Arabic Event Detection in Social Media

Nasser Alsaedi, Pete Burnap

Cardiff School of Computer Science & Informatics, Cardiff University, UK
{N.M.Alsaedi, P.Burnap}@cs.cardiff.ac.uk

Abstract. Event detection is a concept that is crucial to the assurance of public safety surrounding real-world events. Decision makers use information from a range of terrestrial and online sources to help inform decisions that enable them to develop policies and react appropriately to events as they unfold. One such source of online information is social media. Twitter, as a form of social media, is a popular micro-blogging web application serving hundreds of millions of users. User-generated content can be utilized as a rich source of information to identify real-world events. In this paper, we present a novel detection framework for identifying such events, with a focus on 'disruptive' events using Twitter data. The approach is based on five steps; data collection, pre-processing, classification, clustering and summarization. We use a Naïve Bayes classification model and an Online Clustering method to validate our model over multiple real-world data sets. To the best of our knowledge, this study is the first effort to identify real-world events in Arabic from social media.

Keywords: Text mining, Information Extraction, Classification, Online-Clustering, Machine Learning, Event detection.

1 Introduction

In the recent years, microblogging, as a form of social media, has rapidly grown in popularity as a mechanism for expressing opinions, broadcasting news and supporting interaction between people. One of the most representative examples is Twitter, which allows users to publish short tweets (messages within a 140-character limit) about any subject, including commentary on real-world events. Events can be community-specific, such as local gatherings, or can be wide-reaching national or even international level events. At an international level, people use social media to comment on events such as presidential elections, health pandemics, natural and man-made disasters, and major sport events as they are happening, and even before mainstream media release information about the event [8, 11, 14].
Wenwen-Dou defined an event on social media as:
"*An occurrence causing change in the volume of text data that discusses the associated topic at a specific time*." [20].

Here, we use the same definition where events have different degrees of importance causing the different "volume change" when discussed in social media platforms. Thus, an event can be characterized by a 'bursty' increase in particular

terms or words at some point in time. In this paper we are particularly interested in whether we can identify *disruptive* events using social media, and distinguish between these and other events. Examples of such events include protests, terrorist attacks, transport loss and crimes. In [1] disruptive events in the context of social media are defined as:

"*An event that interferes the achieving of the objective of an event or interrupts ordinary event routine. It may occur over the course of one or several days, causing disorder, destabilizing securities and may results in a displacement or discontinuity*."

Our objective is therefore to identify these events so that disruption, security issues, and disorder, can be managed and minimized. As events are typically 'bursty' topics of interest, they can lead to an instant and voluminous social reaction. Identifying events using the public reaction published openly via social media presents a number of benefits for planning and response purposes, but also many challenges. These challenges include: First, the speed and volume at which data arrives, where tweets arrive continuously in chronological order. Second, the nature of "live" events produces a continuously changing dynamic corpus. Third, the significant amount of "noise" presented in the stream constitutes around 40% of all tweets, which have been reported as pointless "babbles" [3] or spam. Finally, each tweet is short (140 characters), which means they often lack the context that would assist text analysis.

The main task that we tackle in this paper is the ability to develop an algorithm to detect disruptive events and test the applicability of our algorithm to Arabic content posted to Twitter. Arabic is a rich Semitic language which is highly productive, both derivationally and inflectionally [2, 4]. The number of Arabic words is estimated to be 60 billion, derived from approximately 10,000 roots. Arabic poses many challenges for data mining tasks [2]. Most of these challenges are due to orthography and morphology. It is true that some of these challenges are shared with other languages but it exhibits considerable complexity from theoretical to computational linguistics. Furthermore, the language processing becomes even more challenging when considering the language used in social networking and microblogging sites, where dialects are heavily used. These dialects may differ in vocabulary, morphology, and spelling from the standard Arabic and most do not have standard spellings.

To overcome these challenges, we propose a novel event detection model that is language-independent. This model is based on frequency or co-occurrence of terms over time. Arabic event detection is enriched using automatically Named Entity Recognition, dictionaries, and Twitter features such as Retweet ratio and Hashtags.

Many researchers have proposed models and techniques for the purpose of identifying real-world events using social media data. In this paper, we propose an online classification-clustering framework, which is able to handle a constant stream of new documents with threshold parameters that can be modified in an experimental manner during training phase. The high volume of tweets from Twitter is the input of the system, which produces a table of the events in a particular region, associated sub-events (details) and *disruptive events* (as defined above) for a particular time (daily or hourly fashion). Social media data are very noisy; hence the first step in our framework after collecting data is preprocessing, which aims to reduce the amount of

noise before classification. The next step is to separate event-related tweets and non-event content. We implement a Naive Bayes machine classifier to achieve this. Then, we compute tweet features in order to extract similar characteristics and apply an incremental online clustering algorithm to assign each message in turn to a suitable event-based cluster by calculating each tweet's similarity to existing clusters, ultimately enabling us to detect a range of events. We focus in this work on real-world event identification for both large scale and rare (disruptive) events such as car accidents in a given location. Our contributions can be summarized as follows:

- Using our framework, we identify the relationship between Twitter activity and real-world events by detecting key events throughout the day;
- Using temporal, spatial and textual features, our framework is able to detect disruptive events at a given place for a particular time.
- Our framework is language independent as we address the challenging task of detecting events in Arabic.
- We validate our model on multiple real-world data sets to show the effectiveness of the framework.

The rest of the paper is organized as follows: Section 2 reviews related work on event detection in social media. In section 3, we discuss the main elements of our proposed framework. In section 4 we discuss several features; temporal, spatial and textual features. Section 5 presents our experiments and discusses the results. Finally, we conclude and highlight the future work of research in section 6.

2 Related Work

In the recent years, many researchers have shown interest in online event detection in social media. For instance, Petrovic et al. [11] presented an approach to detect breaking stories from a stream of tweets. The proposed approach, which is based on the locality-sensitive hashing (LSH), automatically organizes every incoming tweet in an existing story or labels it as a new story. In order to reduce the search space and improve the performance of the LSH, they added a secondary search, which indeed improves the results by19%. Using a different approach, Cordeiro [12] proposed a continuous wavelet transformation based on hashtag occurrences combined with a topic model inference using Latent Dirichlet Allocation (LDA). Instead of using individual words, hashtags are used to build wavelet signals. Wavelet peak and local maxima detection techniques are used to detect peaks in the hashtag signal. Then, LDA is applied to all tweets from the hashtag signal when an event is detected. However, these approaches do not differentiate whether topic detected is event-related or celebrity update. Non-event content such as personal or celebrity updates are not important to the decision-making process and may introduce noise.

Sakaki et al. [14] developed a probabilistic spatio-temporal model to monitor tweets and detect disastrous events such as earthquakes. Their method is based on features such as the keywords “Earthquake!” where they assumed that each user is regarded as a sensor with a function of detecting a target event and reporting it via

Twitter. One requirement of the approach is that to monitor an event we need to know the event in advance to provide representative keyword queries to be detected. This is an issue for detecting dynamic or unexpected events.

Becker et al. [21] proposed an online clustering framework, suitable for large-scale social media sites such as Twitter, to identify different types of real-world events. The online clustering technique groups together topically similar tweets and implements four features (Temporal features, Social features, Topical Features and Twitter-Centric Features) to distinguish between real-world events and non-events. Another study that stresses the importance of proper nouns identification to enhance the similarity comparison between tweets was presented by Phuvipadawat and Murata in [15]. Their method collected, grouped, ranked, and tracked breaking news from Twitter. Nevertheless, these two approaches are limited to widely discussed events and fail to report rare and potentially disruptive events. In addition, none of the aforementioned approaches have been shown to perform well with Arabic content.

The amount of research reported on Arabic information retrieval is considerably limited and immature compared to what is done in other less inflected languages. Most attention is focused on text classification, techniques used for language pre-processing like (stemmers and index tools), filtering and translation [2, 4]. Previous work on Arabic IR has used distance-based algorithms, Learning algorithms, Bayesian classification methods and N-grams for searching Arabic text documents [4].

3 Framework for Event Detection

Figure 1 illustrates our novel framework, which supports the automatic identification of events from social media. The five steps in the framework include; data collection, pre-processing, classification, on-line clustering and summarization. In this section we will explain each step in more detail.

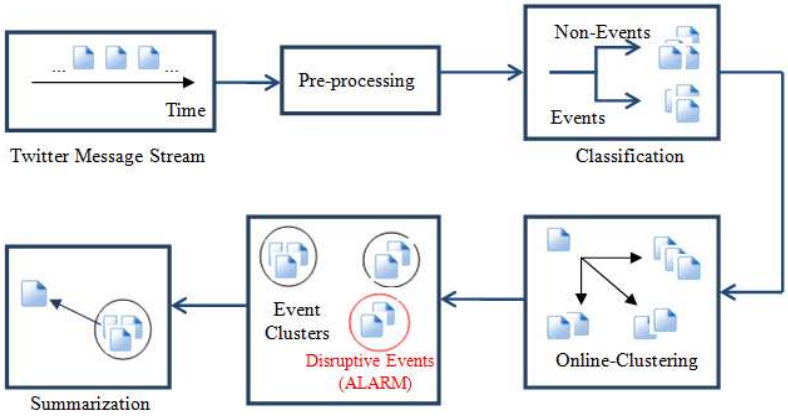


Fig. 1. Twitter Stream Event Detection Framework

3.1 Data Collection

We use Twitter's Streaming API to collect user-generated posts because it allows subscription to a continuous live stream of data. Our goal is to monitor and detect events (including disruptive events) in a given location without prior knowledge of these events. Thus, we collect tweets based on a set of keywords that generally describes a region (for example: Abu Dhabi) using different languages – Arabic and English. We also collect tweets from users who selectively add the required region as their location. In addition, we also make use of geographic Hashtags in the data collection process.

Data is stored using MongoDB [19], an open-source document database, which is easy to use and provides high availability speed and memory. MongoDB has been shown to be suitable for storing tweets, and supports different indices with straightforward queries [19].

3.2 Pre-Processing

The goal of the pre-processing step is to represent data in a form that can be analyzed efficiently and to improve the data quality by reducing the amount of trivial noise (i.e. deleting tweets that are irrelevant to events). We perform text processing techniques such as stop-word elimination (Term frequency and TF-IDF are the criterions used for classifying stop words) and stemming (Khoja stemmer for Arabic tweets [22] and Porter Stemming [25] for English and other Latin tweets). In addition to the Arabic stop word list included in the Khoja stemmer [22], we added to it more stop words which are determined using Term frequencies and TF-IDF of the training corpus. Moreover, posts that were less than 3 words long were removed and tweets with one word accounting for over half of the words are also removed, as these posts are less likely to have useful information.

3.3 Classification

This step aims to distinguish events from noise or irrelevant tweets. Words from each tweet are considered as features and a Naive Bayes classifier was chosen for the classification task over a number of leading methods such as support vector machines (SVMs) or Logistic Regression, due to its performance in previous extensive experiments as demonstrated in [1]. The main reasons for using Naïve Bayes model are; it is relatively fast to compute, easy to construct with no need for any complex iterative parameter estimation schemes. Unlike SVMs or Logistic Regression, Naïve Bayes classifier treats each feature independently. Naïve Bayes also tends to do less overfitting compared to Logistic Regression [1, 14].

We used the R statistical software package (http://www.R-project.org), specifically the e1071 R package, to build and train the Naïve Bayes Classifier on a training corpus of 1500 tweets that have been annotated as "event" or "non-event". Event instances outnumber the non-event ones as the training set consisted of 600 Non-Event tweets and 900 Event-related tweets.

The features and their corresponding category (event or non-event) are provided to the classifier and these constitute the training set. From the training data the likelihood of each tweet belonging to either class is derived based on the occurrence of the tweet's features in the training data. When a new example is presented, the class likelihood for the unseen data is predicted based on the training instances.

Algorithmic steps:

i. Input tweets.
ii. Extract features from tweets.
iii. These features and their corresponding labels are used to train the learning algorithm (Naive Bayes classifier).
iv. New tweets are presented to the trained classifier to predict their label using their extracted features.

3.4 Online-Clustering

The classification step separates event-related documents from non-event posts (such as chats, personal updates, spam, incomprehensible messages). Consequently, non-event posts are filtered. To identify the topic of an event, including determining those that are disruptive events, we define a range of features including temporal, spatial and textual features, which are detailed in the next section. We then apply an online clustering algorithm, which is outlined in Algorithm 1.

Input:
n set of documents $(D_1,...,D_n)$
Threshold τ
Output:
k clusters $(C_1,...,C_k)$
Step 1: **For** a given τ, compute the centroid similarity function $E(D_i, c_j)$ of each cluster c_j
Step 2: **If** centroid similarity $E(D_i, c_j) \geq \tau$ **do**:
1) A new cluster is formed containing D_i
2) The new centroid value = D_i
Step 3: **If** centroid similarity $E(D_i, c_j) < \tau$ **do**:
1) Assign it to cluster which gives maximum value of $E(D_i, c_j)$
2) Add D_i to cluster j and recalculate the new centroid value c_j.

Algorithm 1. Online Clustering Algorithm

Using set of features $(F_1,...,F_k)$ for each document (tweet) $(D_1,...,D_n)$ we compute a similarity measure $E(D_i, c_j)$ between the document and each cluster $(C_1,...,C_k)$ where similarity function is computed in turn against each cluster c_j for $j=1,...,m$ and m is the number of clusters (initially $m=0$). In this paper, we use **the average** weight of each term across all documents in the cluster to calculate the centroid similarity function $E(D_i, c_j)$ of a cluster. The threshold parameters are determined empirically in the training phase.

The decision to use online clustering algorithm was taken for three main reasons: (i) it supports high dimensional data as it effectively handles the large volume of social media data produced around events; (ii) many clustering algorithms such as K-means require the prior knowledge of the number of clusters. As we do not know the number of events and sub-events *a priori* the online clustering is suitable as it does not require such input; (iii) partitioning algorithms are ineffective in this case because of the high and constant sheer scale of tweets [21].

3.5 Summarization

After clustering tweets into clusters, the next natural step would be to automatically summarize or represent topics being discussed within clusters. Each cluster may contain hundreds of tweets, and the task of finding most representative tweets or extracting top terms (topics) is essential to support the identification of events, especially disruptive events, so any potential security and safety issues can be managed. Summarization task is a very challenging task in its own and takes various forms [23]. The simplest approach is to consider each tweet as a document, and then apply a summarization method on this corpus to capture its key features [17, 21, 23]. Voting algorithms [17] are utilized in applications where in the context of microblogging sites take into account the following:

- The average length of a tweet;
- The total frequency of features in a tweet;
- Number of retweets, favorites and mentions;
- The inclusion of multimedia contents such as images.

In this paper, we implement a voting approach where the highest number of retweets in a cluster is used as a criterion for the summarization task. However, we leave the improvement of multilingual summarization of microblogs for future work.

4 Feature Selection

Many researchers have proposed enhancements to models or developed new approaches to optimize the capturing of patterns in the input signals. Here, we introduce several features related to the Twitter in order to reveal characteristics of clusters that are associated with rare real-world events particularly disruptive events.

4.1 Temporal Features

Temporal features are important factors that have been overlooked in many event detection studies using in social media. The volume of tweets, and the continually updated commentary around an event suggests that informative tweets from several hours ago may not be as important as new tweets [21]. For this reason we retain the most frequently occurring terms a cluster in hourly time frames and compare the number of tweets posted during an hour that contain term t to the total number of

tweets posted during that hour. This helps identify terms that enable event clustering and also helps ordering events [8, 11, 14].

4.2 Spatial Features (Geospatial, Regional)

Events are characterized by rich set of spatial and demographic features [1]. In this paper, we make use of three statistical location approaches to extract geographic content from clusters. The first one is from Twitter where the source latitude and longitude coordinates are provided by the user. The second method depends on the shared media (photos and videos) by using the GPS coordination of the capture device (if supported). Third, Open NLP (http://opennlp.sourceforge.net) and Named-Entity Recognition (NER) were implemented for geotagging the tweet content (text) to identify places,, organization, street names, landmarks etc. These approaches rely purely on Twitter with no need for user IP, private login information, or external knowledge bases which give the maximum advantage [5, 24].

Once the geographic content is extracted from each tweet in a cluster, we aggregate them to determine the cluster's overall geographic focus. The higher the volume of tweets from nearly near coordinates, the higher the level of confidence in the location of the event will be. Table 1 presents a disruptive event (loss of communication) happening in the F1 event (from the first dataset) where spatial features are used to determine the cluster (event) overall location (Yas Marina).

Date	Time	User	Original tweet	Translated tweet	RT
04/11/2013	20:13:04	PJoc31		Having problem calling my friends using du in **Yas Island Rotana** hotel #**AbuDhabi** #F1 Grand Prix: The **Yas Marina** Circuit	5
04/11/2013	20:16:41	M7mdAS96	**ياس مارينا** مكان خيالي لكن ما عرف شو مشكلة الاتصال والاشارة دووووم ضعيفة. بليز ساعدوني ضروري#F1 #**AbuDhabi**	The **Yas Marina Circuit** is an awesome venue however I am having trouble with communication and coverage signal. please help #F1 #**AbuDhabi**	2
04/11/2013	20:23:12	BintZayed91	كان الاتصال ممتاز في فترة الظهر ما عرف شو ياها دو من ربع ساعة أحاول اتصل او ارسل رسالة ماشي فايدة **شارع ياس بلازا** قريب #F1 **فندق روتانا #ياس**	Connection was excellent at noon Don't know what happened with Du signal as I am trying to make a call or send sms from quarter of an hour with no success **Plaza st** near **Yas Rotana hotel** #**Yas** #F1	9

Table 1. Spatial features are extracted (bold) from user's tweet to determine the cluster's overall location.

We assume that all locations provided by users are correct however [6] found that 34% of Twitter users had entered fake locations in their profile. Some users may intentionally misrepresent their home location either to cover for their actual location, or for privacy-security issues. On the other hand, some users provided location may differ from their actual location because their locations change frequently due to travel. The virtual sense of community should also be taken into consideration.

4.3 Textual Features

Textual or content features have been identified as contributing to the spread of a post in social media [13]. For example, hashtags are used to generate content features [7, 8], and identify topics affecting retweet likelihood [5, 8, 13, 26]. Here, we introduce the features we derived from tweet text.

Near-Duplicate measure.

The average content similarity over all pairs of tweets posted in a cluster (1-hour) is calculated using:

$$\sum_{a,b \in set\ of\ pairs\ in\ tweets} \frac{similarity\ (a,b)}{|set\ of\ pairs\ in\ tweets|}$$

where the content similarity is computed using the standard cosine similarity over words from tweet a, b vector representation $\overrightarrow{V}(a), \overrightarrow{V}(b)$ of the tweet content:

$$similarity\ (a,b) = \frac{\overrightarrow{V}(a).\overrightarrow{V}(b)}{\left|\overrightarrow{V}(a)\right|\left|\overrightarrow{V}(b)\right|}$$

If the two tweets have a very high similarity, we assume that one of them is a near-duplicate of the other. The original tweet is considered as the first tweet in a particular time frame and/or the shortest tweet in length. Even though, duplicates are less likely to provide additional information about an event, several users independently witnessing an event and tweeting about it would effectively increase the confidence level of an event. An example of tweets with high near-duplicate measure is presented in Table 2.

Date	Time	User	Original tweet	Translated tweet	RT
02/11/2013	6:09:52	hazza_saiff	صباح الخير ..**ضباب** علي **خط** #**ابوظبي العين**	Good morning.. **fog** on #**AbuDhabi alain highway**	2
02/11/2013	6:11:24	BuHazae	**ضباب كثيف** على **خط العين ابوظبي Net_AD@**	**Net_AD@ Heavy fog** on **abu dhabi alain highway**	**3**
02/11/2013	6:12:53	Rose_alduwaila	ارجوا الانتباه **ضباب خط العين ابوظبي** http://t.co/z0siijm WLC	Attention please **fog** on **AbuDhabi alain highway** http://t.co/z0siijm WLC	7
02/11/2013	6:12:58	mzinelsawari	**#برق الامارات ضباب كثيف** في **خط ابوظبي** قبل الخزنه	**#Uaebarq heavy fog** on **abu dhabi highway** before Alkhazna	**4**
02/11/2013	6:14:11	GroupStorms	**ابوظبي ضباب كثيف على خط #ابوظبي_العين**	**Abu Dhabi heavy fog** on #**abudhabi_alain highway**	3
02/11/2013	6:19:23	WALEED625	تنبيه: **ضباب كثيف** على مختلف طرق الخارجية إمارة #**ابوظبي** وبالذات **خط ابوظبي العين** نتمنى من الأخوة أخذ الحيطة	Attention: **heavy fog** on various external ways of #**Abu Dhabi** and **Abu Dhabi alain highway** in particular please brothers take extra caution	0

Table 2. Severe weather alarm from tweets on the 2nd of November 2013

Retweet ratio.

Retweet represent the influence of a tweet beyond one-to-one interaction domain. Popular tweets could propagate multiple hops away from the source as they are retweeted throughout the network [7]. Hence, the number of retweets is an indication of popularity. Furthermore, retweeting in a social network can serve as a powerful tool to reinforce a message when not only one but a group of users repeat the same message [7, 8]. Therefore, retweet ratio indicates tweets surrounding an event where users agree with the message or wish to spread the information (warning, advice, evidence…) with other users. Retweet ratio has been implemented to detect events and to estimate rumors in social media stream [18]. We calculate this attribute by normalizing number of times a tweet appears in a timeframe to the total number of tweets in that timeframe.

Mention ratio.

A mention is a mechanism used in Twitter to reply to users, engage others or to join a conversation in a form of (@username). A user can mention one or more users anywhere in the body of the post. Hence, we calculate the number of mentions (@) relative to the number of tweets in a cluster. Ordinary users show a great passion for celebrities and as a result the most mentioned users are celebrities where sometimes users mention them without necessarily reading their posts [7, 13]. Regarding events reporting, users tend to mention journalists, politicians and official accounts such as news agencies or government official accounts to drive their attention about an event or to add more credibility to their event-related posts.

Hashtag ratio.

Hashtags are an important feature of social networking sites and can be inserted anywhere within a message. Some Hashtags indicate their posted messages (#bbcF1) and some others are dedicated originally to events such as (#abudhabigp). In addition, topic related hashtags are used as an information seeking index on Twitter to search Twitter for more tweets belonging to a topic. The use of hashtags became a coordinating mechanism for disruptive-related activity on Twitter [14, 20]. The Hashtag ratio is the ratio of tweets containing hashtag over the total number of tweets in that timeframe.

Link or Url ratio.

As Twitter is limited to 140 characters per message it is common in the Twitter community to include links when tweeting to share additional information or for referencing. Clusters that have tweets with links from popular websites (news agencies or government sites) may boost level of confidence of that information and hence more adoption to such tweets and clusters. Not all links refer to officials but mostly they are images or videos uploaded by users. Additionally, the co-occurrence of URLs in a cluster confirms that these tweets refer to the same event and improves the level of confidence of an event. This attribute is calculated by the fraction of tweets with URL to the total number of tweets in a timeframe.

Tweet sentiment.

Users express their opinions on a variety of topics in Twitter. They might discuss news, complain about services and express positive or negative sentiment about products [9, 10]. In fact, companies manufacturing such products have developed techniques to analyze these posts to get a sense of sentiment about their products [10].

In prior work, we found that negative sentiment is usually associated when reporting disruptive events (Negative overall cluster).The sudden change of tweets' sentiment is another observed characteristic of a disruptive event cluster. Here we focus on negative sentiment regarding identifying disruptive events, given that negative sentiment tweets are more likely to be retweeted as shown in [6, 8, 9]. We use a semantic classifier based on the SentiStrength model in [9].The SentiStrength algorithm is suitable because it is designed for short informal text with abbreviations and slang. Furthermore, it combines a lexicon-based model with a set of additional linguistic rules for spelling correction, negations, booster words (e.g., very), emoticons, and other factors. Most importantly, SentiStrength support multiple languages including Arabic.

Dictionary-based feature

One main objective of our framework is the ability to automatically detect messages that contain precise information about disruptive events such as labor strike or fire incidences. To enrich such rare event identification, present tense verbs, popular event nouns and adjectives that describe events as they take place are considered as a feature. This bag of words model uses a dictionary of trigger words to detect and characterize events which are manually labeled by experts from several management departments such as traffic control department, crises departments, emergencies and others.

Examples of present verbs are: witness, notice, observe, participate, engage, listen etc. Examples of event nouns and adjectives are; live, urgent, breaking news, latest, update etc.

5 Experimental Evaluation

5.1 Experimental Setup

Data: Our first dataset, which consists of around 1.7 Million tweets (1698517), was collected from 15 October 2013 to 05 November 2013 using Twitter's Streaming API. Our initial aim was to monitor and analyze disruptive events associated with major events in a particular region. We chose the Formula 1 Motor Racing, which was hosted in Abu Dhabi (our input location) between 1st and 4th November 2013. The number of Arabic tweets is 890658 where English tweets are 39191. Around 24% of tweets were published in other Latin script and other languages. Figure 2 shows the language distribution in our first dataset. As our task focuses on Arabic event detection, we restrict our dataset to Arabic tweets and eliminate all non-Arabic tweets.

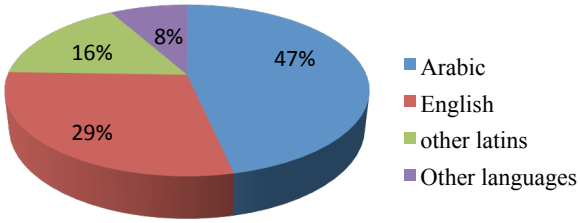


Fig. 2. The distribution of languages used in our dataset

Since then we focused our attention on collecting tweets for the purpose of analyzing disruptive events in the capital Abu Dhabi. In this work, we restrict our search to Arabic tweets. A considerable change of tweets volume was noticed from 2nd to 5th December 2014 due to the famous double-crime (considered as a terrorist attack) on the 2nd December 2014 which was unprecedented in the peaceful Abu Dhabi history. An American woman was murdered in a shopping mall. The second crime was held by the same suspect when she planted a primitive bomb on the doorstep of an American citizen in a different location. The second dataset consists of 1161854 Arabic tweets. Figure 3 shows the tweets volume in Abu Dhabi which clearly indicates the rise of posts' volume and discussions during the terrorist attack.

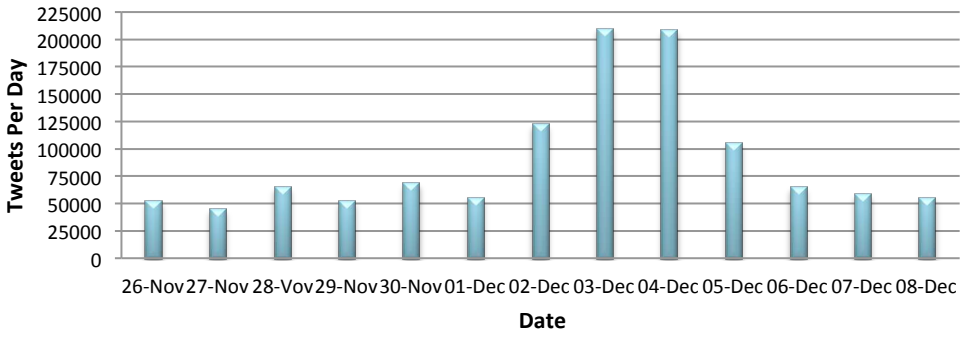


Fig. 3. The volume of tweets in the second data set from (26th Nov to 8th Dec) in Abu Dhabi

Annotation: To evaluate the framework, we evaluate the two main stages: classification and clustering. For classification, three human annotators manually labeled 1200 tweets in to two classes "Event" and "Non-Event" to train our classifiers (500 Non-Event tweets and 700 Event-related tweets). The agreement between our three annotators, measured using Cohen's kappa, was substantial (kappa = 0.807).

The resulting dataset after classification contained approximately 62,000 event-related tweets which we used to train, test and evaluate the clustering algorithm. We used the first 15 days of data (from 15/Oct until 29/Oct from the first dataset) to train the clustering algorithm and to tune the thresholds using the validation set. Then we tested the clustering algorithm on unseen data of the last 6 days from the 30th of Oct

until the 4th of Nov. Threshold values were varied from 0.10 to 0.90 at graded increments of 0.05% with a total of 17 tests in order to find the best cut-off of τ =0.55 (77 character difference). Figure 4 illustrates the F-measure for different thresholds where the best performing threshold τ =0.55 seems to be reasonable because it allows some similarity between posts but does not allow them to be nearly identical.

In order to evaluate the clustering performance, we employed three human annotators to manually label 637 clusters based on the highest number of retweets a post gets to represent a cluster. The task of the annotators was to choose one of the eight different categories: politics, finance, sport, entertainment, technology, culture, disruptive event and others. The agreement between annotators was calculated using Cohen's kappa (K=0.772) which indicates an acceptable level of agreement. We used only **492 clusters** on which all annotators agreed as the **gold standard**.

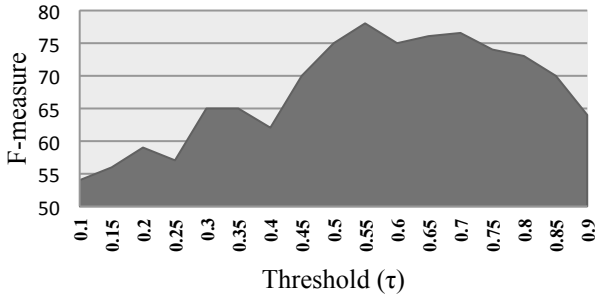


Fig. 4. F-measure of online clustering over different thresholds

5.2 Evaluation Matrices

To measure the effectiveness of classifiers based on our proposed features, we used a set of well-known classification metrics: precision, recall, accuracy, and F1 measure. Precision is how often are our predictions of a class are correct —a measure of false positives. Recall is how often tweets are classified correctly as the correct class — a measure of false negatives. F-measure is a harmonic mean of precision and recall. Accuracy is the proportion of the correctly classified tweets to the total number of tweets. A false positive is when the outcome is incorrectly predicted as X class when it is actually Y class. A true positive is when actual X class events are correctly predicted as X class events.

$$\text{Precision(P)} = \frac{\text{tp}}{\text{tp+fp}} \qquad \text{Recall(R)} = \text{True positive rate} = \frac{\text{tp}}{\text{tp+fn}}$$

$$\text{F} - \text{measure} = \frac{2\times\text{P}\times\text{R}}{\text{P+R}} \qquad \text{False positive rate} = \frac{\text{fp}}{\text{tn+fp}}$$

$$\text{Accuracy} = \frac{\text{tp} + \text{tn}}{\text{tp} + \text{fp} + \text{fn} + \text{tn}}$$

To evaluate the quality of clusters we compute average cluster precision (AP) [16] on the gold standard. The average precision measures how many of the identified clusters

are correct averaged over hours per day and calculated based on the precision of each cluster per hour per day. Average precision is a common evaluation metric in tasks like ad-hoc information retrieval where only the set of returned documents and their relevance judgments are available [1, 16, 20, 21].

5.3 Experimental Results

To evaluate the overall framework, we have to evaluate the two main elements. Starting with Classification: we found in [1] that the Naive Bayes classifier outperformed other machine learning algorithm (SVMs classifier and Logistic Regression) in classifying events using the English language. Furthermore, Naive Bayes classifier achieves better results using combination of attributes (Unigrams+ Bigrams+ part-of-speech (POS) + Named Entity Recognition (NER)) with F-measure value of 85.43%. Here we repeat the same experiment comparing the same machine learning algorithms but with only Arabic input and the new annotation. A ten-fold cross validation approach is adopted to train and test the methods using the WEKA machine learning toolkit for the classification task. Table 3 gives the F-measure results of the three machine learning algorithms using combination of attributes.

	Naive Bayes classifier	SVMs classifier	Logistic Regression classifier
F-measure	80.24	78.53	76.85

Table 3. F-scores of different classification algorithms

We obtain similar results to [1] as the Naïve Bayes classification method outperforms others. There is an overall drop in the performance of all three methods, which we expected due to the limitation of the used attributes. For example, Part-of-speech (POS) and Named Entity Recognition (NER) are very limited for Arabic language.

In order to evaluate the clustering performance, we used similar techniques to [1, 16, 21]. Average precision is calculated with respect to eight categories: politics, finance, sport, entertainment, technology, culture, disruptive event and other-event. Table 4 shows the average precision percentages of clusters in the test set.

Date	Politics	Finance	Sport	Entertainment	Technology	Culture	Disruption Events	Average Per Day
30-Oct	83.26	82.19	79.50	78.64	73.20	75.93	82.35	79.30
31-Oct	81.34	82.47	85.33	69.91	72.37	77.43	80.58	78.49
⋮								
4-Nov	79.75	81.86	81.93	79.38	80.46	81.51	83.02	81.13
Average Per Topic	81.39	80.62	79.57	73.23	76.13	77.54	82.26	78.68

Table 4. Average precision of the online clustering algorithm, in percent.

While the online clustering algorithm achieves a good performance, the results are sometimes inconsistent with respect to topics. Not surprisingly, the average precision

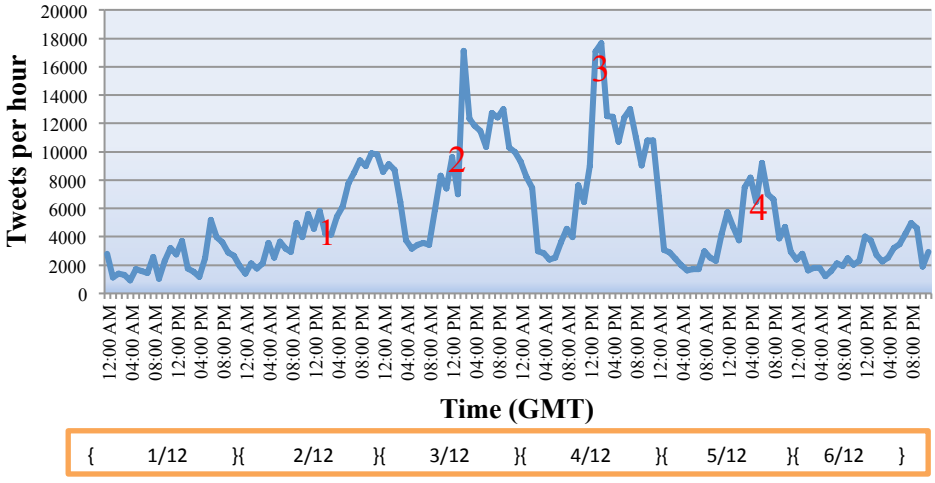
of identifying political events is greater than the average precision of identifying entertainment related events by about 9%. Since it is easier to extract and categorize events like politics, finance, sport and disruptive events than events like entertainment, technology or cultural events even for humans which cause the main disagreement between annotators in the annotation task. Finally, it is important to notice that the framework is able to automatically identify disruptive events with the best performance of 82.26%.

One of the frameworks' objectives is to identify disruptive events and send a notification to the administrators. Table 5 shows the top 3 emerging disruptive events identified by the framework based on the number of retweet counts for the second dataset. For space limitation, we only present results of the disruptive incidents on the 2nd of Dec as an example of the system's output. The system can produce results with different level of time granularity (per hour, 3 hours, …, per day).

Date	User	Tweet	Translation	RT
Dec 2	AbuDhabiPolice	مشاجرة في دورة مياه تسفر عن مصرع سيدة بجزيرة الريم http://www.securitymedia.ae/ar/media.center/News/4202109.aspx	Woman Dies after Public Toilet Fight on Reem Island http://www.securitymedia.ae/ar/media.center/News/4202109.aspx	76
	Mona_Alraesi	حريق ضخم في محطة لتوزيع الكهرباء في ابوظبي بالقرب من مصفح الصناعيه ونسال الله السلامة للجميع pic.twitter.com/kLLc4L0hoJ	A huge fire in an electricity distribution station in Abu Dhabi near musaffah industrial area we ask God for everyone's safety	49
	NET_AD	أبوظبي الان : حادث تدهور على خط دبي_بوظبي بعد محطة السمحه مع وجود اصابات... نرجو أخذ الحيطه والحذر	Abu Dhabi now: there is a multiple car crashes on the Abu Dhabi_Dubai highway after Alsamha petrol station with several injuries ... please take caution	22

Table 5. Top 3 emerging disruptive events identified by the system on the 2nd of Decmber 2014.

To provide further validation for our system, we evaluated it using the second dataset which contains more disruptive events than the first dataset. We were able to compare our disruptive event identification results with the official record of events, as the authorities released 2 videos on Youtube with the exact time of these events (shown in Figure 5). All of these events were detected successfully by the framework. Figure 6 shows the clustering output of two time-frames (2-3PM on 2nd of Dec and the same time of the next day 3/12/2014). The results suggest that the number of disruptive events (clusters in the red) increased dramatically over the same period from previous day as people discussed the murder.



5. An American woman was murdered in a Shopping mall in Abu Dhabi. (Based on the CCTV which was released by the officials on Youtube. Time of the crime between 1:12pm-2:45pm on the 2nd of Dec)
6. The Ministry of Interior released CCTV (On the 3rd of Dec at 12pm) footage of the suspect "Reem Island Ghost" and ask public for information.
7. Abu Dhabi Police reveal the second video (on youtube on the 4th pf Dec at 1pm) which contains the double-crime, search, inspection procedures and the arrest of the suspect.
8. Minister of Interior made the announcement at a press conference about Reem Island Crime and that the suspect has been arrested (5 Dec at 3pm).

Fig. 5. The volume of tweets in the second dataset from (1st Dec to 6th Dec) in Abu Dhabi with the main events detection.

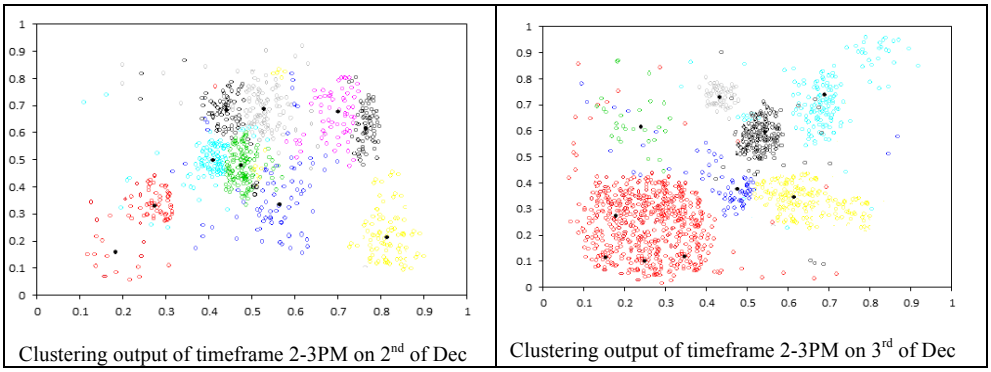


Fig. 6. The clustering output of two time-frames (2nd - 3rd/Dec/2014)

6 Conclusions and Future Work

In this paper we have presented an integrated framework to detect real-world events in Arabic from social media platform (Twitter). The event identification was performed through several stages; data collection, preprocessing, classification, clustering and summarization. We have also shown that our approach is able to reveal disruptive events for a certain location using rich set of features. Extensive experiments were conducted to evaluate the effectiveness of the proposed framework using two real-world datasets.

This framework can be generalized to develop a social awareness system or for the purposes of decision making enrichment which can be implemented in many fields such as crises management or information intelligence. Our results support the claim that the use of social media for the purposes of information gathering could be utilized as a complementary to traditional intelligence and not to be used independently. In future we aim to compare our results with other works in the area of event detection on Twitter. This is a challenge due to the differences between datasets as each dataset has different size, time and characteristics. We also aim to validate our results against real-time complete official reports or official news streams.

There are many directions for future work. One of the main directions is to compare and validate the performance of the proposed framework against other well-known algorithms such as the state-of-the-art Labeled Dirichlet Allocation (LDA) method. Another direction is to study the contributions and limitations of various feature types to event detection in social media. Finally, detection of rumors in social media with deep analysis of the distinctive characteristics of rumors and the way they propagate in the microblogging communities will be carried out in the near future.

7 References

1. Alsaedi, N., Burnap, P. and Rana, O. 2014. A Combined Classification-Clustering Framework for Identifying Disruptive Events. Proceedings of 7th ASE International Conference on Social Computing (SocialCom 2014), pp. 1–10. DOI= http://ase360.org/handle/123456789/71
2. Darwish, K. and Magdy, W. 2014. Arabic Information Retrieval. Foundations and Trends® in Information Retrieval 7, pp. 239–342. DOI=http://www.nowpublishers.com/articles/foundations-and-trends-in-information-retrieval/INR-031.
3. PearAnalytics. Twitter study - august 2009. http://www.pearanalytics.com/wpcontent/uploads/2009/08/Twitter-Study-August-2009.pdf, 2009.
4. Larkey, L., Ballesteros, L. and Connell, M. 2007. Light stemming for Arabic information retrieval. Arabic Computational Morphology, pp. 221–243.
5. Cheng, Z., Caverlee, J. and Lee, K. 2010. You are where you tweet: a content-based approach to geo-locating twitter users. Proceeding CIKM ’10 pp. 759–768. DOI= http://dl.acm.org/citation.cfm?id=1871535
6. Hecht, B., Hong, L., Suh, B. and Chi, E. 2011. Tweets from Justin Bieber’s heart: the dynamics of the location field in user profiles. Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, pp. 237–246.

7. Cha, M., Haddadi, H., Benevenuto, F. and Gummadi, P. 2010. Measuring User Influence in Twitter: The Million Follower Fallacy. ICWSM-2010.
8. Ma, Z., Sun, A. and Cong, G. 2013. On predicting the popularity of newly emerging hashtags in twitter. Journal of the American Society for Information Science and Technology 64(7), pp.1399-1410.
9. Thelwall, M., Buckley, K. and Paltoglou, G. 2011. Sentiment in Twitter events. Journal of the American Society for Information Science and Technology 62(2), pp. 406–418.
10. Agarwal, A., Xie, B., Vovsha, I., Rambow, O. and Passonneau, R. 2011. Sentiment analysis of twitter data. Proceedings of the ACL 2011 Workshop on Languages in Social Media, pp. 30–38.
11. Petrović, S., Osborne, M. and Lavrenko, V. 2010. Streaming first story detection with application to twitter. Annual Conference of the North American Chapter of the Association for Computational Linguistics, pp. 181–189.
12. Cordeiro, M. 2012. Twitter event detection: combining wavelet analysis and topic inference summarization. Doctoral Symposium on Informatics Engineering, DSIE'2012.
13. Cheng, J., Adamic, L., Dow, P., Jon, K. and Jure, L. 2014. Can cascades be predicted? WWW '14. DOI= http://dl.acm.org/citation.cfm?id=2567997
14. Sakaki, T., Okazaki, M. and Matsuo, Y. 2010. Earthquake Shakes Twitter Users : Real-time Event Detection by Social Sensors. 19th International World Wide Web Conference (WWW '10).
15. Phuvipadawat, S. and Murata, T. 2010. Breaking news detection and tracking in Twitter. Proceedings - 2010 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology - Workshops, WI-IAT 2010, pp. 120–123.
16. Bollmann, P. 1977. A comparison of evaluation measures for document retrieval systems. Journal of informatics (1977), pp. 97-116.
17. Bauer, E. and Kohavi, R. 1999. An empirical comparison of voting classification algorithms: Bagging, boosting, and variants. Machine learning 38 (1998)
18. Takahashi, T. and Igata, N. 2012. Rumor detection on twitter. SCIS '6 and ISIS '13, pp. 452–457.
19. Kumar, S., Morstatter, F. and Liu, H. 2014. Twitter Data Analytics. Springer
20. Dou, W., Wang, X., Skau, D., Ribarsky, W. and Zhou, M.X. 2012. LeadLine: Interactive visual analysis of text data through event identification. (VAST 2012), pp. 93–102.
21. Becker, H., Naaman, M. and Gravano, L. 2011. Beyond Trending Topics: Real- Event Identification on Twitter. ICWSM, pp. 1–17.
22. Khoja, S., Garside, R. and Knowles, G. 2001. Stemming arabic text. NAACL2001.
23. Chua, F. and Asur, S. 2012. Automatic Summarization of Events from Social Media. ICWSM-2013.
24. Mahmud, J., Nichols, J. and Drews, C. 2012. Where Is This Tweet From? Inferring Home Locations of Twitter Users. ICWSM, pp. 511–514.DOI= http://www.aaai.org/ocs/index.php/ICWSM/ICWSM12/paper/viewFile/4605/5045
25. Porter, M. An algorithm for suffix stripping. Program: electronic library & information systems 40 (3), pp. 211 – 218.
26. Burnap, P., Williams, M.L., Sloan, L., Rana, O., Housley, W., Edwards, A., Knight, V., Procter, R. and Voss, A. (2014), 'Tweeting the Terror: Modelling the Social Media Reaction to the Woolwich Terrorist Attack', *Social Network Analysis and Mining 4:1*