Detecting and ranking events in Twitter using diversity analysis

Daoud M. Daoud
Department of Software Engineering, Princess Sumaya University for Technology, Amman, Jordan
Email: d.daoud@psut.edu.jo

Abstract: In Twitter and in other social media channels, detecting events is very important and has many applications. However, this task is very challenging because of the huge number of tweets that are posted every minute and the massive scale of the spamming activities. In this paper, we present an innovative approach for detecting events using data posted to Twitter. The proposed approach is based on the concept of user’s attention by quantitatively modelling the diversity of hashtags using Shannon’s index. Our method records the diversity values on an hourly basis time-series. Using statistical techniques, the method locates the intervals having diversity values that fall outside the range of forecasted ones (normal state). We also present the labelling and ranking techniques that are implemented in this research. Experimental results on a dataset consisting of 15 million Arabic tweets show that our proposed approach can effectively detect real-world events in Twitter.

Keywords: social media; event detection; diversity index; Twitter; Arabic; hashtags; time-series analysis; z-score; events labelling; events ranking.

Reference to this paper should be made as follows: Daoud, D.M. (2018) ‘Detecting and ranking events in Twitter using diversity analysis’, Int. J. Business Intelligence and Data Mining, Vol. 13, Nos. 1/2/3, pp.129–146.

Biographical notes: Daoud M. Daoud received his BSc degree in Electrical and Computer Engineering from Kuwait University in 1988, MSc in Computing Science from the University of Glasgow – UK and PhD in Computing Science from Joseph Fourier University – France. He is currently serving as an Assistant Professor at Princess Sumaya University for Technology. He also served in Institute of Advanced Studies – United Nations University (1998–1999). He also worked as a Principal Investigator for the Arabic part of Universal Networking Language project (1996–1999). He also served as the Director for Next Generation Services Department at Paltel (1999–2001). His main research interests are natural language processing, machine translation, information extraction, information retrieval, data mining and social media analysis.

1 Introduction

Nowadays, social media is considered one of the most important channels that people are expressing their opinions and interacting with the events that occurred around them.
Many researchers have observed that when a real-world event attracts people attention, their views and comments are reflected promptly in different social media networks. For example, in the case of a disaster or a terrorist attack people express their opinions by retweeting, using hashtags, or engaging in discussions. The greater the significance of an event, from the perspective of the users, the more is the amount of interaction through social media networks (Rogers, 2009; Jungherr and Jürgens, 2013).

In this paper, we specifically study Twitter because of its popularity and its real-time nature (Sakaki et al., 2010). Identification of what is happening on Twitter is a challenging task as millions of tweets are posted every day. Other challenges are stemmed from the massive scale of the spamming activities.

Several approaches and techniques have been discussed by researchers to detect events on Twitter and other social media such as unsupervised learning (Becker et al., 2011), signal processing techniques (Weng and Lee, 2011), latent dirichlet allocation (Pan and Mitra, 2011), clustering (Alsaedi et al., 2016), the burstiness of terms (Zhang et al., 2013) and aging theory (Cataldi et al., 2010). According to Weng and Lee (2011), current event detection approaches can be classified into two groups: document-based methods and feature-based methods. The first one identifies events by clustering documents based on the semantics distance between documents, while the second studies the distributions of words and finds out real-world events by bringing words together. Many of the existing approaches require explicitly identifying the events to be detected. Others need to be trained for the task of detection. Some of the approaches failed to locate the time of the events.

In this work, we adopt a different approach which is based on the concept of user attention and social media focus. For demonstration, a journalist might be interested in retrieving events and memes that attracted users’ attention in Twitter in the past week. Alternatively, consider a news channel is looking for setting a comparison and measuring the impact of two events on Twitter. Clearly in both cases, one needs to quantifying the state of the social media channel at a given time by measuring group parameters: such as the number of tweets during a period, the number of active users, the variety of included links within tweets, or through the diversity of hashtags. When something extraordinary happened in the real world that captured the attention of users of a social media network, they diverge from their normal interaction patterns and make them concentrate on these events. Consequently, these real world happenings will have an impact on the dynamics and variables of the social media channel. Thus, to detect an event, we have to measure both the normal and the actual states of the system (during the time in question). If both values are significantly different, we could assume that an event had occurred that made users deviate from their regular engagement patterns.

Although we worked with Arabic tweets in our research, the technique can be extended to any other language.

To perform this task, we calculate the diversity of hashtags in an hour-by-hour time-series and attempt to find out if it is possible to detect any deviations during a time span. This raises the following questions: Is there a reference or normal value for diversity in a social media channel? Can we sense the deviations from the standard state in diversity values during events?

This paper is organised as follows: in the next section we explain the diversity of hashtags, then we present the statistical model using the diversity index. In Section 3, we describe our techniques in detecting events. Then, we describe the experiment we have conducted with a focus on the dataset used and the generated time-series. In the results
section, we provide a detailed description of the obtained results and the approaches we use to label and rank events. Finally, we evaluate the proposed approach with real case studies.

2 The diversity of hashtags

Hashtags are used on a large scale through different social media channels. Users become aware of their importance in labelling and classification of their tweets. They can start conversations and make a tweet findable by a much higher audience and increase engagement. In this research, we are going to use hashtags diversity to measure the state of Twitter at any specified time.

To give an example, we traced the four top hashtags during the hour right before the Nice attack which has occurred on 14th July 2016 in France.

<table>
<thead>
<tr>
<th>Hashtag</th>
<th>Volume</th>
</tr>
</thead>
<tbody>
<tr>
<td>#&gt;Welcome you do in your life.</td>
<td>221</td>
</tr>
<tr>
<td>#We have being fooled.</td>
<td>170</td>
</tr>
<tr>
<td>#Would you marry a divorced woman.</td>
<td>167</td>
</tr>
<tr>
<td>#Iraq</td>
<td>146</td>
</tr>
</tbody>
</table>

One hour later, the top four hashtags became:

<table>
<thead>
<tr>
<th>Hashtag</th>
<th>Volume</th>
</tr>
</thead>
<tbody>
<tr>
<td>#Nice</td>
<td>351</td>
</tr>
<tr>
<td>#France</td>
<td>300</td>
</tr>
<tr>
<td>#Khaled Aldosary in critical condition.</td>
<td>165</td>
</tr>
<tr>
<td>#Breaking news</td>
<td>165</td>
</tr>
</tbody>
</table>

Obviously, as a result of the deadly attack, Arab users replaced their interaction patterns and started to use different hashtags such as #France, #Nice.

So, it seems that hashtag usage and distribution during a given period do express the state of the social network.

The shift in users interest is remarkably shown in Figure 1 which traces four hashtags:

<table>
<thead>
<tr>
<th>Hashtag</th>
<th>Volume</th>
</tr>
</thead>
<tbody>
<tr>
<td>#Welcome you do in your life.</td>
<td></td>
</tr>
<tr>
<td>#Would you marry a divorced woman.</td>
<td></td>
</tr>
<tr>
<td>#Nice</td>
<td></td>
</tr>
<tr>
<td>#Nice attack</td>
<td></td>
</tr>
</tbody>
</table>

Figure 1(B) shows clearly that the volume of hashtags that are created and used in response to the event in Nice was increasing rapidly. On the same time, hashtags that they were popular before this event started to fade away [Figure 1(A)]. Additionally, the volume of some hashtags grows at the cost of other hashtags. The degree to which the volume of each hashtag spread at a time of interest is usually called ‘diversity’.
Figure 1  A timeline of some hashtags after and before Nice attack (see online version for colours)
Different levels of external forces have different influences on diversity. If our goal is to detect an event in a given time, we need to be able to know how diversity is affected by various variables. Because diversity indices provide more information than simply the number of hashtags present (i.e., they account for some hashtags being rare and others being frequent), they serve as valuable tools that enable researchers to quantify diversity in a social media network and describe its numerical structure.

As we have seen above, when users stopped using many of the available hashtags and start focusing on emerging ones will lead to low diversity value.

The usage of hashtag frequency alone as a measure to specify the state of the system is error prone since it is a measure of the hashtag itself and cannot be used to detect unusual activities. That is because events are usually represented by several hashtags. Additionally, many of the hashtags enjoying high frequency reflect marketing and spamming activities. On the other hand, diversity measure takes into account numerous hashtags at the same time. Thus, it is possible to monitor any change in this value at different periods.

2.1 The diversity indices

We aim to validate our assumptions by quantitatively modelling the diversity of Arabic hashtag extracted from Twitter. A diversity measurement is a mathematical measure of species variety in a community. Diversity indices provide more information about community composition than simply species richness (i.e., the number of species present). Diversity indices offer vital information about rarity and commonness of species in a community.

In this work, we use a diversity measure well-studied in ecology i.e. Shannon’s diversity index (Shannon, 1948; McDonald and Dimmick, 2003).

2.1.1 Shannon’s diversity index

To find Shannon’s diversity index we use the following formula:

$$H = \sum_{i=1}^{S} (P_i \times \ln P_i)$$

where

- $H$ is the Shannon diversity index
- $P_i$ is the fraction of $S$ made up of the $i^{th}$ species, which is the proportion ($n / N$) of items of one species found ($n$) divided by the total number of items found ($N$)
- $S$ is the total number of species found (richness)
- $\ln$ is the natural logarithm.

Shannon’s evenness ($E_h$) can be calculated by dividing $H$ by $H_{\text{max}}$ (here $H_{\text{max}} = \ln S$). Evenness assumes a value between 0 and 1 with 1 being complete evenness.

In this research, we adapt this formula to find the diversity of hashtags for a given interval. The species are the hashtags and the number individuals within each species are the number of occurrences (frequency) for each hashtag. For a demonstration, suppose we have a sample of five hashtags having the following frequencies 60, 10, 25, 1 and 4.
Thus, $S$ is equal to 5 and $N$ is equal to 100 (60 + 10 + 25 + 1 + 4). The value of $P_1$ is equal to 60 divided by 100, making the $P_1 \times \ln P_1$ expression equals to $-0.31$. Similarly, the values of $P_2$, $P_3$, $P_4$, and $P_5$ are calculated. The absolute value of their summation represents the Shannon diversity index ($H$), which is 1.07 for this particular example.

3 Our approach

When something unusual in the real world captures the attention of social networking users, they abandon their normal patterns of interaction focusing on these events. As a result, these real-world events have an effect on the dynamics and variables of the social media channel. Our approach is based on detecting emerging topics in Twitter by measuring the deviations of hashtags activity relative to some baseline.

Since we are focusing on Arabic content, and giving the difficulty of handling Arabic, we use the hashtags diversity as an indicator of human activities. This choice enables us to avoid complex processing of Arabic and utilise the widespread usage of hashtags.

3.1 Building the time-series

We create a time-series data (Truong and Anh, 2015; Malik et al., 2016) where time is discrete, and $D(t)$ is the score of hashtags diversity over the time interval $[t-1, t]$. The time-series can be implemented based on different scales (hourly, daily, weekly, etc.) to reflect the patterns of underlying human behaviour. As shown later we will use the hourly scale. For each hour in question we perform the following steps:

- retrieve the set of hashtags and their frequencies from the dataset
- calculate the Shannon diversity index for the retrieved set of hashtags
- record the calculated diversity index in the hourly time-series.

So, we obtain the hourly rhythms of the diversity of hashtags. Occasionally, it is suffered from drop periods of the diversity values, which we will refer to events. Low diversity scores indicate that there are unusual activities relative to the normal pattern of interaction such as a terror attack or a national disaster.

From the time-series, we can determine the regular and normal state of the system based on the historical data stored in the timeline. If the actual scores deviate from the normal state, we can say that something extraordinary has happened.

3.2 Finding the normal state

The first step in finding the normal state, is to convert the raw diversity scores in the hourly time-series into $Z$-scores (Yin, 1994) using the following definition:

$$ Z = \frac{D(H) - \text{Mean}}{\text{STD}} $$

where $D(H)$ is the diversity value of hour $H$, the mean is the average diversity value, and STD is the standard deviation of the diversities. Therefore, $Z$-score expresses the distant between a raw score and a reference (the mean) as multiples of standard deviation units.
Thus, each raw diversity score is redefined in terms of how distant it is from the collection mean. It is useful for data analytics, especially when one needs to compare and rate some raw results (Ertl et al., 2012).

To give an example, if a system returns a Z-score of –3.5 it is interpreted as “–3.5 standard deviations away from the mean”.

If a group of raw scores is converted into a group of Z-scores, the new distribution would have a mean of zero and a standard deviation of one (Weiss, 2016). A key concept here is that the values in the mid of the distribution, represent the forecasted outcome (the norm or the normal state). When the absolute value of the Z-score is large (in the tails of the distribution), it indicates something abnormal. We assume the resulted Z-score distribution is normal or approximately normal, or the sample size is sufficiently large. Since we are interested in low diversity values as an indicator of unusual activities, we perform left-tailed test. In this type of test, the critical region is on the left side of the distribution where the values of Z-score are all negatives as shown in Figure 2. Thus, for a Z distribution and for a left-tailed test:

\[
p - \text{value} = P(Z \leq z^*)
\]

where \(z^*\) is the test statistics of a \(z\) test (Sprinthall, 2011). The \(p\)-value is the probability that the recorded diversity is within the normal values. When the \(p\)-value is very small, it means it is very unlikely that the observed diversity is within the forecasted values. Thus, very small Z-scores are linked with very low \(p\)-values.

\[\text{Figure 2} \quad Z\text{-score distribution (see online version for colours)}\]

To determine whether a Z-score is within the normal state or not, one needs a subjective judgment by selecting a confidence level. Commonly used values for confidence levels are 80, 90, 95, or 99%. \(Z\) table is used to obtain the probability for a given Z-score at a specified confidence level.

For example, the \(p\)-value for a Z-score of –1.4 value and a 90% confidence level is 0.0807: \(p\)-value = \(P(Z \leq -1.4) = 0.0807\). This \(p\)-value is less than 0.1; indicating that the Z-score falls outside the range of normal values. In this case, it should be treated as an
abnormal score. On the other hand, if the selected confidence level is 99% for the same Z-score, the obtained p-value is greater than 0.01. Thus, it should be handled as a forecasted outcome.

**Table 1** Z-score for confidence intervals (left-tailed testing)

<table>
<thead>
<tr>
<th>Confidence level</th>
<th>Z-score</th>
</tr>
</thead>
<tbody>
<tr>
<td>99%</td>
<td>-2.33</td>
</tr>
<tr>
<td>95%</td>
<td>-1.65</td>
</tr>
<tr>
<td>90%</td>
<td>-1.29</td>
</tr>
<tr>
<td>80%</td>
<td>-0.84</td>
</tr>
<tr>
<td>70%</td>
<td>-0.52</td>
</tr>
</tbody>
</table>

As listed in Table 1, for a confidence level of 90%, Z-score values that are less or equals –1.29 are significant and representing significant activities. In the same manner, for a confidence level of 95%, a –1.65 value is required to consider a Z-score abnormal. Finally, a maximum Z-score value of –2.33 is needed to meet a confidence level of 99%.

### 3.3 Labelling

Event labelling is the process of picking descriptive, human-readable labels for the abnormalities produced by an event detection algorithm; standard events algorithms do not typically produce any such labels. Events labelling find a label that summarises the topic of each event and distinguish the events from each other.

Since we are using the diversity of hashtags to detect any unusual activity, it is most likely that any change in user behaviour is mirrored in the hashtags that are used at a given moment as shown Figure 1. For example, many hashtags are emerged and circulated during the deadly truck attack in Nice as shown in Table 3.

**Table 2** Sample of hashtags used during the truck attack in Nice

<table>
<thead>
<tr>
<th>Hashtag</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>#فرنسا</td>
<td>France</td>
</tr>
<tr>
<td>#نيس</td>
<td>Nice</td>
</tr>
<tr>
<td>#هجوم_نيس</td>
<td>Nice attack</td>
</tr>
<tr>
<td>#حادثة_نيس</td>
<td>Nice event</td>
</tr>
<tr>
<td>#اعتداء_نيس</td>
<td>Nice assault</td>
</tr>
<tr>
<td>#دهس_نيس</td>
<td>Nice run-over</td>
</tr>
</tbody>
</table>

However, we notice that despite the usage of many hashtags during this particular event, #هجوم_نيس ‘Nice attack’ became the dominant one as shown in Figure 3. As a matter of fact, the availability of many hashtags that are thematically connected during the large-scale event is an indicator of a major shift in user interaction behaviour. The study of other events shows that they share the same pattern of having several related hashtags, yet only one becoming the most frequently used.

Based on the above, and once we locate an interval having a low diversity value (the -score below the normal state value), the most frequent hashtag in that interval is selected and used to label the detected event. To give an example, the deadly attack in Nice is labelled by #هجوم_نيس.
In summary, we use the dominant hashtags to label detected events. In our approach, we first locate abnormalities using $Z$-score measurements. Then, we label those detected intervals by finding the highest frequent hashtags in those periods.

### 3.4 Ranking

Some events last longer than others. More precisely, the impact of one event through its representative hashtags has a longer duration than others.

To consider the duration factor in ranking events, we introduce the following formula:

$$Z_{total} = -\sum_{t=1}^{N} zscore_t$$

where

- $Z_{total}$ is summation of all $Z$-scores in all the intervals
- $N$ is the number of intervals
- $zscore_t$ is the $Z$-score of an interval.

Since $zscore_t$ is negative, the minus sign is used to insure a positive value for $Z_{total}$.

Thus, the $Z_{total}$ is the summation of all (hot range $Z$-scores) in all the intervals that an abnormality (represented by the dominating hashtag) exists.

### 4 The experiment

#### 4.1 Dataset

We have collected tweets over a 60 days period from 1-July-2016 to 1-September-2016. More than 15 million tweets were collected using Twitter APIs. Twitter provides 1% random sample of all the tweets via its sample API in real time.
For analysis, we performed filtering of the tweets by a language detection software to remove any non-Arabic tweets in our dataset (Daoud et al., 2016).

4.2 Hour-by-hour diversity analysis

The study of our dataset shows that 69% of the collected Arabic tweets did not use any hashtag at all. Only 31% did use one or more hashtags. We could not validate these figures from other sources for Arabic tweets, but (Hong et al., 2011) shows that hashtags usage was different across different languages. Some studied indicate that only 24% of tweets contain hashtags (Feng and Wang, 2014).

In our dataset, we have 281,640 unique hashtags. As shown in Figure 4, popular hashtags are widely used, while 53% of hashtags are used once; around 88% of them are not used more than 16 times, which shows that the large proportion of the hashtags is limited to only one user or a tiny community of users.

Figure 4 The distribution of distinct hashtags in the dataset (see online version for colours)

In other words, the usage of the most common hashtags tends to rise faster than the usage of the less common ones. This is similar to findings of Zipf (1949) in which he concluded that the frequency of words in English and in other languages follows a power-law distribution. Zipf’s law (Ha et al., 2002; Seetha et al., 2015) states that the frequency in a dataset of $i$th most common word $cf_i$ is proportional to $1/i$:

$$cf_i \propto \frac{1}{i}$$  \hspace{1cm} (5)

Zipf’s law can be written as:

$$cf_i = ci^k$$  \hspace{1cm} (6)

where $k$ and $c$ are constants that have to be fit from the collection. If we take the log of both sides we get the following formula:
Detecting and ranking events in Twitter using diversity analysis

\[ \log c_f = \log c + k \log i \]  

(7)

where

\( k \) is the slope of the line and its value should be closed to –1.

Zipf's law does not hold exactly because it is not an exact law, but a statistical one.

**Figure 5** Log-log plot of hashtags frequency vs rank (see online version for colours)

To establish that Zipf's law holds for the collected Arabic hashtags, we plot \( \log \) (frequency) on the y-axis and \( \log \) (rank) on the x-axis (Figure 5). The plot shows a line appearing approximately linear, indicating that the distribution of Arabic hashtags in Twitter follows the general trend of Zipf's distribution.

The number of Arabic hashtags varies from hour to hour. On average, we processed 1,818 distinct Arabic hashtags per hour. They also follow Zipf's law regarding distribution as most of them were used very infrequently.

Since diversity is expressed by a single value, we have calculated the Shannon diversity index on hourly-based during the two months time span. The calculated diversity used all available hashtags that are retrieved from our dataset. In this experiment, we were keen to consider all hashtags even uncommon one in our calculations. The investigation on the effect of focusing on hashtags with high frequency only is beyond the scope of this work. The generated hourly times-series of diversity values is transformed into Z-score one based on the previously discussed formula. In this experiment, we set the confidence level at 70%. The corresponding threshold of the Z-score value is –0.52. If the Z-score is less or equals the specified threshold, it is considered abnormal, and it would be labeled and ranked. Otherwise, it would be treated as a normal value.

5 The results

We have calculated the diversity hour-by-hour covering a two months time span. To visually inspect the results, we plot the corresponding trend timeline as shown in Figure 6. By examining the generated timeline, we can notice sudden drops in diversity
values suggesting major deviations from the normal state. The more is the distance from the mean; the less is the diversity.

We notice in the timeline several points having very low diversity. To validate the obtained result, we did further investigation to link those deviations with actual happenings. Table 3 highlights the most prominent events in the timeline.

**Figure 6** Hour-by-hour Shannons diversity for two months

![Hour-by-hour Shannons diversity for two months](image)

It is noticeable that the dataset under study contains similar events, but captured the attention of users differently. For example, the winning of a Kuwaiti player a gold medal in Rio2016 causes more interaction on Twitter than the winning of a Jordanian one a gold medal. The first event recorded diversity score of 4.1, while the second one recorded a diversity score of 5.8. This disparity might be explained by the fact that the number of Kuwaiti users in Twitter is larger than the Jordanians ones.

**Table 3** The top detected events

<table>
<thead>
<tr>
<th>Event</th>
<th>Date</th>
<th>Dominating hashtag</th>
</tr>
</thead>
<tbody>
<tr>
<td>Explosion in the US consulate in Jeddah</td>
<td>2016-07-03T23:49:18Z</td>
<td>#تفاعل_السفارة_الأمريكية_في_جدة</td>
</tr>
<tr>
<td>Explosion in the Medina</td>
<td>2016-07-04T16:53:25Z</td>
<td>#تفاعل_المنورة</td>
</tr>
<tr>
<td>The Nice attack</td>
<td>2016-07-14T22:34:52Z</td>
<td>#هجم_نيس</td>
</tr>
<tr>
<td>The attempted coup in Turkey</td>
<td>2016-07-15T19:53:16Z</td>
<td>#تركيا</td>
</tr>
<tr>
<td>A Kuwaiti athlete won a gold medal in Rio2016 Olympics</td>
<td>2016-08-10T16:54:32Z</td>
<td>#فهد_الديحاني</td>
</tr>
</tbody>
</table>

**Figure 7** The Z-score of hourly time series

![The Z-score of hourly time series](image)
For a more efficient quantifying, processing, rating, ranking and comparing results, we calculated the corresponding Z-scores of the raw diversity scores as shown in Figure 7. Low negative Z-scores indicate unusual activities. Table 4 highlights some of the features of the hourly generated Z-score time series. The central tendency is between –0.34 to 0.672 values.

Table 4  Quartile of Z-scores distribution

<table>
<thead>
<tr>
<th>Quartile</th>
<th>Z-score</th>
</tr>
</thead>
<tbody>
<tr>
<td>100.0%</td>
<td>Maximum 1.6884581914</td>
</tr>
<tr>
<td>99.5%</td>
<td>1.4344110453</td>
</tr>
<tr>
<td>97.5%</td>
<td>1.1803638991</td>
</tr>
<tr>
<td>90.0%</td>
<td>0.926316753</td>
</tr>
<tr>
<td>75.0%</td>
<td>Quartile 0.6722696068</td>
</tr>
<tr>
<td>50.0%</td>
<td>Median 0.1641753145</td>
</tr>
<tr>
<td>25.0%</td>
<td>Quartile –0.343918978</td>
</tr>
<tr>
<td>10.0%</td>
<td>–1.106060416</td>
</tr>
<tr>
<td>2.5%</td>
<td>–2.630343293</td>
</tr>
<tr>
<td>0.5%</td>
<td>–4.869768887</td>
</tr>
<tr>
<td>0.0%</td>
<td>Minimum –5.424861901</td>
</tr>
</tbody>
</table>

Table 5 lists the obtained results based on different confidence levels. For a 99% confidence level, 13 statistically significant values are identified. After further investigation, we find 12 values signify real-life large-scale events such as Turkey attempted coup, Nice attack, Al-Madina bombing, etc. For a less conservative confidence level, the number of detected values becomes higher. Evidently, the findings show that the percentage of non-events abnormalities increases dramatically by selecting lower levels of confidence values. For example, at 99% confidence level, the percentage of detected real-life events is 92%. While it is only 55.9% for a confidence level of 70%.

Table 5  Detected abnormalities at different confidence levels

<table>
<thead>
<tr>
<th>Confidence level</th>
<th>Real-life events</th>
<th>Non-events</th>
</tr>
</thead>
<tbody>
<tr>
<td>99%</td>
<td>12</td>
<td>1</td>
</tr>
<tr>
<td>95%</td>
<td>19</td>
<td>2</td>
</tr>
<tr>
<td>90%</td>
<td>24</td>
<td>5</td>
</tr>
<tr>
<td>80%</td>
<td>44</td>
<td>25</td>
</tr>
<tr>
<td>70%</td>
<td>52</td>
<td>41</td>
</tr>
</tbody>
</table>

As discussed before, the hashtag with the highest frequency is selected to label the detected abnormality.

After removing duplicate hashtags from the result set, we ended with 93 labelled abnormalities (using 70% confidence level). Some of the hashtags denote real-life events as we have shown above. However, others signify non-events, or they are meme hashtags (Shapp, 2014) such as the Egyptian education needs’ and “a restaurant that you regretted to eat in it”. Whether memes are the real cause of the abnormalities or they are simply overshadowing other real-life events is something needs more investigation. In the next section, we discuss further how to identify them and minimise their effects.
We noticed that some events last longer than others. More precisely, the impact of one event through its representative hashtags has a longer duration than others. Figure 8 shows the $Z$-score for the hashtags #هجوم_نيس which is the dominating hashtag of the Nice attack incident. This hashtag dominates the scene for 16 hours. Some other dominating hashtags representing events has a high absolute $Z$-score (within the hot range), yet they have a shorter life. We observe that 41.9% of the detected anomalies has a one-hour time duration. For example, the hashtag #فيديو_الدحاني has a duration of nine hours. If we consider the value of $Z$-score alone, #فيديو_الدحاني is ranked before #هجوم_نيس as shown in Figure 9.

In the same line, we observe that most of the abnormalities that are labelled with meme hashtags have a short life. Our findings show that 60% of detected potential events with one-hour duration are memes.

**Figure 8** The timeline for nice attack

![Figure 8](image)

**Figure 9** Ranking based on highest $Z$-score value

![Figure 9](image)

In Figure 8, #فيديو_الدحاني is ranked the second event. However, when we consider the duration, it is ranked the seventh event as shown in Figure 10.

Using the $Z_{total}$ score for ranking increases the relevancy. Our findings show that 15 of the top ranked hashtags denote real-life events.
6 Evaluation

6.1 Methods of measuring the efficiency

We will use a comparative approach for the evaluation: the output of the experimental system is compared with a gold standard (Paroubek et al., 2007) (Reference list) that is an external resource.

6.2 Comparison with a reference list of events

Let us suppose that we have reference list containing a set of events for a specified period. The list of events that have been detected by the experimental system can be compared with this list, and the two measures of recall and precision (Salton and McGill, 1983) can be calculated.

\[
\text{recall} = \frac{\text{number of detected events in the reference list}}{\text{umber of events in the reference list}} = \frac{n(A \cap R)}{n(R)} \quad (8)
\]

\[
\text{precision} = \frac{\text{number of detected events in the reference list}}{\text{umber of detected events}} = \frac{n(A \cap R)}{n(A)} \quad (9)
\]

where

- \( A \) is the set of events detected by the experimental system.
- \( R \) is the set of events of the reference list.
- \( n(X) \) is the number of elements of any set \( X \).

To perform this evaluation, we used two reference lists of events during the month of July 2016. The source of both lists is Al-Jazeera news channel. The first reference list was published by Al-Jazeera under the title *July 2016... A Month That Will Be*
Remembered by the World (Al-Jazeera, 2016). This list contains 13 large scale events such Al-Karrada bombing, Nice attack, Al-Madina blast, Jeddah explosion.

The second reference list contains 46 events that range from small to significant ones. During July 2016, our experimental system detected 17 real events such as Turkey attempted coup, Nice attack, Al-Madina bombing, Jeddah blast, Dallas shooting, Munich attack, Al-Karada bombing, etc.

6.3 Evaluation results based on the first reference list

1. Recall 62% (8 out of 13) of events in the first reference list were extracted by our experimental system. This means that five events were missed by our tools. For example, Arab League summit event is not detected, yet it is in the reference list. In other words, this event did not attract users attention despite its significance (at least from the professional point of view). In a sense, the recall rate expresses the proportion of events that have a high impact within Twitter and exist in the reference list.

2. Precision 47% (8 out of 17) of the detected events by our system are found in the first reference list. This means that nine of the extracted events are not included in this list. Among the missing events Munich shooting and Aleppo bombardment.

6.4 Evaluation results based on the second list

The second reference list contains daily events posted on Twitter by Al-Jazeera breaking news service.

1. Recall 30.4% (14 out of 46) of events were detected by our system. This suggests that many of the posts in the second list did not attract much attention by the users. These events are missed for several reasons:
   - they did not cause users to change their normal behaviour
   - the hashtags representing those events are masked by other dominating hashtags.

2. Precision 82% (14 out of 17) of the extracted events are found in the second reference list. Three events which have a high impact on Twitter are not reported by Al Jazeera. Some of the non-political events are detected (such as admission to universities, or scandals of some celebrities) and are not reported by a news agency.

7 Conclusions

In this paper, we presented an event detection and examination technique for Twitter. The system uses the diversity of hashtags scheme based on Shannon’s diversity measure. We introduced a labelling and ranking approaches. We evaluate the system using gold standard by comparing our results with two reference lists obtained from Al-Jazeera news channel.

For future work, we will further investigate hashtag-based diversity analysis and improve the current detection approach to allow for faster and a more accurate analysis. We will consider eliminating noisy hashtags from processing to improve detection of small events.
Detecting and ranking events in Twitter using diversity analysis

References


