Social Network Analysis and Mining (2022) 12:78
https://doi.org/10.1007/s13278-022-00911-x

ORIGINAL ARTICLE

Event prediction in social network through Twitter messages analysis

A. Yavari[1] · H. Hassanpour[1] · B. Rahimpour Cami[2] · M. Mahdavi[3]

Received: 25 March 2022 / Revised: 10 June 2022 / Accepted: 20 June 2022

Abstract
Event detection using social media analysis has attracted researchers' attention. Prediction of events especially in the management of social crises can be of particular significance. In this study, events are predicted through analyzing Twitter messages and examining the changes in the rate of Tweets in a specified subject. In the proposed method, the Tweets are initially preprocessed in consecutive fixed-length time windows. Tweets are then categorized using the non-negative matrix factorization analysis and the distance dependent Chinese restaurant process incremental clustering. The categorization results show that a high rate of Tweets entering a cluster represents the occurrence of a new event in near future. Finally, a description of the event is presented in the form of some frequent words in each cluster. In this paper, investigations on a Tweet dataset during a 6-month period indicate that the rate of sending Tweets about predictable events considerably changes before their occurrence. The use of this feature can make it possible to predict events with high degrees of precision.

Keywords Social network analysis · Event detection · Event prediction · Twitter · Incremental clustering

1 Introduction

The emergence of the web 2 has considerably changed the ways in which people interact with one another through the internet and social networks. The data generated on social networks is very valuable since it reflects aspects of real-world societies. In the meantime, these data are easily accessible through web crawlers and the application programming interface (API). These two features constitute the most important reasons why researchers study and analyze virtual social networks. There are many applications in the field of virtual social network analysis. Among all these applications, the issue of event detection and prediction has attracted more attention due to its complexity and social consequences as well as its impact on news media and crisis management in natural and unnatural disasters (Panagiotou et al. 2016).

Various definitions for event have been provided in the literature. In Aggarwal and Subbian (2012), an event refers to something that happens at a specific time and place and is addressed by the news media. The authors in Becker et al. (2011a) consider the event as a real-world incident taking place over a period of time and concurrently discussed in Twitter. In McMinn et al. (2013), an event refers to anything which is remarkable and eye-catching and occurs at a specific time and place. That is, it refers to anything which is discussed in the news media. In Dou et al. (2012), anything that changes the volume of textual data at a particular point of time is considered an event. In Fedoryszak et al. (2019), it refers to any important and significant incident that is discussed by a considerable number of people, literally becoming a trend, about which the amount of discussions changes over time. In the current study, an event in social networks is defined as a factor causing significant changes in some of the parameters and features of the social network. Naturally, the changes will be proportionate to a specific time and place.

Events can be divided into different categories such as natural events, human-made events, recent news events, public events, social and non-social events, local and global events, and more specific events such as traffic and disease outbreaks (Panagiotou et al. 2016; Daly and Geyer 2011; Goswami and Kumar 2016; Nurwidyantoro and Winarko 2013). In terms of origin, events are generally divided into

✉ A. Yavari
Abulfazl.Yavari@gmail.com

[1] Faculty of Computer Engineering and IT, Shahrood University of Technology, Shahrood, Iran

[2] Faculty of Computer and IT Engineering, Mazandaran University of Science and Technology, Babol, Iran

[3] Sydney International School of Technology and Commerce, Sydney, Australia

Published online: 09 July 2022

two categories: predictable events and unpredictable events. Predictable events are the ones occurring under the influence of factors within the social network. In fact, it is through examining the behavior of these factors that one can detect and predict events. For example, the members of a social network discuss an event by sending Tweets. But, some events, which fall into the unpredictable category, are the ones caused by factors outside the network. Such events can only be detected during or after their occurrence by social network analysis. As examples of these events, references can be made to earthquakes or the sudden death of a globally-recognized character.

The motivation that can be mentioned for event prediction is that it is done automatically from a valuable and publicly available source such as Twitter. This process can have several applications, for example, in news agencies to automatically identify and plan for future events, manage social crises and then prepare for relief, use in recommendation systems, and rank future events by their importance.

In recent years, numerous studies have been conducted addressing event detection and identification using social media analysis, for example predicting election results (Yavari et al. 2022), the spread of disease (Fu et al. 2020), and the changes in the stock market (Kolasani and Assaf 2020). Yet, these predictions are specifically designed for a specific purpose. After the filtering of Tweets, many of these studies aimed to predict specific events based on a series of keywords and statistical analyses. This study, however, was an attempt to predict all types of predictable events by analyzing the changes in the rate of Tweet sending. For this purpose, due to the dynamic nature of Twitter, incremental clustering of Tweets in a fixed time window was used. Then, based on the changes in the rate of Tweet entry to each cluster and their comparison with the threshold value, future events are predicted.

Investigation of millions of Tweets over a 6-month period revealed that, before the occurrence of real-world events, there would be significant changes in some social network features such as the number of Tweets sent in each time interval or the Tweet sending rate. These changes can form the basis for predicting events.

In the following sections, first, the related literature on the prediction of events will be reviewed in Sect. 2. Section 3 deals with the proposed method for predicting events, and Sect. 4 reports the results of the experiments conducted on the dataset.

2 Literature review

In this section, the studies related to the proposed method for predicting events are reviewed in three parts. The first part deals with the studies related to event prediction. In the second part, as mentioned in the introduction, because the proposed method detects and predicts events based on incremental clustering, several methods for event detection based on incremental clustering are reviewed. Finally, in the third part, the distance-dependent Chinese restaurant process (ddCRP) (Blei and Frazier 2011) method of incremental clustering and the non-negative matrix factorization (NMF) (Lee and Seung 1999) method as used in the proposed method are presented in more detail.

2.1 Event prediction methods

So far, various studies have addressed the prediction of specific events in social networks. These include the prediction of election results, disease outbreaks, the stock market, the sale of specific products, the number of article citations, and the outcome of sports competitions.

In Singh et al. (2017), Twitter data related to the 2017 Punjab assembly elections in India were examined. The authors, ran a set of statistical analyses of hashtags, mentions, and sentiments besides the results of previous elections, aimed at predicting the number of seats occupied by political parties. Similarly, in Budiharto and Meiliana (2018), the authors predicted the results of the Indonesian presidential election based on counting important data and analyzing sentiments. In Skoric et al. (2011), using Twitter messages, a network was formed based on how individuals disseminated political information, and then user's political tendencies were predicted. In Conover et al. (2011), Tsakalidis et al. (2015), similar studies have aimed at predicting the election results.

In Fu et al. (2020), a system was proposed to predict the outbreak of COVID-19 based on the detection and tracking of Twitter events. The system includes a pipeline to generate event-based knowledge graphs from Twitter data streams related to COVID-19 based on which to predict the way the disease will spread. In Alkouz et al. (2019), a new model called Tweetluenza, which is based on linear regression, was introduced to predict how the flu will spread by using Twitter data. In Ram et al. (2015), a new method was introduced based on the decision tree classification to predict the number of visits to medical centers for Asthma. This prediction is based on the data collected from Twitter, Google searches, and environmental sensors.

In Kolasani and Assaf (2020), apart from investigating the effectiveness of Twitter in predicting stock prices, the stock prices of the Apple company were predicted through sentiment analysis using Support Vector Machine (SVM). In Oliveira et al. (2017), Groß-Klußmann et al. (2019), the stock prices of different companies were predicted by using sentiment analysis of Twitter data.

In Huang and Pai (2020), the authors predicted movie sales, using the Least square support vector regression

(LSSVR) model. They used data from the Box Office and IMDb movie datasets and the Twitter dataset. In Asur and Huberman (2010), Asghar et al. (2018), the revenue and success of a film were predicted, using sentiment analysis after its release.

In Patel and Passi (2020), the researchers conducted a sentiment analysis of Twitter data related to the 2014 Brazil World Cup and investigated its relationship with the results of the matches. In Radosavljevic et al. (2014), the strength of each team in the 2014 World Cup was predicted by analyzing the posts in Tumbler. The prediction was made by analyzing hashtags and mentions about the teams and their players.

The problem with most event prediction methods is that they only predict a specific event. In addition, they usually have short prediction intervals. While the proposed method uses incremental clustering, it is possible to predict multiple events. Also, as will be shown in the experiments section, in the proposed method, the forecast period is extended to several weeks before the events occur. In other words, they can be predicted a few weeks before different events occur.

2.2 Incremental clustering

Incremental clustering has attracted considerable attention as one of the most widely used methods for social network analysis and event detection. Some instances are described below.

In Becker et al. (2011b), an incremental clustering algorithm is used to detect events in the Twitter data stream. The similarity of each Tweet is compared to the existing clusters. If the similarity of the incoming Tweet with none of the existing clusters reaches a threshold value, a new cluster containing that Tweet is created. At the end of the clustering process, an SVM algorithm is used to identify clusters that refer to real-world events.

In Boom et al. (2015), the incremental clustering method used in Becker et al. (2011b) has been developed based on the semantic information contained in each Tweet. The authors used the TwitterLDA method (Zhao et al. 2011) to assign a semantic topic to each Tweet. Using hashtags, they were also able to obtain higher accuracy and recall than the method in Becker et al. (2011b).

The incremental clustering method in Phuvipadawat and Murata (2010) was used to detect the events which are based on the first story. Each Tweet is compared with the first Tweet and the K top words of each cluster, and if the similarity is more than a threshold value, it is inserted in that cluster. If it is not similar to any cluster, it is inserted in a new cluster. In McMinn and Jose (2015), incremental clustering is performed based on the named entities in the Tweets. The named entities include individuals, places, and organizations. In this study, the burst detection technique was used to identify clusters related to occurring events. Of course, the accuracy of the method strongly depends on the system detecting the named entities.

In TwitterNews+ Hasan et al. (2019), a two-step operation is performed to detect events. The first step involves identifying the burst and the second step involves clustering the Tweets that discuss the same event. The first step is taken via a search module. The search module uses an inverse index in which each entry contains a word and a finite list of the latest Tweets that contain that word and are updated regularly. The operation in step 2 is done by a module called EventCluster. This module is responsible for a quick decision to allocate a Tweet to a cluster using a method based on incremental clustering. A cosine similarity comparison with the centroid of these event clusters is performed to find the cluster with the most compliance above a certain threshold value. If no such cluster is found, a new cluster will be created to which the Tweet will be assigned. One of the advantages of this method is the online detection of news events with high accuracy.

There are several incremental clustering methods, each of which can be used in the proposed method. But the reason that the proposed method uses ddCRP clustering is that, firstly, it is based on the known CRP method and secondly, it is based only on one parameter that is easily adjustable.

2.3 NMF and the ddCRP methods

In linear algebra, matrix analysis is used as a tool for data analysis. Each factorization provides a different interpretation of the implicit structure of the data which are mathematically equivalent. singular value decomposition (SVD) (Lathauwer et al. 2000), principal component analysis (PCA), and NMF are among the most popular methods of matrix analysis. In the NMF algorithm, the non-negative matrix $\mathbf{A} \in \mathbb{R}^{m*n}$ is as input data. This algorithm obtains the two non-negative matrices of $\mathbf{H} \in \mathbb{R}^{r*n}$ and $\mathbf{W} \in \mathbb{R}^{m*r}$ so that approximately $\mathbf{A} \approx \mathbf{WH}$. The number r is smaller than the minimum values of the numbers m and n. For example, in text processing, where the matrix $\mathbf{A}$ is considered as a word × document, the output of the NMF algorithm is conceptually two matrices of word × subject and subject × document. In this paper, the subject × document matrix is used at the first step in clustering. Based on the evaluations in Chen et al. (2019), the NMF matrix analysis algorithm operates better than the latent Dirichlet allocation (LDA) probabilistic algorithm in detecting the subjects of short documents. The ddCRP algorithm is an incremental clustering method based on probability theory and the Bayesian network framework. The ddCRP algorithm is theoretically capable of generating clusters of an infinite number using the Dirichlet modeling process (Gershman and Blei 2012). This algorithm is a development of the Chinese restaurant process (CRP) algorithm, which is itself one of the main developments of the Bayesian network framework. In the CRP algorithm, it is assumed that there is a restaurant with an unlimited number and capacity of tables. The first customer sits at the first table. New customers can sit at tables where one or more persons are already sitting or choose a new table. This selection is made using a probability function that depends on the number of people around a table:

$$P\left(c_{N+1} = k|\alpha, n\right) = \begin{cases} \frac{n_k}{N+\alpha} & k \in K \\ \frac{\alpha}{N+\alpha} & k = K+1 \end{cases} \quad (1)$$

Equation (1) calculates the probability of the $N+1$ customer (c_{N+1}) sitting at each of the K tables where customers are already sitting or at a new table. α is called the concentration parameter which regulates the concentration of the customers sitting at the tables. N is the total number of customers at the tables, and K is the total number of occupied tables so far. n_k is the number of customers around table k.

The ddCRP algorithm makes the probability dependent on the degree of similarity among the customers around the tables rather than their number (Eq. 2). Figure 1 shows how ddCRP functions. Each incoming tweet is linked to an existing tweet based on a similarity such as cosine similarity (in other words, they are placed in a cluster) and each cluster is considered as an event E_i.

$$P\left(c_{N+1} = k|\alpha, S\right) = \begin{cases} s_{N+1,c_k} & k \in K \\ \frac{\alpha}{\alpha+N} & k = K+1 \end{cases} \quad (2)$$

In Eq. (2), s_{N+1,c_k} is the degree of similarity between the new customer ($N+1$) and the customers around table k, and S is the set of all similarity values among the customers. Since in Eq. (2), there is alpha in the numerator, the larger its value, the more likely it is to select a new cluster and as a result, the number of clusters increases faster, and vice versa. In the proposed method, $\alpha = 0.2$ was selected experimentally. In fact, in text document clustering, each document is considered as a customer and the clusters are considered as tables.

The cosine similarity McMinn and Jose (2015) is used to calculate the similarity between the two Tweets, in this study. To this end, the text vector of Tweets from the matrix term frequency inverse document frequency (TF-IDF) is retrieved and the similarities between the two Tweets t_i and t_j is calculated according to Eq. (3).

$$s\left(t_i, t_j\right) = \frac{t_i.t_j}{t_i t_j} \quad (3)$$

3 The proposed method

This section describes the proposed method for predicting events. To make an accurate prediction, effective features for prediction must be defined. Among the features which can be evaluated in the process of detecting and predicting events, one can refer to features such as new user registration, the rate of Tweets on the network, the number of hashtags, re-Tweets, mentions, links, user responses, and the rate of sending texts by influential characters.

This study draws upon the changes in the rate of sending messages to the social network. The goal is to show that events can be predicted by examining the message sending rate. For this purpose, it is first shown that before the occurrence of predictable events, the Tweets related to those events are sent to the network. Then, an event predicting method is proposed which predicts events before their occurrence by utilizing the fluctuation feature in the rate of sending Tweets.

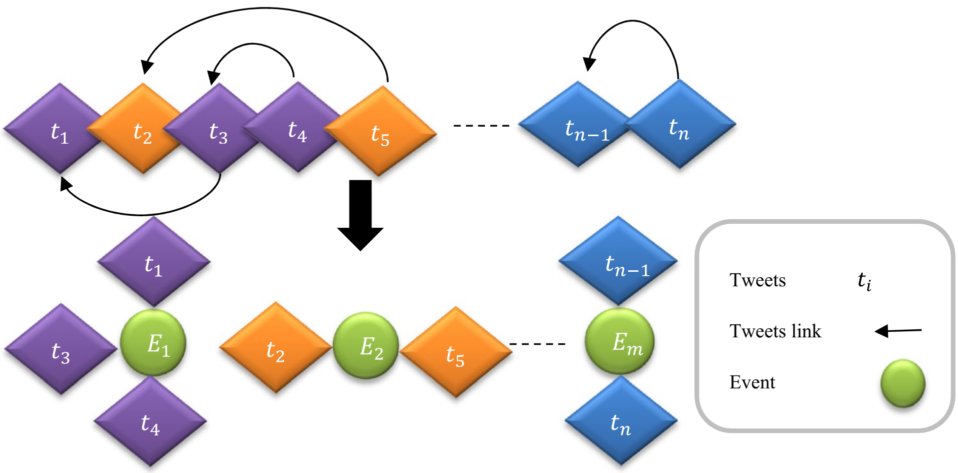


Fig. 1 ddCRP incremental clustering

3.1 Network behavior before the event

In the real world, various trends, talks, and movements about an event usually happen before it actually occurs. For example, months before the elections, the talks begin about it. Before the coming of a flood, numerous warnings and messages are issued about the weather. Coordination meetings and messages are arranged before the break-out of street riots. The World Cup and concerts are also examples of such events. Therefore, before the occurrence of predictable events, signs of their future occurrence will appear in society. Naturally, no sign of unpredictable events such as plane crashes or earthquakes will appear before their occurrence, and they are discussed after they occur.

Another common feature of predictable events is that, as we get closer to their occurrence, the emergence of their signs gets faster until it reaches its maximum during or after the occurrence of those events. Then, this rate of emergence slows down until it gradually disappears. This rise and fall in the emergence of signs are observable when real-world events occur. This section aims to discuss the existence of such similar behaviors in Twitter.

Figure 2, for instance, shows the number of Tweets about two different events. These events have been selected from a set of predictable events. The first event was the recognition of homosexuals' marriage in Utah on December 20, 2013, for which the first Tweets were sent to the network on December 16, 2013. The second event was the peace talks between Palestine and Israel on April 23, 2014, the first signs of which were Tweeted four days before. The horizontal axis in the graphs shows the half-hour intervals from the first Tweet to the last Tweet about the event. The interval with the maximum number of Tweets shows the time of occurrence. As depicted in Fig. 2, expected behaviors with different intensities can be seen in the graphs. This means that messages about an event were disseminated on the network before its occurrence. The time-lapse between the first and the last signs of an event and the number of Tweets associated with it depends on the type of the event. Some events have more or less social effects; therefore, changes in the number of messages, i.e., changes in the message sending rate, will have different trends.

As an example of unpredictable events, Fig. 3 displays the number of Tweets in half-hour intervals for an unpredictable event. The event is related to the bomb blast in Baghdad market on December 25, 2013. As it can be seen, there are almost no prior signs of occurrence in this part either, but immediately after occurrence, Tweets have been sent to the network at a high rate.

3.2 Event prediction

In the previous section, the existence of changes in the rate of sending Tweets about an event in Twitter was discussed. This section aims to propose a method for predicting events based on the feature of change in the rate of Tweets about them.

The architecture of the proposed method is shown in Fig. 4. This method consists of three main components. The first component, called pre-processing, consists of two steps. First, incoming Tweets are grouped based on a fixed-length time window. Next, the usual text-data preprocessing operations in data mining systems are performed to clear them, and finally, the TF-IDF weight vector is created. In the second component of the proposed method, in addition to clustering the Tweets, the feature of change in the rate of Tweet entry into each cluster is examined to predict the occurrence of the event.

Finally, in the last component of the proposed method, the predicted events are visualized. In the following section, each of the main parts of the method is explained in more detail.

3.2.1 Pre-processing

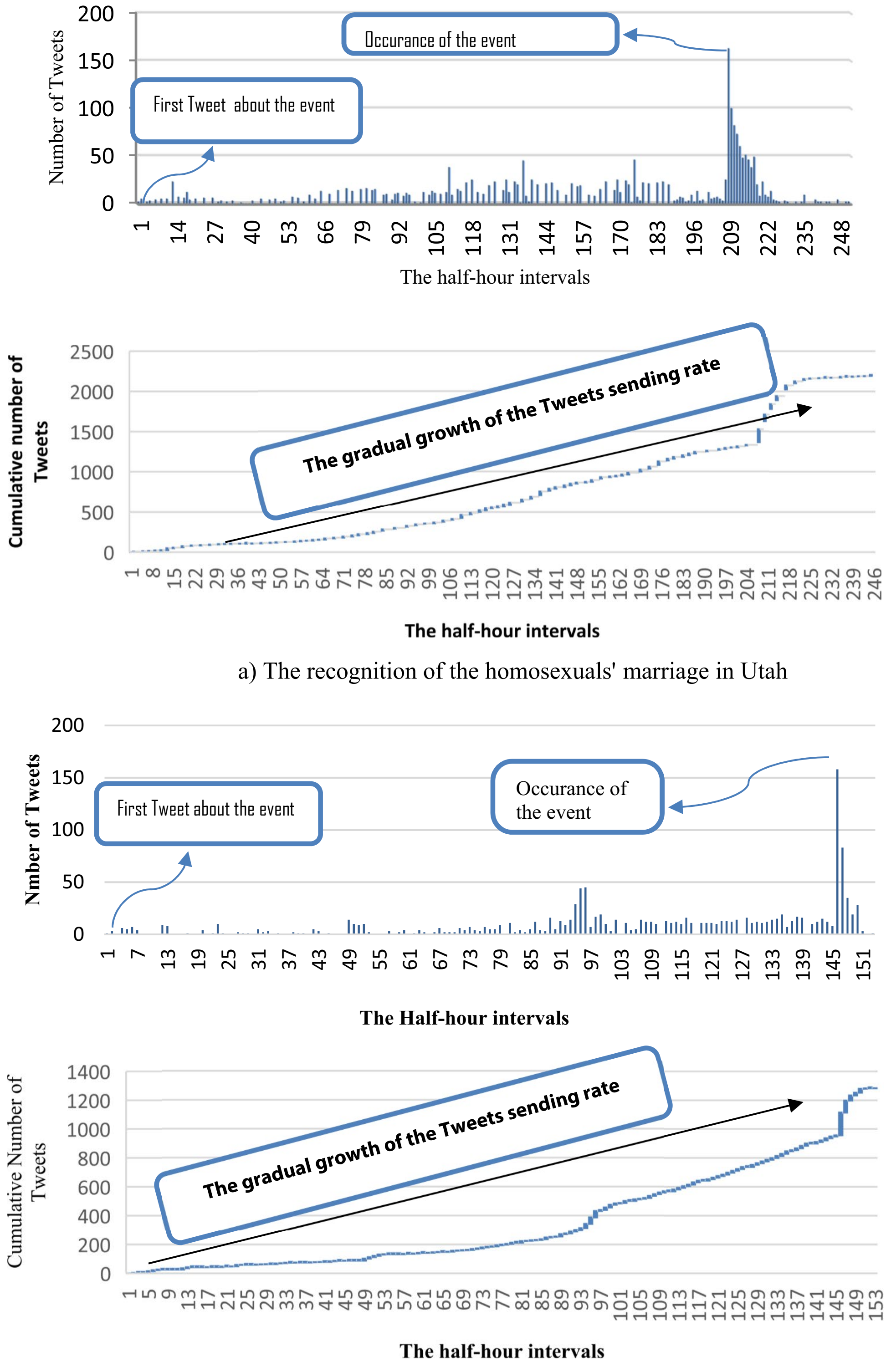
In general, text documents often contain useless data that not only negatively influence the final accuracy but also influence data volume and processing time. Therefore, they should be cleaned before processing.

In this study, before cleaning the data, Tweets are categorized based on a time window whose length is adjustable at system startup. This grouping of Tweets is important in two respects. First, since subsequent system analysis is based on the changes in the rate at which Tweets enter the system, it simplifies calculations. This is because the duration of time is constant during the intervals, and changes in the number of sent Tweets can represent changes in the rate at which Tweets are sent. Second, the online reading of Tweets can be simulated for the system in this way. After grouping the Tweets according to the time of their dissemination, each group is read respectively and the usual actions of cleaning and preprocessing the data such as tokenizing, removing stop words, numbers and links, and lemmatizing are done on Tweets. At the end of this step, using the method TF-IDF (Eq. 4), the weight matrix of the term×document is created, which contains the importance of each term in the Tweets, and is used as input to the prediction step.

$$\text{TF-IDF}\left(w, t, N, n_w\right) = f_{w,t}.\log\left(\frac{N}{n_w}\right) \tag{4}$$

In Eq. (4), $f_{w,t}$ is the frequency of the word w in Tweet t, N is the total number of Tweets and n_w is the number of Tweets in which w appears.

Fig. 2 Changes in the rate of sending Tweets for two predictable events: **a** the recognition of the homosexuals' marriage in Utah, **b** Palestine-Israeli peace talks



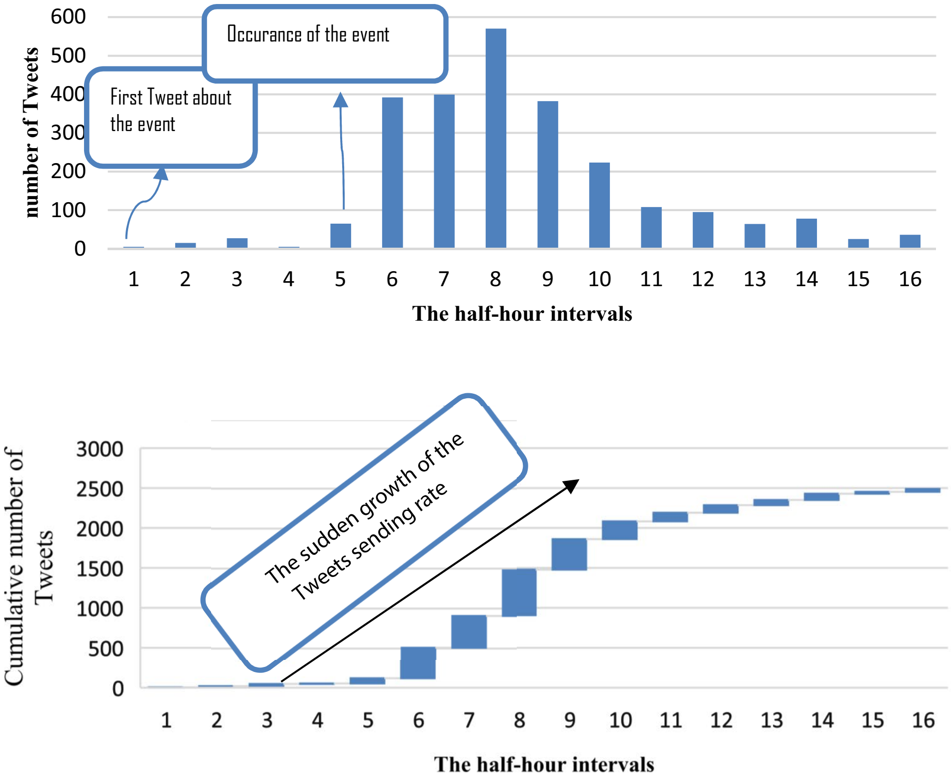
a) The recognition of the homosexuals' marriage in Utah

b) Palestine-Israeli peace talks

3.2.2 Event prediction

Traditional clustering methods such as *k*-means clustering are not very useful in the field of stream textual data clustering, since the number of clusters needs to be known at the beginning and for stream data, they have high computational complexity. Another logically plausible reason is that, in dynamic operational environments such as social networks, information is constantly increasing and evolving. Therefore, considering a high limit for the number of clusters does not

Fig. 3 Changes in the rate of sending Tweets about a sudden event



seem to be correct. So, we should find an alternative solution which is compatible with the evolution of previous information and the emergence of new information. This is why incremental clustering algorithms are used for this purpose.

This study uses the incremental clustering method. At the beginning of the incremental clustering process, the first group of Tweets is clustered using the NMF algorithm to create a good initialization for the initial clusters with a suitable initial structure as fast as possible. Then, by reading each new Tweet set, the incremental clustering procedure is performed. In this method, the ddCRP incremental clustering method is used. At the end of each incremental clustering step, a set of clusters is available. In fact, each cluster represents an event. After clustering each group of Tweets, the Tweet entry rate to each cluster is calculated and stored. Of course, since the grouping of Tweets in the preprocessing phase was based on a time window with a fixed length, the rate of Tweet entry in the time interval i, ($V_{t,i}$) have a direct correlation with the number of incoming Tweets to each cluster in the same interval ($n_{t,i}$) (Eq. 5)

$$V_{t,i} = {}^{n_{t,i}}\!/_{i} \quad (5)$$

In Eq. (5), i is the length of the interval.

Now, it is possible to predict the future occurrence of an event based on the amount of changes in the rate of Tweets entering each cluster and comparing this change rate with a series of thresholds that are obtained experimentally. It needs to be stated that based on the amount of changes in the rate of messages entering the clusters, in general, the system encounters three types of clusters. The first type includes the clusters that represent unpredictable events with a highly accelerated growth rate. The second type includes the clusters that have a growth acceleration rate below a threshold, do not actually refer to a specific event, and are sometimes created due to the existence of noise or outlier data. And finally, the third type refers to the clusters which represent valid events in the future. In the implementation, the second type of clusters are removed from the cluster set because they may not contain useful information or refer to very trivial events. Consequently, the reduction in the number of clusters improves the processing time and accuracy of the system.

3.2.3 Displaying the future events

The last part of the proposed method is the visualization of the predicted events. From each cluster displaying future events, the K most frequent words are sent to the output as a description of the event. Table 1 shows the output of the proposed method that is applied to the dataset described in the next section.

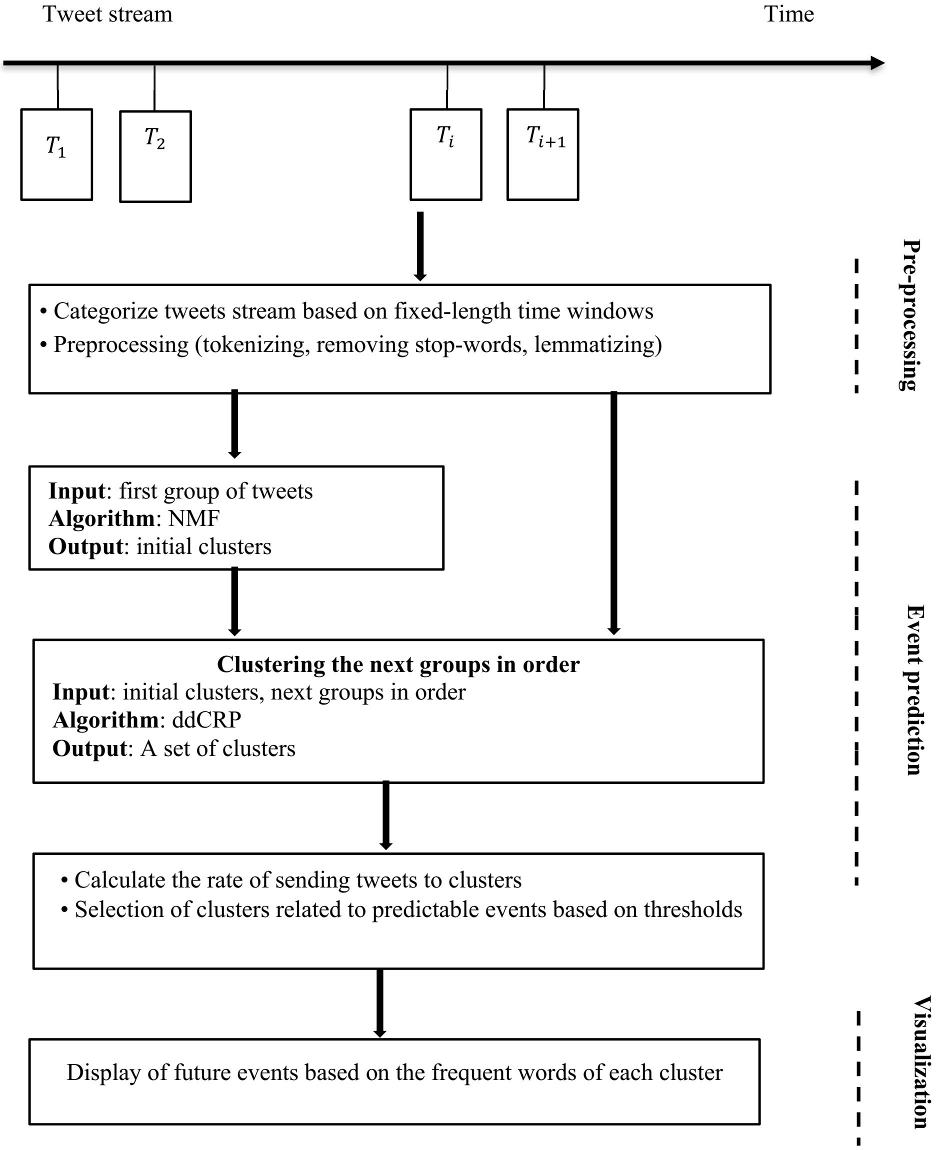

Tweet stream
Time
T_1
T_2
T_i
T_{i+1}
• Categorize tweets stream based on fixed-length time windows
• Preprocessing (tokenizing, removing stop-words, lemmatizing)
Pre-processing
Input: first group of tweets
Algorithm: NMF
Output: initial clusters
Clustering the next groups in order
Input: initial clusters, next groups in order
Algorithm: ddCRP
Output: A set of clusters
Event prediction
• Calculate the rate of sending tweets to clusters
• Selection of clusters related to predictable events based on thresholds
Display of future events based on the frequent words of each cluster
Visualization

Fig. 4 General architecture of the proposed method for event prediction

Table 1 The output of the proposed method as the predicted events

	K most frequent words of some clusters as a description of predicted events ($K = 6$)
1	Chemical, Kerry, use, house, weapons, stand
2	Divorce, old, request, wife, George, make
3	Fire, hospital, immediate, dead, sandy, topic
4	Merkel, exit, angel, third, car, stop
5	Mayor, healthcare, new, india, first, york

4 Experimental results

In this section, the experiments on a Tweet dataset are elaborated on. In the following, the dataset used in the experiments is introduced. Then, the baseline methods used for comparison and the evaluation metrics are described. Finally, the experiments performed on the dataset are given. The first experiment aimed to show that the rate of sending Tweets about events on Twitter increases before its occurrence. The second experiment attempted to evaluate the precision of the proposed method in predicting the occurrence of events based on the changes in the rate of sending Tweets.

4.1 Dataset description

The dataset used in this research is selected from the article (Kalyanam et al. 2016). This dataset includes Tweets taken from popular news media such as CNN, BreakNews, and BBC in Twitter during a period of 6 months. The dataset contains approximately 43 million Tweet identities referring to 5234 events. Each event in the dataset is described by several keywords which are used to evaluate the proposed method.

For each Tweet, the required information includes ID, text, and timestamp. Unfortunately, approximately only 27 million Tweets were collected, because some accounts had been blocked and some Tweets had been deleted. But, fortunately, this number of Tweets was enough to cover the complete set of the 5234 events. However, as the proposed method was highly dependent on the number of Tweets for each event, only the events for which more than 80% of the Tweets were available were selected in the experiments. Therefore, the number of final events was 1940 and the number of total Tweets reduced to almost 12 million. The statistical summary of the final dataset is given in Table 2.

4.2 Baseline methods

The proposed method is compared to three methods MABED (Guille and Favre 2015), Embed2Detect (Hettiarachchi et al. 2021), and FSD (Petrović et al. 2010). The criteria related to each of these methods are calculated based on their global implementation in GitHub. MABED detects events by a statistical approach as well as by detecting anomalies in the number of words sent to Twitter. FSD detects the occurrence of an event by an incremental approach and the first-appearance detection method. The Embed2Detect method is a new event detection method that detects events based on semantic features embedded in words and hierarchical clustering.

Table 2 Statistical summary of the final dataset

Statistics of the dataset	Min	Avg	Max
The number of Tweets per event	1120	6625	38,436
The number of keywords per event	2	3.65	29

4.3 Evaluation metrics

The efficiency of the proposed method is evaluated based on the criteria of the standard evaluation of precision, recall, and $F1$-score. The precision criterion indicates what percentage of the output of the proposed method accurately represents the event. The recall criterion specifies what percentage of the events in the dataset are correctly predicted by the proposed method. Therefore, to calculate precision (P) and recall (R), Eqs. (6) and (7) are used, respectively.

$$P = \frac{|\{\text{DS Events}\} \cap \{\text{Predicted Events}\}|}{|\{\text{Predicted Events}\}|} \tag{6}$$

$$R = \frac{|\{\text{DS Events}\} \cap \{\text{Predicted Events}\}|}{|\{\text{DS Events}\}|} \tag{7}$$

$F1$-score, a kind of average between precision and recall, is obtained through Eq. (8).

$$F1\text{-Score} = 2 \times \frac{P \times R}{P + R} \tag{8}$$

4.4 Experiments

In the first experiment, the behavior of Twitter users before the occurrence of predictable events and sudden events was examined. Labeling events as predictable or sudden is done manually in the dataset. The results of the experiment are shown in Fig. 5. The horizontal axis shows the number of half-hour intervals until the occurrence of the event. The vertical axis represents the average number of Tweets about events. As seen in the figure, for predictable events, from at least 2161 half-time intervals or 45 days before their occurrence, some Tweets were sent about them. In addition, the closer we get to the time of occurrence, the higher the rate of sending Tweets will be.

In the second experiment, the proposed method was examined for its predictive accuracy. In order to identify

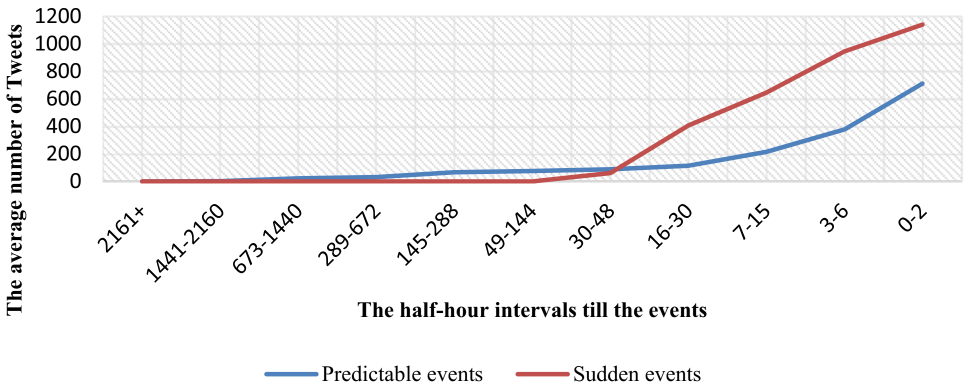
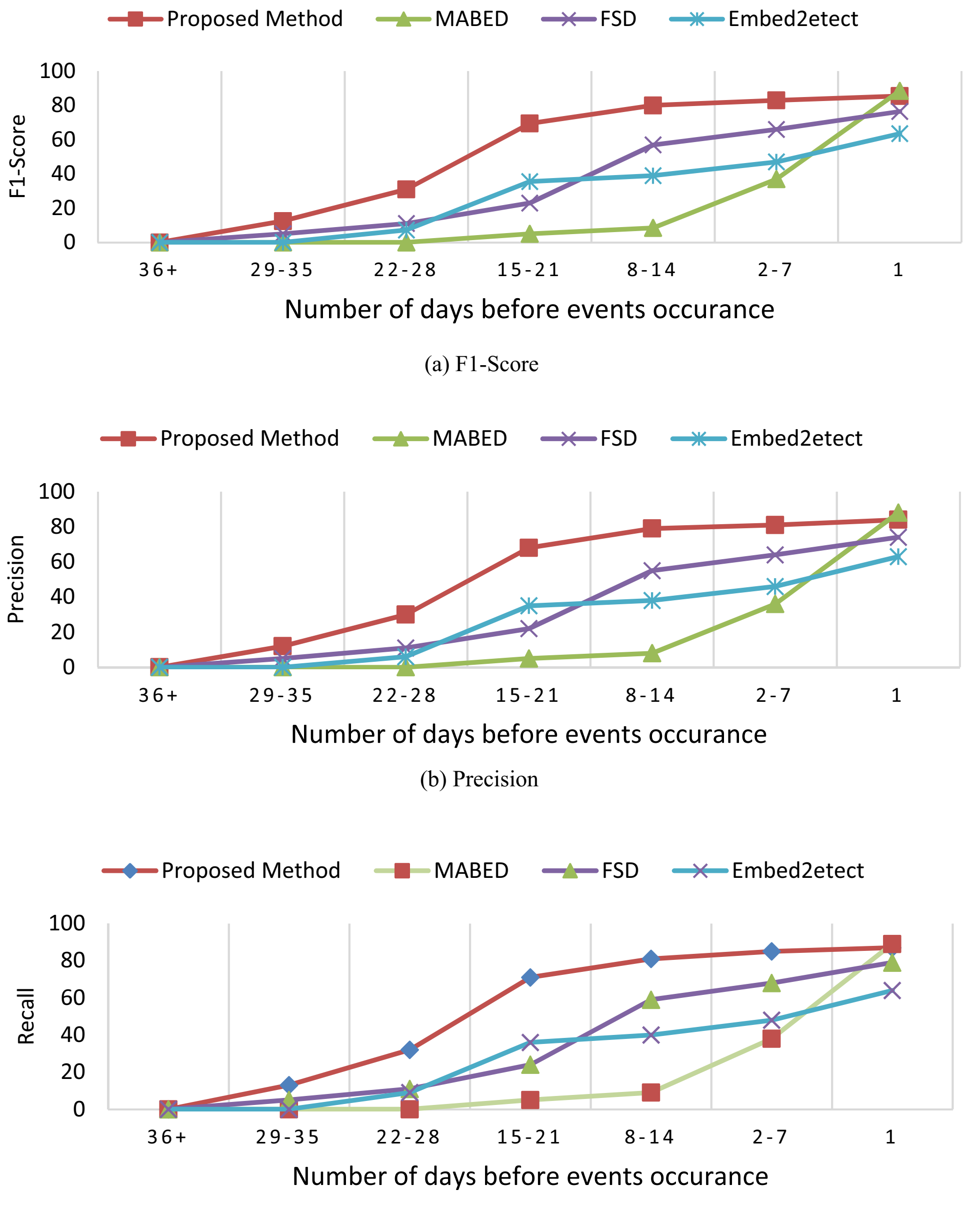


Fig. 5 The average number of Tweets sent in half-hour intervals

method output events, the most frequent words of each cluster, which are considered as a description of the predicted events, were compared with the keywords of the events in the dataset. In case the system output keywords have a minimum similarity (s) to the keywords associated with each event in the dataset, it is considered an identified event. The accuracy of the proposed method was calculated based on setting the incremental clustering density parameter to $\alpha = 0.2$ and similarity to $S = 70\%$. The results of comparing the proposed method with the baseline methods (MABED, Embed2Detect, FSD) in terms of accuracy, recall, and $F1$-score are shown in Fig. 6.

The higher predictive precision of the proposed method, from 5 weeks before the occurrence, in comparison to the other methods, is evident in Fig. 6b. The average precision of the proposed method in predicting events from 3 weeks before occurrence was 71%, which rose to 81% and 85%, from 2 weeks and 1 week before the occurrence, respectively. While in other compared methods, the accuracy of predicting the occurrence of the event is less. On the day leading to the event, mostly true of sudden events, the accuracy of the proposed method in predicting and detecting events rose to 87%. Similarly, on the day of occurrence, there was an increase in the precision of the other three event detection methods, especially the MABED method, which operates based on the increase in the anomaly. According to Fig. 6c, the proposed method predicts a higher percentage of events in the dataset. Since MABED only checks for

Fig. 6 The results of comparing the proposed method with other existing models **a** $F1$-measure results, **b** precision results, **c** recall results



(a) F1-Score

(b) Precision

(c) Recall

tweets that have a mention, it is faster than other methods but has little accuracy in the prediction process. For example, in the 2 weeks before the event, its precision is 8%. While the precision of the proposed method in the 2 weeks left to the event is 81%. However, as the event approaches since MABED is based on the detection of anomalies, its precision increases and eventually becomes approximately equal to the accuracy of the proposed method.

Embed2Detect as a new method in event detection, combining the semantic feature of words and hierarchical clustering, deals with event detection. The time complexity of this method is equal to $O\left(N^2 \log N\right)$ (Hettiarachchi et al. 2021). In the proposed method, the ddCRP clustering method is the most time-consuming phase of it, and if the number of tweets is equal to N since each tweet is compared to all the tweets in the clusters to find a similar tweet, the time complexity of the proposed method is $O\left(N^2\right)$. Although the time order of the proposed method is less than Embed-2Detect, Both do not seem to be real-time for big data. However, they can be executed in real-time by sampling or using thresholds to control the size of the data. In general, the proposed method is more accurate in predicting the events and its traceability is more practical.

5 Conclusion and suggestions for further research

In this article, the prediction of real-world events was made by analyzing Twitter. Initially, it was shown that before predicting the predictable events, Tweets about those events are sent to the network. The changes in Tweet rates before the occurrence of events were used as an effective feature in predicting events. Experiments on millions of Tweets indicated that it is possible to predict events on Twitter with high precision. The experimental results showed that the proposed method has a better performance in predicting the event than the three existing methods in this field. Based on these results, monitoring the rate of sending of related tweets, which are likely to belong to an event, will be effective in predicting events several weeks before they occur. Future studies can aim at predicting the approximate time of the occurrence of events. In addition, other features of the social network such as hashtags, mentions, influential people, and user ranking (Abu-Salih et al. 2019) can be examined to improve the precision.

References

Abu-Salih B, Wongthongtham P, Chan KY, Zhu D (2019) CredSaT: credibility ranking of users in big social data incorporating semantic analysis and temporal factor. J Inf Sci 45(2):259–280. https://doi.org/10.1177/0165551518790424

Aggarwal CC, Subbian K (2012) Event detection in social streams. In: Proceedings of the 2012 SIAM international conference on data mining, pp 624–635. https://doi.org/10.1137/1.9781611972825.54

Alkouz B, Al Aghbari Z, Abawajy JH (2019) Tweetluenza: predicting flu trends from Twitter data. Big Data Min Anal 2(4):273–287

Asghar MZ, Khan A, Khan F, Kundi FM (2018) Rift: a rule induction framework for Twitter sentiment analysis. Arab J Sci Eng 43(2):857–877

Asur S, Huberman BA (2010) Predicting the future with social media. In: The 2010 IEEE/WIC/ACM international conference on web intelligence and intelligent agent technology, vol 1. IEEE, pp 492–499

Becker H, Naaman M, Gravano L (2011a) Beyond trending topics: real-world event identification on twitter. In: Proceedings of the international AAAI conference on web and social media, vol 5, no 1

Becker H, Naaman M, Gravano L (2011b) Beyond trending topics: real-world event identification on Twitter. In: Proceedings of the international AAAI conference on web and social media, vol 5, no. 1

Blei DM, Frazier PI (2011) Distance dependent Chinese restaurant processes. J Mach Learn Res 12(8):2461–2488

De Boom C, Van Canneyt S, Dhoedt B (2015) Semantics-driven event clustering in Twitter feeds. In: Making sense of microposts, vol 1395, pp 2–9

Budiharto W, Meiliana M (2018) Prediction and analysis of Indonesia presidential election from Twitter using sentiment analysis. J Big Data 5(1):1–10

Chen Y, Zhang H, Liu R, Ye Z, Lin J (2019) Experimental explorations on short text topic mining between LDA and NMF-based schemes. Knowl Based Syst 163:1–13

Conover MD, Gonçalves B, Ratkiewicz J, Flammini A, Menczer F (2011) Predicting the political alignment of Twitter users. In: Privacy, security, risk and trust and IEEE third international conference on social computing (SocialCom). IEEE, pp 192–199

Daly EM, Geyer W (2011) Effective event discovery: using location and social information for scoping event recommendations. In: Proceedings of the fifth ACM conference on recommender systems, pp 277–280

De Lathauwer L, De Moor B, Vandewalle J (2000) A multilinear singular value decomposition. SIAM J Matrix Anal Appl 21(4):1253–1278

Dou W, Wang X, Ribarsky W, Zhou M (2012) Event detection in social media data. In: IEEE VisWeek workshop on interactive visual text analytics-task driven analytics of social media content, pp 971–980

Fedoryszak M, Frederick B, Rajaram V, Zhong C (2019) Real-time event detection on social data streams. In: Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining, pp 2774–2782

Fu X, Jiang X, Qi Y, Xu M, Song Y, Zhang J, Wu X (2020) An event-centric prediction system for COVID-19. In: 2020 IEEE international conference on knowledge graph (ICKG). IEEE, pp 195–202

Gershman SJ, Blei DM (2012) A tutorial on Bayesian nonparametric models. J Math Psychol 56(1):1–12

Goswami A, Kumar A (2016) A survey of event detection techniques in online social networks. Soc Netw Anal Min 6:107

Groß-Klußmann A, König S, Ebner M (2019) Buzzwords build momentum: global financial Twitter sentiment and the aggregate stock market. Expert Syst Appl 136:171–186

Guille A, Favre C (2015) Event detection, tracking, and visualization in Twitter: a mention-anomaly-based approach. Soc Netw Anal Min 5(1):18

Hasan M, Orgun MA, Schwitter R (2019) Real-time event detection from the Twitter data stream using the TwitterNews+ Framework. Inf Process Manag 56(3):1146–1165

Hettiarachchi H, Adedoyin-Olowe M, Bhogal J, Gaber MM (2021) Embed2Detect: temporally clustered embedded words for event detection in social media. Mach Learn 111:1–39

Huang YT, Pai PF (2020) Using the least squares support vector regression to forecast movie sales with data from Twitter and movie databases. Symmetry 12(4):625

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

Kolasani SV, Assaf R (2020) Predicting stock movement using sentiment analysis of Twitter feed with neural networks. J Data Anal Inf Process 8(4):309–319

Lee DD, Seung HS (1999) Learning the parts of objects by non-negative matrix factorization. Nature 401(6755):788–791

McMinn AJ, Jose JM (2015) Real-time entity-based event detection for Twitter. In: International conference of the cross-language evaluation forum for European languages. Springer, Cham, pp 65–77

McMinn AJ, Moshfeghi Y, Jose JM (2013) Building a large-scale corpus for evaluating event detection on Twitter. In: Proceedings of the 22nd ACM international conference on information & knowledge management, pp 409–418

Nurwidyantoro A, Winarko E (2013) Event detection in social media: a survey. In: ICT for smart society (ICISS). IEEE, pp 1–5. https://doi.org/10.1109/ICTSS.2013.6588106

Oliveira N, Cortez P, Areal N (2017) The impact of microblogging data for stock market prediction: using Twitter to predict returns, volatility, trading volume, and survey sentiment indices. Expert Syst Appl 73:125–144

Panagiotou N, Katakis I, Gunopulos D (2016) Detecting events in online social networks: definitions, trends, and challenges. In: Solving large-scale learning tasks. Springer, Cham, pp 42–84. https://doi.org/10.1007/978-3-319-41706-6_2

Patel R, Passi K (2020) Sentiment analysis on twitter data of world cup soccer tournament using machine learning. IoT 1(2):218–239

Petrović S, Osborne M, Lavrenko V (2010) Streaming first story detection with application to Twitter. In: Human language technologies: the 2010 annual conference of the North American Chapter of the Association for Computational Linguistics (2010)

Phuvipadawat S, Murata T (2010) Breaking news detection and tracking in Twitter. In the 2010 IEEE/WIC/ACM international conference on web intelligence and intelligent agent technology. IEEE, pp 120–123

Radosavljevic V, Grbovic M, Djuric N, Bhamidipati N (2014) Large-scale World Cup 2014 outcome prediction based on Tumblr posts. In: KDD workshop on large-scale sports analytics

Ram S, Zhang W, Williams M, Pengetnze Y (2015) Predicting asthma-related emergency department visits using big data. IEEE J Biomed Health Inform 19(4):1216–1223

Singh P, Dwivedi YK, Kahlon KS, Pathania A, Sawhney RS (2020) Can Twitter analytics predict election outcome? An insight from 2017 Punjab assembly elections. Gov Inf Q. https://doi.org/10.1016/j.giq.2019.101444

Skoric M, Poor N, Achananuparp P, Lim EP, Jiang J (2012) Tweets and votes: a study of the 2011 Singapore general election. In: System science, 45th Hawaii international conference on. IEEE, pp 2583–2591

Tsakalidis A, Papadopoulos S, Cristea AI, Kompatsiaris Y (2015) Predicting elections for multiple countries using Twitter and polls. IEEE Intell Syst 30(2):10–17

Yavari A, Hassanpour H, Rahimpour Cami B, Mahdavi M (2022) Election prediction based on sentiment analysis using Twitter data. Int J Eng Trans B Appl. https://doi.org/10.5829/ije.2022.35.02b.13

Zhao WX, Jiang J, He J, Song Y, Achanauparp P, Lim EP, Li X (2011) Topical keyphrase extraction from Twitter. In: Proceedings of the 49th annual meeting of the association for computational linguistics: human language technologies, pp 379–388

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.