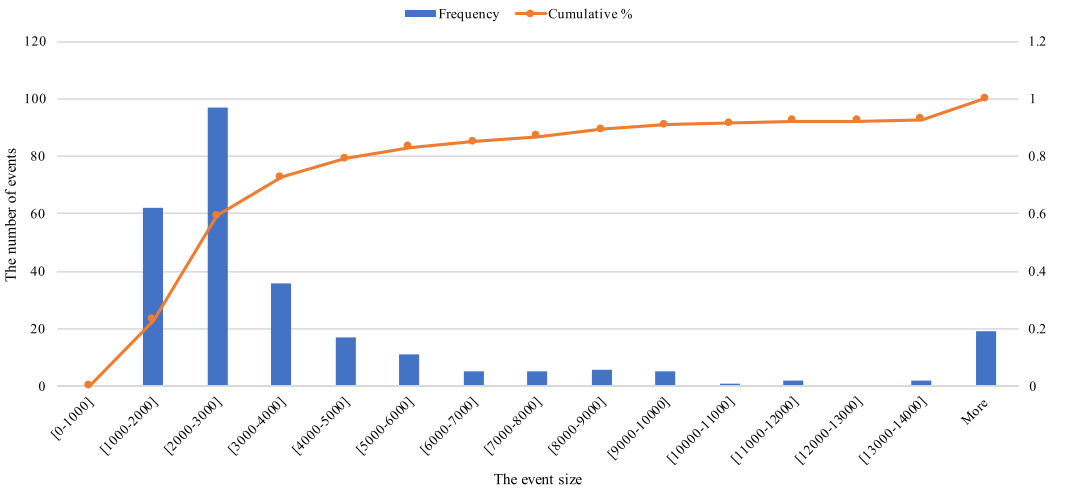

Figure 5. The histogram of event sizes which is the number of tweets having the same event label in ground truth tweet dataset.

4. Experimental results

4.1 Datasets

To evaluate the performance of our framework, we needed a dataset of tweets with event labels and a dataset of news articles that are published during the same time interval with tweets. To this aim, we utilized a subset of the tweet dataset with event labels introduced in Kalyanam *et al.* (2016) and also a subset of the breaking news dataset (Ramisa *et al.* 2018) that was published in the same time period as the tweet dataset. The used tweet dataset contains 1,369,876 tweets that account for 268 real-world events and contains tweets published between the 1st and the 13th of January 2014. The description of the data collection methodology is provided in Kalyanam *et al.* (2016). As a brief description, the authors of Kalyanam *et al.* (2016) used the news headlines reported by a list of well-known news media accounts on Twitter (e.g., @CNN, @BBCNews, etc.) and collected tweets written about news headlines that are collected periodically every hour over approximately 1 year using Twitter API.[b] The collected events are also validated in Kalyanam *et al.* (2016) to ensure that each group of tweets corresponds to a meaningful and cohesive news event. Figure 5 shows the histogram of the frequency distribution of events' sizes by bin. The size of an event is measured by the number of tweets having the same event label in the ground truth tweet dataset. The size of the bins is 1000. According to Figure 5, for 195 events (72.76% of the dataset) in the ground truth data, the event's size is less than 4000 tweets and less than 9% of events have sizes higher than 10,000 tweets. The minimum, maximum, and average event size in the ground truth data is 1019, 165,785, and 5111.47, respectively. Some more statistics about the tweet dataset are as follows.

- The total number of sentences in the tweet dataset is 1,808,052 across all tweets and the average is 1.31 per tweet.
- The tweet dataset contains 19,501,927 words of which 525,402 are unique.
- The average number of words per tweet is 14.23.

The news articles dataset contains 5473 news articles from several major newspapers and media agencies collected between the 1st and the 19th of January 2014. Therefore, the simultaneity

[b]https://dev.twitter.com.

https://doi.org/10.1017/S1351324921000462 Published online by Cambridge University Press

requirement, which is having both tweets and news articles published in the same time frame, is met. We have done stop-word removal on both datasets. Some more statistics about the news article dataset are as follows.

- The total number of sentences in the news article dataset is 276,192 across all documents and the average is 50.46 per document.
- The news article dataset contains 5,395,437 words of which 218,832 are unique.
- The average number of words per document is 985.82.

4.2 Evaluation metrics

The tweet dataset contains the event labels of the tweets so the performance of the proposed method can be evaluated using supervised metrics including precision (P), recall (R), and $F1$. To compute these measures, we need to calculate true positive (TP), false positive (FP), true negative (TN), and false negative (FN) values for each pair of data. For example, TP shows how many pairs of data (i.e. tweets) that are in the same class (i.e. have the same event label) in gold data, are predicted to be in the same cluster by the clustering method. The measures are computed according to the following formulas.

$$P = \frac{TP}{TP + FP}$$

$$R = \frac{TP}{TP + FN}$$

$$F1 = \frac{2PR}{P + R}$$

4.3 Baseline method

The baseline method used for comparison with the proposed method is an event detection method based on hashtags (Yang and Rayz 2017). In Yang and Rayz (2017), the feature extraction method proposed in Feng *et al.* (2015) and K-means clustering algorithm (Lloyd 1982) is employed for hashtag clustering. Each hashtag h_i is represented as a weighted vector by combining two vectors: a word vector representing the words of the tweets that contain the hashtag h_i and a hashtag vector representing other hashtags co-occurring with hashtag h_i. The co-occurring hashtags are the ones that appear in the same tweets. The K-means algorithm is used to detect hashtags' clusters. Finally, the hashtag clusters are used as their corresponding tweets' clusters.

In the following subsections, the experiments evaluating the performance of the proposed framework from different viewpoints are presented. First, the impacts of the two main parameters of the framework, k and γ, on the proposed framework's performance are investigated. Afterward, the impact of the enhancement method on event detection performance is evaluated and the event detection result is compared with the baseline. Finally, the impact of using the Hungarian method to find the most similar news article cluster to each tweet cluster on the performance compared to a baseline method is studied.

4.4 Parameter tuning

We randomly selected one-fifth of the tweet dataset and used it as the validation set for tuning the parameters of the proposed method including α, β, γ, p, and k. Using grid search, the optimal values of the parameters that lead to the highest $F1$ value on the validation set are $\alpha = 0.5$, $\beta = 0.1$, $\gamma = 0.17$, $p = 15$, and $k = 40$. These parameter values are used in all of the next experiments on the test set, that is the remaining four-fifth of the tweet dataset. Figure 6 shows the highest $F1$

https://doi.org/10.1017/S1351324921000462 Published online by Cambridge University Press

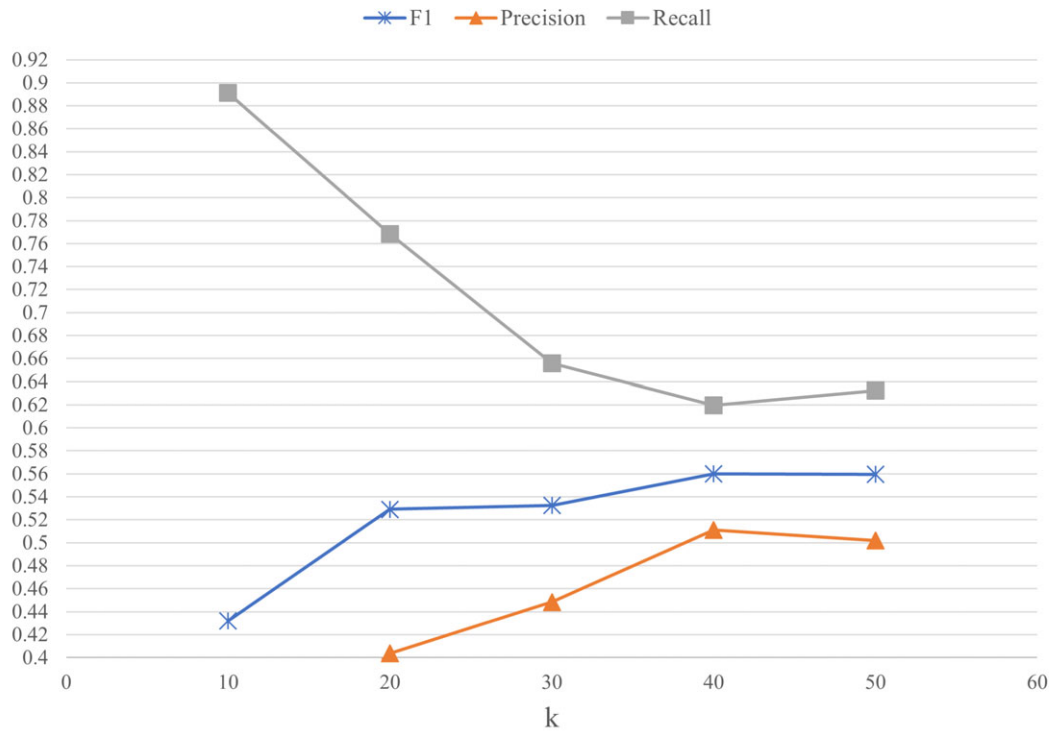

Figure 6. The impact of k on the proposed framework performance on the validation set.

obtained from the grid search for each value of k, that is the number of clusters, on the validation set. As shown in Figure 6, by increasing k from 10 to 40, $F1$ score rises until it reaches its maximum value at $k = 40$. The changes of *precision* by changing the values of k are similar to $F1$ while the *recall* scores at lower values of k are higher compared to larger k values. Higher *recall* values at low values of k is because of having many tweets in the same cluster due to the low number of clusters. In Yin and Wang (2014), the authors recommend setting k to larger values than the actual number of the clusters. To try this, we repeated the grid search using larger values of k from 200 to 300 increased by 10 and obtained the results. The results showed that the optimal value of k remains 40 and larger k values would not lead to higher performance of the clustering method.

4.5 Impact of γ

In this experiment, we investigate the influence of γ on $F1$ score of the proposed framework on the validation set. This parameter is defined to control the proportion of expansion documents to existing tweets in the cluster in the expansion step.

The experiment is done for six different values of γ ($\gamma \in \{0.01, 0.05, 0.09, 0.13, 0.17, 0.21\}$). Other parameters are set to the optimal values explained in Section 4.4. As shown in Figure 7, by increasing the value of γ from 0.01 to 0.17, the performance of the proposed method increases in terms of all three measures and reaches the maximum at $\gamma = 0.17$ and then all measures including *precision*, *recall*, and $F1$ decrease. This phenomenon can be explained considering how the event clusters change due to changes in γ value. For γ values smaller than the optimal value (0.17), the impact of the expansion is restricted while the clusters still have improvement capacity and can include more event-related words from news articles. So, by increasing γ from 0.01 to 0.17, the *precision*, *recall*, and $F1$ increase. By increasing γ more than 0.17, the event's distribution of

https://doi.org/10.1017/S1351324921000462 Published online by Cambridge University Press

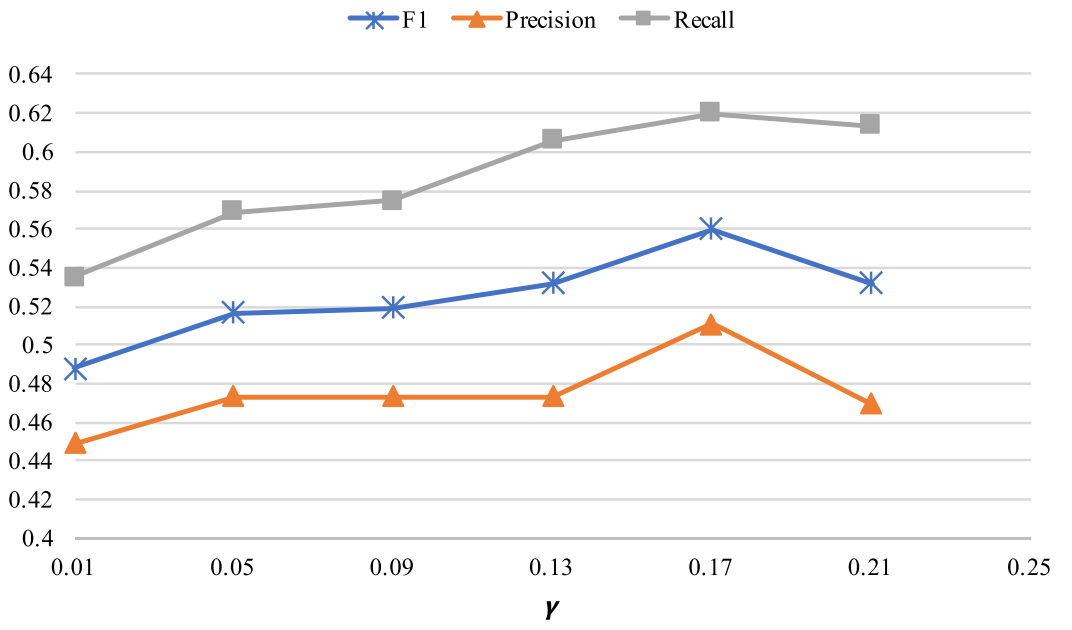

Figure 7. The impact of γ on the proposed framework performance.

words deviates from the ideal distribution for that event by adding extra expansion words so the performance decreases.

4.6 Event detection enhancement (step 3) evaluation

In this section, the performance of the proposed framework is evaluated. In the proposed framework, after finding the tweet clusters and news clusters, the expansion documents are built for each tweet cluster as explained in Section 3.2. The proper functioning of the expansion document construction, that is the second step, directly impacts the framework's performance. The improvement in the final results of the proposed framework shows that the second step performs well in finding suitable expansion words. We also investigated the content of the tweet clusters before the expansion as well as the expansion words found by the framework for those tweet clusters. As an example, the top 10 words with the highest weights in the language model of the tweet cluster number 4 without expansion are "Duggan", "Mark", "http", "police", "killed","verdict", "jury", "lawfully", "inquest" and "duggan". Using this set of words as a query and checking the top result retrieved by Google shows that no related document is retrieved that reports the particular event occurred in the time interval of our dataset (between the 1st and the 13th of January 2014). If we carry out the same Google search using the top 10 words followed by the corresponding expansion words, found by the framework, which are "Guardian", "home", "year", "said", "two", "years", "He", "would", "told", "She", "old", "family" as a query, the top retrieved result is a news article published by the Guardian on Jan 8, 2014, about the event.[c] To demonstrate the impact of expansion words in retrieving the relevant news article, we repeated the Google search using the expanded query excluding the "Guardian" term which can help retrieve documents from this news website. Again, we obtained the relevant news article about the event as the first retrieval result.

As another example, the top 10 words with the highest weights in the language model of the tweet cluster number 28 such as "Best", "Golden", "TV", "Series", "Globe", "wins", and "Drama"

[c]https://www.theguardian.com/uk-news/2014/jan/08/mark-duggan-verdict-live-coverage.

https://doi.org/10.1017/S1351324921000462 Published online by Cambridge University Press

Table 1. Example tweet clusters and their computed expansion words

Cluster number	Cluster top 10 words	Expansion words	Google News link
	Best, Golden, GoldenGlobes, http, Globes, Wins, Comedy, Globe, Red, carpet	2014, Street, Wall, Oscar, Wolf	https://www.independent.co.uk/arts-entertainment/films/news/golden-globes-2014-leonardo-dicaprio-wins-best-actor-for-the-wolf-of-wall-street–9055662.html, Date: Jan 13, 2014
7	http, Alex, Nick, Griffin, bankrupt, declared, Rodriguez, Yankees, Suspension,162	News, latest, said, London, Street, work, one, says, people, officer, police, Mr, denied, results	https://www.telegraph.co.uk/news/politics/bnp/10548708/BNPs-Nick-Griffin-declares-bankruptcy.html, Date: Jan 03, 2014
9	Cristiano, Ronaldo, Ballon, d'Or, FIFA, 2013, Congratulations, wins, winning, http	Year, French, round, World, team, man, Ireland, Cup, League, Ireland's, Republic, manager, O'Neill, Martin, double, jockey	https://www.theguardian.com/football/2014/jan/13/cristiano-ronaldo-ballon-dor-winner-real-madrid, Date: Jan 13, 2014
21	http, helicopter, explosion, crash, US, Palestinian, killed, bit.ly, Norfolk, fire	Year, US, said, night, woman, two, work, Paul, one, attack, Thousands, baby, volcano, Guatemala, Clinton	https://pilotonline.com/news/military/article_738fefd9-d777-53f5-9cf3-f9b4bfe74dcb.html, Date: Jan 9, 2014

does not clearly indicate the golden globe ceremony held in 2014, while its corresponding expansion words such as "years", "Slave", "12" that indicate the name of the movie that won the Golden Globe 2014 best motion picture drama award, enhances the clusters' content regarding this event.

Expansion document construction is one of the principal parts of the proposed framework and its proper functioning directly impacts the framework's performance. In Table 1, the content of a sample of four tweet clusters before the expansion as well as the expansion words found by the framework for those tweet clusters are reported to provide some more examples from the dataset in addition to the computed expansion terms for each of them. To represent the tweet cluster's content, we show the top 10 words with the highest weights in each cluster's language model. The expansion words for each tweet cluster, the links to the news webpages that report the corresponding event in addition to their publication dates are also shown in Table 1. The expansion words computed for each event provide more details reported about each event in news articles. For example, based on the top words in cluster number 0, such as "Best", "Golden", "Globes", "Wins", and "Comedy", its corresponding expansion terms include the name of the movie whose actor won the best actor in Comedy award in Golden Globes 2014.

In the next experiment, the event detection result on the tweet dataset, that is the results of the first step, is compared with the enhanced event detection results obtained from the proposed framework, that is the results of the third step, to evaluate the performance of the event detection enhancement framework. Furthermore, the performance of the proposed method is compared with another event detection method used as a baseline (Yang and Rayz 2017). Table 2 shows the results of the original event detection method, which is GSDMM in this paper, in comparison with the proposed enhanced event detection and the baseline method results on the test set.

According to Table 2, F1 value for the enhanced method on the test set achieves 47.13% and improves the *F*1, 2.03% compared to the original event detection method and 2.56% compared to the baseline method. It also shows that the proposed method's improvement is due to improving the recall measure from 51.62 to 55.10. Comparing the results of the baseline method with the

https://doi.org/10.1017/S1351324921000462 Published online by Cambridge University Press

Table 2. The performance of the event detection enhancement framework in comparison with event detection without enhancement and the baseline method (Yang and Rayz 2017) on test set

Method	Precision (%)	Recall (%)	*F*1 (%)	% Improvement on *F*1
Baseline	37.51	59.28	45.95	2.56
Original	41.79	51.62	46.19	2.03
Enhanced	41.18	55.10	47.13	

Table 3. The performance of the proposed framework using the Hungarian method compared to using Closest matching on test set

Matching method	Precision (%)	Recall (%)	*F*1 (%)
Hungarian	41.18	55.10	47.13
Closest matching	37.65	52.99	44.02

enhanced method shows that the precision has increased in the enhanced method. Hence, despite the higher recall value in the baseline method, the enhanced method outperforms the baseline in terms of *F*1.

4.7 Hungarian method performance in expansion evaluation

In this experiment, we try to study the performance of the Hungarian method in finding the most similar news article cluster to each tweet cluster. To this aim, we compare the results of the proposed framework using the Hungarian method with the results through which each tweet cluster is matched with its most similar news article cluster that can also be matched with other tweet clusters. We name the second run Closest Matching in Table 3. In both runs, the similarity between tweet clusters and news article clusters are computed using Jensen-Shannon (JS) divergence as explained in Section 3.2.

As shown in Table 3, the result using the Hungarian method outperforms the result obtained from matching each tweet cluster with its most similar news cluster. This can be because of several tweet clusters being matched with the same news cluster in the second run and consequently having several tweet clusters expanded similarly that do not occur in the Hungarian method. Moreover, news clusters with general content can cause more decrease in the performance of the second run rather than the first run (Hungarian method) since those clusters are similar to most tweet clusters and if they are the topmost similar clusters to tweet clusters, they can only be selected as the matched news cluster for one tweet cluster in the Hungarian method while they can be matched with several tweet clusters in Closest Matching run and cause more decrease in the performance.

The results of our experiments show that modifying the distribution of words to achieve better representations of events (i.e. better events' language models) is an effective approach to improve the event detection performance (according to Table 2). One of the key points in this approach is to find event-related words and limit irrelevant or less-relevant words when updating the event's language model. Our investigation into the results reveals that having some expansion words that are highly relevant to the event's topic is more effective than adding more expansion words that are moderately relevant to the event (according to Figure 7).

https://doi.org/10.1017/S1351324921000462 Published online by Cambridge University Press

5. Conclusions and future works

In this paper, we proposed an event detection enhancement framework that utilizes the joint information in news media content and Twitter content as user-generated content to improve the event detection performance. The proposed framework consists of three main steps. In the first step, events are detected from tweets using an event detection method. In the second step, detected events are expanded through finding and extracting event-related content from news streams published during the same time with tweets and based on a language modeling approach. In the third step, the expanded representations of events are used as a new initialization of the event detection method to run further iterations of the method and consequently improve the event detection results.

One of the challenges of event detection in Twitter is the short length of tweets. Although the tweet character length limit has increased from 140 to 280 characters, the tweets are still short and it restricts the content available in each tweet for processing. The proposed framework addresses this issue by introducing a framework for event detection enhancement that utilizes the event-related information available in news stream to enrich the events' language models. According to the experimental results, the proposed framework improves the event detection results and outperforms the baseline method.

Another achievement of this research is proposing a graph-based approach for discovering the tweet–news relations using the Hungarian algorithm. Furthermore, a new language model based expansion method for clusters is proposed in this study which can be employed in other research areas relevant to text clustering.

One of the characteristics of the proposed event detection enhancement framework is that it is not limited to a specific event detection method and any method that can provide a language model representation of events can be used in this framework to achieve improvement. Therefore, one of the future works for this research is employing other event detection methods such as topic modeling as the base event detection method and investigating the performance of the framework. Another future direction is to include the temporal information of tweets and news articles in the expansion process. Precisely speaking, the similarity between the publish time of tweets and news articles can be used in finding the event-related information and constructing better expansion documents.

In this study, tweets generated during 3 weeks have been used for experiments. Addressing the topic drift problem and considering models that allow for such possibility is another direction for future works. Furthermore, the validity of events reported on Twitter can also be studied as another interesting direction for future work.

Acknowledgments. Research of the second author was supported in part by a grant from the Institute for Research in Fundamental Sciences (no. CS1400-4-237). Research of the third author was supported in part by grants NSF DGE 1433817, ARO W911NF20-1-0254, and ONR N00014-19-S-F009. Verma is the founder of Everest Cyber Security and Analytics, Inc. The views and conclusions contained in this document are those of the authors and not of the sponsors.

References

Abel F., Gao Q., Houben G.-J. and Tao K. (2011a). Analyzing user modeling on twitter for personalized news recommendations. In *International Conference on User Modeling, Adaptation, and Personalization*. Springer, pp. 1–12.

Abel F., Gao Q., Houben G.-J. and Tao K. (2011b). Semantic enrichment of twitter posts for user profile construction on the social web. In *Extended Semantic Web Conference*. Springer, pp. 375–389.

Ahn D. (2006). The stages of event extraction. In *Proceedings of the Workshop on Annotating and Reasoning about Time and Events*, pp. 1–8.

Allan J. (2012). *Topic Detection and Tracking: Event-based Information Organization*, vol. **12**. NJ, United States: Springer Science & Business Media.

Allan J., Carbonell J.G., Doddington G., Yamron J. and Yang Y. (2018). Topic Detection and Tracking Pilot Study Final Report.

https://doi.org/10.1017/S1351324921000462 Published online by Cambridge University Press

Atefeh F. and Khreich W. (2015). A survey of techniques for event detection in twitter. *Computational Intelligence* **31**(1), 132–164.

Balalau O., Castillo C. and Sozio M. (2018). Evidense: A graph-based method for finding unique high-impact events with succinct keyword-based descriptions. In *Twelfth International AAAI Conference on Web and Social Media*.

Barthel M., Shearer E., Gottfried J. and Mitchell A. (2015). The evolving role of news on twitter and facebook. Pew Research Center, 14.

Blei D.M., Ng A.Y., Jordan M.I. and Lafferty J. (2003). Latent Dirichletallocation. *Journal of Machine Learning Research* **3**, 993–1022.

Broersma M. and Graham T. (2013). Twitter as a news source: How dutch and british newspapers used tweets in their news coverage, 2007–2011. *Journalism Practice* **7**(4), 446–464.

Chakraborty A. (2018). Enhanced contextual recommendation using social media data. In Proceedings of the 41st International ACM SIGIR Conference on Research & Development in Information Retrieval. ACM, pp. 1455–1455.

Chen L., Zhang H., Jose J., Yu H., Moshfeghi Y. and Triantafillou P. (2017). Topic detection and tracking on heterogeneous information. *Journal of Intelligent Information Systems*, **51**(1), 115–137.

Chua A.Y., Razikin K. and Goh D.H. (2011). Social tags as news event detectors. *Journal of Information Science* **37**(1), 3–18.

Dubey A., Hefny A., Williamson S. and Xing E.P. (2013). A nonparametric mixture model for topic modeling over time. In *Proceedings of the 2013 SIAM International Conference on Data Mining*. SIAM, pp. 530–538.

Farzindar A.A. and Inkpen D. (2020). Natural language processing for social media, third edition. *Synthesis Lectures on Human Language Technologies* **13**(2), 1–219.

Feng W., Zhang C., Zhang W., Han J., Wang J., Aggarwal C. and Huang J. (2015). Streamcube: Hierarchical spatio-temporal hashtag clustering for event exploration over the twitter stream. In *2015 IEEE 31st International Conference on Data Engineering*. IEEE, pp. 1561–1572.

Filatova E. and Hatzivassiloglou V. (2003). Domain -independent detection, extraction, and labeling of atomic events.

Fuglede B. and Topsoe F. (2004). Jensen-Shannon divergence and Hilbert space embedding. In *International Symposium on Information Theory, 2004. ISIT 2004. Proceedings.* IEEE, p. 31.

Guo W., Li H., Ji H. and Diab M. (2013). Linking tweets to news: A framework to enrich short text data in social media. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume* **1:** *Long Papers)*, vol. 1, pp. 239–249.

Hoffman M., Blei D. and Bach F. (2010). Online learning for latent dirichlet allocation. vol. **23**, pp. 856–864.

Hofmann T. (2001). Unsupervised learning by probabilistic latent semantic analysis. *Machine Learning* **42**(1), 177–196.

Ji H. and Grishman R. (2008). Refining event extraction through cross-document inference. In *Proceedings of ACL-08: HLT*, pp. 254–262.

Kalyanam J., Quezada M., Poblete B. and Lanckriet G. (2016). Prediction and characterization of high-activity events in social media triggered by real-world news. *PloS One* **11**(12), e0166694.

Krestel R., Werkmeister T., Wiradarma, T.P. and Kasneci G. (2015). Tweet-recommender: Finding relevant tweets for news articles. In *Proceedings of the 24th International Conference on World Wide Web*. ACM, pp. 53–54.

Kuhn H.W. (1955). The hungarian method for the assignment problem. *Naval Research Logistics Quarterly* **2**(1–2), 83–97.

Kwak H., Lee C., Park H. and Moon S. (2010). What is twitter, a social network or a news media? In *Proceedings of the 19th International Conference on World Wide Web*. ACM, pp. 591–600.

Li Z., Wang B., Li M. and Ma W.-Y. (2005). A probabilistic model for retrospective news event detection. In *Proceedings of the 28th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 106–113.

Liang Y., Caverlee J. and Cao C. (2015). A noise-filtering approach for spatio-temporal event detection in social media. In *Proceedings of the 37th European Conference on Information Retrieval*, pp. 233–244.

Lloyd S. (1982). Least squares quantization in PCM. *IEEE Transactions on Information Theory* **28**(2), 129–137.

Lourentzou I., Dyer G., Sharma A. and Zhai C. (2015). Hotspots of news articles: Joint mining of news text & social media to discover controversial points in news. In *2015 IEEE International Conference on Big Data (Big Data)*. IEEE, pp. 2948–2950.

McCreadie R., Macdonald C. and Ounis I. (2013). News vertical search: When and what to display to users. In *Proceedings of the 36th International ACM SIGIR Conference on Research and Development in Information Retrieval*. ACM, pp. 253–262.

Mele I., Bahrainian S.A. and Crestani F. (2019). Event mining and timeliness analysis from heterogeneous news streams. *Information Processing & Management* **56**(3), 969–993.

Paltoglou G. (2016). Sentiment-based event detection in twitter. *Journal of the Association for Information Science and Technology* **67**(7), 1576–1587.

Petrovic S., Osborne M., McCreadie R., Macdonald C., Ounis I. and Shrimpton L. (2013). Can twitter replace newswire for breaking news? In *Seventh International AAAI Conference on Weblogs and Social Media*.

Popescu A.-M. and Pennacchiotti M. (2010). Detecting controversial events from twitter. In *Proceedings of the 19th ACM International 'Conference on Information and Knowledge Management*. ACM, pp. 1873–1876.

Ramisa A., Yan F., Moreno-Noguer F. and Mikolajczyk K. (2018). Breakingnews: Article annotation by image and text processing. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **40**(5), 1072–1085.

https://doi.org/10.1017/S1351324921000462 Published online by Cambridge University Press

Rudra K., **Goyal P.**, **Ganguly N.**, **Mitra P. and Imran M.** (2018). Identifying sub-events and summarizing disaster-related information from microblogs. In *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval*. ACM, pp. 265–274.

Shi B., **Ifrim G. and Hurley N.** (2014). Be in the know: Connecting news articles to relevant twitter conversations. arXiv preprint arXiv:1405.3117.

Verma R., **Karimi S.**, **Lee D.**, **Gnawali O. and Shakery A.** (2019). Newswire versus social media for disaster response and recovery. In *2019 Resilience Week (RWS)*, vol. **1**. IEEE, pp. 132–141.

Xie W., **Zhu F.**, **Jiang J.**, **Lim E.-P. and Wang K.** (2016). Topicsketch: Real-time bursty topic detection from twitter. *IEEE Transactions on Knowledge and Data Engineering* **28**(8), 2216–2229.

Xu G., **Meng Y.**, **Chen Z.**, **Qiu X.**, **Wang C. and Yao H.** (2019). Research on topic detection and tracking for online news texts. *IEEE Access* **7**, 58407–58418.

Yang S.-F. and Rayz J.T. (2017). An event detection approach based on twitter hashtags. In *The 18th International Conference on Computational Linguistics and Intelligent Text Processing (CICLing 2017)*.

Yang Y., **Carbonell J.G.**, **Brown R.D.**, **Pierce T.**, **Archibald B.T. and Liu X.** (1999). Learning approaches for detecting and tracking news events. *IEEE Intelligent Systems and Their Applications* **14**(4), 32–43.

Yin J. and Wang J. (2014). A dirichlet multinomial mixture model-based approach for short text clustering. In *Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, pp. 233–242.

Cite this article: Karimi S, Shakery A and Verma RM (2023). Enhancement of Twitter event detection using news streams. *Natural Language Engineering* **29**, 181–200. https://doi.org/10.1017/S1351324921000462