EveTAR: Building a Large-Scale Multi-Task Test Collection over Arabic Tweets

Maram Hasanain · Reem Suwaileh · Tamer Elsayed · Mucahid Kutlu · Hind Almerekhi

Abstract This article introduces a new language-independent approach for creating a large-scale high-quality test collection of tweets that supports multiple information retrieval (IR) tasks without running a shared-task campaign. The adopted approach (demonstrated over Arabic tweets) designs the collection around *significant* (*i.e., popular*) *events*, which enables the development of topics that represent frequent information needs of Twitter users for which rich content exists. That inherently facilitates the support of multiple tasks that generally revolve around events, namely event detection, ad-hoc search, timeline generation, and real-time summarization. The key highlights of the approach include diversifying the judgment pool via interactive search and multiple manually-crafted queries per topic, collecting high-quality annotations via crowd-workers for relevancy and in-house annotators for novelty, filtering out low-agreement topics and inaccessible tweets, and providing multiple subsets of the collection for better availability. Applying our methodology on Arabic tweets resulted in *EveTAR*, the first freely-available tweet test collection for multiple IR tasks. *EveTAR* includes a crawl of 355M Arabic tweets and covers 50 significant events for which about 62K tweets were judged with substantial average inter-annotator agreement (Kappa value of 0.71). We demonstrate the usability of *EveTAR* by evaluating existing algorithms in the respective tasks. Results indicate that the new collection can support reliable ranking of IR systems that is comparable to similar TREC collections, while providing strong baseline results for future studies over Arabic tweets.

Keywords Twitter; Microblogs; Event Detection; Ad-hoc Search; Real-time Summarization; Timeline Generation; Dialects; Evaluation

1 Introduction

Since the start of the Arab Spring in 2011, the Arab world has faced a continuous stream of events that changed the way Arab people are using online micro-blogging services such as Twitter. Several key events were initiated over those platforms and real-time updates were more frequent there than in typical mainstream media (Bruns et al. 2013; Wasike 2013). With an average of 19.7M Arabic tweets posted every day as of March 2016 (Salem 2017), the Arab users became more interested in several online tasks that are related to events. Among those tasks, they might be interested in automatically detecting events as soon as they occur (Alsaedi et al. 2016)

Maram Hasanain, Reem Suwaileh, Tamer Elsayed, Mucahid Kutlu
Department of Computer Science and Engineering, College of Engineering, Qatar University, Qatar
e-mail: maram.hasanain,reem.suwaileh,telsayed,mucahidkutlu@qu.edu.qa

Hind Almerekhi
Hamad Bin Khalifa University, Qatar
e-mail: halmerekhi@qf.org.qa

(*Event Detection*), or once an event occurs, they become interested in searching for tweets that are discussing that event (Darwish et al. 2012) (*Ad-hoc Search*), or getting real-time updates about the events as they are developing over time (Magdy and Elsayed 2016) (*Real-time Summarization*), or even asking for a timeline of the event when it concludes (*Timeline Generation*). That makes the need for automatic tools that can effectively search huge tweet stream, detect events before being publicly announced, and summarize their timelines obvious and imperative.

To design and develop high-quality Information Retrieval (IR) systems over Twitter, evaluation mechanisms such as *test collections* are evidently required. A test collection typically consists of a set of documents (e.g., tweets), topics representing information needs (e.g., events), and relevance judgments (i.e., labels) that specify which documents are relevant to the topics (Sanderson 2010). There are several publicly-available test collections over Twitter that support a broad range of retrieval tasks (Ounis et al. 2011; Soboroff et al. 2012; Lin and Efron 2013; Lin et al. 2014, 2015, 2016) and hence motivated a large body of research; however none of those collections is in Arabic, which severely hinders the advancement of research on Arabic tweets. Moreover, the majority of those collections were constructed using shared-task evaluation campaigns, which benefit from the participation of multiple research teams that contribute to the pool of judged documents, in addition to the employment of experienced dedicated annotators; that is costly and not easily feasible in many languages of relatively-low IR research resources such as Arabic.

In this article, we address the problem of building a large-scale high-quality test collection of tweets that supports multiple IR tasks without running a shared-task campaign. To tackle the problem, we adopted a new language-independent approach that designs the collection around *significant* events. A significant event is defined as an occurrence that happens at a particular time in a specific location and is discussed by the media (e.g., covered by an online news article). Focusing on significant events has several advantages. First, it enables the development of topics that represent frequent information needs of Twitter users as they often search Twitter for timely and up to date information about events (Teevan et al. 2011; Zhao and Mei 2013). Second, rich content exists for popular events. Third, the popularity of the topics might make the annotations much easier and thus consistent. Finally, it inherently facilitates the support of multiple tasks that generally revolve around events.

Applying our approach on *Arabic* tweets resulted in **EveTAR**, our **Eve**nt-centric **T**est Collection of **Ar**abic Tweets[1]. *EveTAR* is built over a one-month crawl of 355M Arabic tweets and supports four IR tasks: event detection, ad-hoc search, timeline generation, and real-time summarization. The topics cover 50 events in that month selected using Wikipedia's Current Events Portal[2]. A diversified judgment pool was constructed via multiple queries per topic that are manually-crafted using interactive search. To ensure large-scale but high quality judgments, we first employed crowd-workers to collect relevance judgments, then we filtered out the topics with the lowest inter-annotator agreement and removed inaccessible tweets, reaching an average *substantial* agreement indicated by Kappa value of 0.71 over 62K judgments. As two of the supported tasks require the identification of novel tweets per topic, we finally added novelty annotations via in-house annotators to ensure the integrity of the annotations.

We demonstrate the usability of *EveTAR* by evaluating a number of strong existing techniques in each of the supported tasks. The performance results of those systems constitute reference baselines for future studies in the respective research problems. Moreover, we show that ranking those systems over *EveTAR* is at least as reliable as in comparable TREC collections.

With the huge size of the collection and the sharing restrictions mandated by Twitter, we have provided a smaller sample (about 15M tweets) of the collection that is much easier to crawl and showed that it is as reliable as the full collection in ranking the tested systems. Furthermore, two subsets of that sample that include only Modern Standard Arabic (MSA) tweets and only dialectal tweets are also provided to encourage IR research on Arabic dialects, which naturally exhibit different unique characteristics from MSA and are widely used by Arab users in social media platforms (Darwish et al. 2012; Darwish and Magdy 2014). Performance results of the evaluated systems are presented on all subsets of *EveTAR* as well as the full set (whenever possible).

Our contribution in this work is 5-fold:

1. We propose a language-independent approach to build a multi-task test collection without the need of a shared-task campaign.
2. We introduce *EveTAR*, the *first* large-scale test collection over *Arabic* tweets that supports event detection, ad-hoc search, timeline generation, and real-time summarization IR tasks.

---

[1] pronounced as evi-tar.

[2] https://en.wikipedia.org/wiki/Portal:Current_events

3. We make the full test collection publicly-available for research[3], including ids of 355M Arabic tweets, 50 events and 62K relevance judgments, novelty annotations, inter-annotator agreements, queries used to identify potentially-relevant tweets for the events, and documented design of the crowdsourcing tasks. We also provide the annotations per tweet to further support relevant studies on crowdsourcing in IR.
4. We introduce four versions of *EveTAR* that make accessibility of the collection easier and encourage research over MSA vs. dialectal tweets.
5. We demonstrate the usability of *EveTAR* by evaluating existing techniques over the four supported tasks. The resulting performance scores constitute reference baselines for future studies.

The remainder of the article is organized as follows. We first discuss the related studies on building Arabic and tweet test collections in Section 2. We present our proposed approach in Section 3, and then we show how it is applied in practice on Arabic tweets (resulting in *EveTAR*) in Section 4. Section 5 covers the evaluation of *EveTAR* from usability and reliability perspectives. Finally, Section 6 concludes our study and gives some guidelines for future work.

2 Related Work

As our test collection is over Arabic tweets, in this section we review existing Arabic test collections and survey literature on constructing tweet test collections.

2.1 Arabic Test Collections

Test collections supporting Arabic IR can be categorized into two groups: multi-lingual collections and Arabic-only collections, and both are discussed next. For further information about Arabic test collections and Arabic IR, please refer to Darwish and Magdy (2014).

2.1.1 Multi-lingual Test Collections

Many evaluation campaigns used Arabic documents in multi-lingual test collections for various IR tasks such as cross-language retrieval (Gey and Oard 2001; Oard and Gey 2002), topic detection and tracking (Strassel et al. 2005; Glenn et al. 2006), identifying nuggets (i.e., text snippets)[4], thematic grouping of retrieved results[5], question answering for machine reading (Peñas et al. 2012), and community question answering (Nakov et al. 2015, 2016, 2017). These test collections focus on genres other than tweets in various languages including Arabic, while *EveTAR* focuses only on Arabic tweets.

2.1.2 Arabic-only Test Collections

To the best of our knowledge, there is no prior work that constructs an Arabic-only test collection as a primary contribution. Therefore, researchers working on Arabic IR systems had to construct their own test collections to conduct their experiments. Examples of such collections are presented in this section.

Darwish et al. (2012) explored issues making Arabic microblog search harder and investigated solutions for some of those issues. In order to perform their experiments, they constructed an Arabic tweet test collection for ad-hoc search that contains 112M tweets, 35 topics and 19,824 relevance judgments. For each topic, they manually judged top 30 results returned by each of 7 IR systems they employed in their experiments. In comparison, *EveTAR* is larger in all aspects. To detect potentially-relevant documents, we used multiple queries per topic instead of multiple IR systems. Moreover, *EveTAR* is designed to evaluate several IR tasks including ad-hoc search.

Alsaedi and Burnap (2015) proposed an Arabic event detection method and labeled the output of their system for evaluation. Experiments were conducted over 1M Arabic tweets restricted to Abu Dhabi. 3 annotators were hired to label the system output represented by event clusters. Differently, *EveTAR* supports multiple IR tasks, not only event detection, and is more reusable because events in *EveTAR* are not output of a specific system. In addition, events in *EveTAR* are not restricted to a specific location, which helps increase the linguistic diversity of the collection and the chance of having different types of events.

---

[3] http://qufaculty.qu.edu.qa/telsayed/evetar
[4] https://www.ldc.upenn.edu/collaborations/current-projects/bolt
[5] http://www.darpa.mil/program/broad-operational-language-translation

Azmi and Alshenaifi (2016) developed a question answering test collection to evaluate a system for answering why-questions. Given 700 Arabic articles from different genres, they recruited an annotator to develop 100 Arabic why-questions and answers. The annotator created that set after browsing the articles to ensure that the developed questions can be answered from the collection. The system was finally evaluated by manually comparing the system answers to the human-provided answers. *EveTAR* allows automatic evaluation for different IR tasks and targeting evaluation over Arabic tweets.

Overall, it is very challenging for researchers to work on today's IR research problems over Arabic data by employing existing Arabic test collections. The collections are usually old and contain limited number of documents and judgments. To the best of our knowledge, there is no publicly available test collection of Arabic tweets that serves the evaluation of event detection, real-time summarization, ad-hoc search and timeline generation tasks.

2.2 Tweet Test Collections

Twitter provides a rich resource of information but the huge size of the stream invokes the need for IR systems to help users sift through the stream. However, available tweets test collections for IR system evaluation are limited compared to other document types. This can be due to tweets redistribution policy of Twitter (i.e., redistribution of tweet texts at a scale larger than few tens of thousands is prohibited) and also being a more recent source compared to others such as Web pages and news articles. Existing tweets test collections were constructed in two ways: shared-task evaluation campaigns that develop collections to evaluate systems from teams participating in the same IR task, and centralized efforts by individual research teams who created test collections to develop and evaluate their IR systems or for the sake of providing test collections for the community. In this section, we present examples under each and also discuss what distinguishes *EveTAR* from available collections.

2.2.1 Shared-Task Evaluation Campaigns

One of the leading efforts to develop tweet test collections was carried by TREC evaluation campaigns. TREC developed large-scale tweet test collections as part of its Microblog track which ran from 2011 to 2015 and Real-time Summarization track which ran in 2016. The tracks targeted various IR tasks over English tweets such as ad-hoc search (2011-2014), filtering (2012), tweet timeline generation (2014) real-time filtering (2015), and real-time summarization (2016). 4 different collections were produced:

- Tweets2011 corpus, which contains 16M tweets have been used in 2011 (Ounis et al. 2011) and 2012 tracks (Soboroff et al. 2012). TREC organizers employed *download-tweets-yourself* approach, in which each participant has to download the tweets from twitter.com site given the ids of the 16M tweets in the collection. A total of 115 topics and 133K relevance judgments were created along that collection.
- Tweets2013 corpus, which contains 243M tweets has been used in 2013 (Lin and Efron 2013) and 2014 (Lin et al. 2014) tracks. Differently from Tweets2011, in order to distribute the tweet collection, TREC employed evaluation-as-a-service paradigm, in which tweets are stored centrally and the participants can access the tweets via an online search API. In total, 108 topics and 129K relevance judgments are released with that corpus.
- In 2015, instead of distributing the collection or providing online access to it ahead of time, each participant obtained the tweets directly from Twitter for a defined 10-day evaluation period (Lin et al. 2015). The participants in the track were responsible for collecting the 1% random sample of tweets made freely-available by Twitter daily[6]. The test collection included 51 topics and 94K relevance judgments in addition to novelty judgments.
- In 2016, TREC introduced the real-time summarization track (Lin et al. 2016), which also requires collecting tweets posted during a predefined evaluation period, instead of providing pre-crawled tweets. 51 topics and 67K relevance judgments in addition to novelty judgments where released with the collection.

Another evaluation campaign that targeted IR tasks over tweets is the FIRE 2016 Microblog track (Ghosh and Ghosh 2016) which focused on retrieval of tweets during disaster events. Organizers of the campaign collected tweets for the two weeks following the earthquake that occurred in Nepal and parts of India on the 25th of April 2015, by tracking the keyword "nepal". After eliminating duplicate and near-duplicate tweets, their dataset contained 50K English tweets. They developed 7 topics for the evaluation. In contrast, *EveTAR* contains much

---

[6] `https://dev.twitter.com/streaming/reference/get/statuses/sample`

more tweets, and topics cover multiple independent events, while their collection focused on sub-events under a single large-scale event.

To the best of our knowledge, there is no prior work on building multi-task collections except in TREC 2012 and 2014 Microblog tracks. Each track included a test collection to evaluate two tasks: ad-hoc search and filtering in 2012 (Soboroff et al. 2012), and ad-hoc search and timeline generation in 2014 (Lin et al. 2014) . Differently from those collections, *EveTAR* focuses on Arabic tweets while TREC collections focused on English only. More importantly, *EveTAR* is built without running a shared-task evaluation campaign where a diversified judgment pool was constructed using multiple manually-crafted queries per topic. Additionally, *EveTAR* supports the event detection task while it has never been targeted by any TREC tracks related to tweets. Evaluating multiple systems per task using *EveTAR* and TREC tweets test collections showed that *EveTAR* is as reliable as TREC collections in ranking the tested systems (Section 5.2). We provide an extensive comparison covering TREC test collections and *EveTAR* in Section 5.3.

2.2.2 Centralized Efforts

Many studies proposing IR systems for tweets also build their own test collections but for the purpose of developing and evaluating their systems. Collecting judgments for the test collection is generally done by evaluating the output of the proposed systems, reducing the reusability of the labeled data. Some of the tasks targeted by those studies include aligning tweets with events (Rowe and Stankovic 2012), tagging tweets (Ma et al. 2014), event monitoring (Zhang et al. 2015), event detection (Liu et al. 2016; Kunneman and Bosch 2014; Becker et al. 2011; Qin et al. 2013), adaptive filtering (Magdy and Elsayed 2016) and first story detection (Petrovič et al. 2012). The test collections used in the cited event-detection studies are smaller than *EveTAR* since the main focus of those studies was to develop IR systems instead of building test collections.

There are also studies mainly focusing on constructing English test collections. McMinn et al. (2013) built a publicly-available test collection for evaluating event detection systems over English tweets. The authors crawled around 120M tweets, covering more than 500 events identified using automatic and manual ways, and collected labels for over 150K tweets via crowdsourcing. *EveTAR* is designed for multiple tasks, not only for event detection. Additionally, we crawled about 3 times more tweets (355M vs. 120M) and labeled more tweets per event on average. CREDBANK is another test collection developed without running an evaluation campaign (Mitra and Gilbert 2015). The collection was designed to evaluate tasks focusing on information credibility on social media and includes more than 60M tweets and 1049 events with crowd-sourced judgments. But we target different IR tasks.

3 Our Approach

A typical test collection comprises three major components: a document collection, a set of topics, and a set of judgments per topic. There are many challenges in test collection construction process. First, the document collection should be a good representative of the domain in which IR systems will be applied. Second, the topics should be carefully designed to represent real-world information needs. Third, the document-topic pairs to be judged should be carefully selected and the collected judgments should be consistent to achieve reliable evaluation. On top of these challenges, multi-task test collections impose additional challenges: both the topic set and collected judgments should be reusable across the different tasks, i.e., the topics represent real information needs in all tasks and the collected judgments should be reusable across the tasks to minimize the human effort and cost. Moreover, we aim to construct a collection that is large-scale, easily-accessible, and highly-available.

In this section, we outline our methodology to construct a multi-task tweet test collection that addresses those challenges and meets the target objectives. Section 4 provides more details on how we applied the methodology to construct *EveTAR*, our Arabic test collection.

3.1 Significant Events as Topics

We elected to choose *significant* (i.e., popular) events to be the core of our test collection. A significant event is defined as an occurrence that happens at a particular time in a specific location and is discussed by the media (e.g., covered by an online news article). Focusing on significant events has several advantages. First, it enables the development of topics that represent frequent information needs of Twitter users as they often search Twitter for timely and up-to-date information about events (Teevan et al. 2011; Zhao and Mei 2013). Second,

rich content (i.e., relevant tweets) is more available for popular events. Third, the popularity of the topics might help annotators to do more consistent judgments due to their probable familiarity with the significant events. Finally, it inherently facilitates the support of multiple tasks that generally revolve around events, namely event detection, ad-hoc search, timeline generation, and real-time summarization, which are defined in more detail next.

3.2 Multi-Task Collection

With the choice of events to be at the heart of our test collection, **Event Detection (ED)** task is naturally supported. The task requires the detection of events over a *stream* of tweets without prior knowledge on what events to expect and when they might happen. Our definition of an *event* (given in Section 3.1) is similar to definitions introduced in (McMinn et al. 2013); however, we emphasize event significance, which is often neglected in most event definitions (Alsaedi and Burnap 2015; Becker et al. 2011; Petrović et al. 2010). An event in our collection is *represented* as a set of tweets that are relevant to it within a time period surrounding the time of that event. The event *detected* by the event detection system would be represented as a subset (not necessarily all) of the tweets that are relevant to that event.

ED task requires relevance judgments for each event. This inherently enables our test collection to support the **Ad-hoc Search (AS)** task as well. Ad-hoc search is the typical search task in IR in which an ad-hoc query (representing a topic of interest) is issued at a search system which is required to retrieve a ranked list of documents (i.e., tweets) that are relevant to that topic over a collection of documents. The query is usually tagged with a time stamp that indicates the time at which it is issued. In *EveTAR*, each topic is represented by multiple different queries that can be used in ad-hoc search, e.g., the title of the event.

As Twitter has become a valuable information source from which we can get instant updates about topics (or events) of interest, the massive amount of data posted on Twitter makes it very challenging to track the topics manually. Therefore, an IR system that helps users filter out redundant information and elect relevant updates in real-time is crucial. The **Real-time Summarization (RTS)** track was introduced in TREC 2016 to tackle this exact problem (Lin et al. 2016). According to TREC definition, an RTS system monitors a stream of tweets in real-time to detect relevant but non-redundant tweets for a given topic (called interest profile). Those tweets are then pushed (in real-time) to the mobile of the interested user. The pushed tweets represent the developed summary of the topic over time. In TREC, the maximum number of tweets pushed per day is limited to 10 to avoid overwhelming users. Note that a relevant tweet can be redundant if another tweet covering the same information has already been pushed and RTS systems are indeed penalized when they return redundant tweets. Therefore, in addition to relevance judgments, we also need *novelty* judgments to support the RTS task. *EveTAR* adopts the same exact definition but applied on events; however, the time for each event is limited to its active period only (5 days as explained in Section 4), simulating the fact that the user would be interested in the event only during that period.

Having relevancy and novelty judgments, we can also support **Tweet Timeline Generation (TTG)** (Lin et al. 2014) task which is the retrospective version of the RTS task. Given a topic at a specific time, a TTG system is required to return a *set* of relevant but non-redundant tweets in a chronological order (called a *timeline*). The timeline is considered as a retrospective summary of the topic at the issuing time.
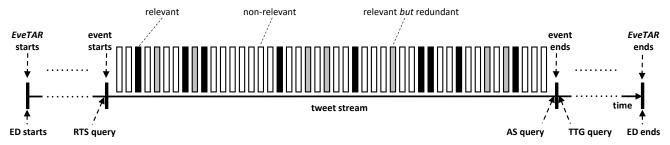
Fig. 1 The four supported retrieval tasks over *EveTAR*. ED and AS tasks target relevant tweets (black and gray) while TTG and RTS tasks target relevant but not redundant tweets (black only)

Both ED and RTS are considered streaming tasks since each processes the tweets as a stream; whereas AS and TTG are considered retrieval tasks since each processes the tweets as a collection. Figure 1 illustrates the relationship between the four tasks with respect to time and also type of targeted tweets for an example event.

For that event, sets of relevant and also novel tweets are identified over its specified time period. The evaluation of both ED and AS tasks requires only the relevant tweets (black- and gray-colored), while the evaluation of both TTG and RTS requires relevant and novel tweets (black-colored). The ED task starts with the beginning of the collection since all events should be detected, while RTS task per topic starts when the topic becomes of interest, i.e., when the event starts. Queries of both AS and TTG tasks are issued at the end of the event because they require retrospective retrieval of tweets. Table 1 contrasts the four tasks with respect to the nature of the input and output, and whether the tasks consider the novelty of retrieved tweets.

Table 1 Supported IR Tasks

Task	Input Data	Novelty	Expected Output
Event Detection	Tweet stream	✗	Events (represented by sets of tweets)
Ad-hoc Search	Tweet collection	✗	Ranked list of tweets
Real-time Summarization	Tweet stream	✓	10 tweets per day
Timeline Generation	Tweet collection	✓	A set of tweets

3.3 Large-Scale and Dense Dataset

To properly support the evaluation of the four IR tasks, we need to acquire a tweet dataset that is both *large-scale* and *dense*. In other words, we need to ensure breadth, to cover a good number of significant events, and depth, to have enough relevant content per event.

To achieve the needed breadth, the dataset has to span a time period that is long enough to cover several events. Considering that those events represent our topic set, we have to ensure the coverage of a number of events that is typically used in TREC to construct a *reliable* test collection, which is roughly 50. For Arabic collection, as an example, a month is expected to cover that number.

To achieve the needed depth, the Twitter's streaming API[7] typically provides around 1% random sample of tweets posted in Twitter and covering all languages used in the platform; however, tweets of a language that is not dominant (e.g., Arabic) will be an even lower percentage of that sample. To avoid sparseness, the API can then be used to track (and thus collect) tweets that include *frequent* keywords, ensuring a focused set of tweets that are (almost) all of the target language, while abiding by Twitter's tweet crawling rate limit. Since the set of keywords being tracked controls which tweets will be added to the final dataset, a special attention has to be paid to the selection of those words in order to maximize the chance of obtaining a representative set of tweets of the target language on Twitter.

3.4 High-Coverage and Diversified Judgment Pool

In our test collection, we need two types of judgments, relevance and novelty; however, relevance judgments have to be collected first before determining which of the relevant ones are also novel. This requires getting first a set of potentially-relevant documents per topic that constitutes the judgment pool.

Ideally, each document-topic pair should be judged manually. However, due to the high cost of relevance judging, we can judge only a subset of documents in practice. Clearly, the choice of documents to be judged has a significant impact on the quality of the test collection. Traditionally, unjudged documents are assumed to be irrelevant; therefore, in collecting the relevance judgments for a topic, it is crucial to acquire relevant documents/tweets as much as possible. Furthermore, the method of selecting tweets to be judged should not be biased to a specific retrieval algorithm in order to achieve reliable evaluation. Many methods have been proposed to select documents to be judged, such as pooling (Jones et al. 1975), statAP (Pavlu and Aslam 2007), MTC (Carterette et al. 2006), and interactive judging (Cormack et al. 1998). Most of these methods require a diverse set of systems to enrich the document pool. Developers of TREC test collections overcome this problem by running a shared-task evaluation campaign where research teams participate with their own systems. On one hand, organizing a shared-task is a very costly process. On the other hand, developing many systems is also a very challenging process, especially for languages with limited IR resources such as Arabic.

---

[7] `https://dev.twitter.com/streaming/reference/get/statuses/sample`

Moffat et al. (2015) showed that query variations for a given topic are as strong as system variations in producing a diverse document pool. Therefore, in order to diversify the tweet pool, we adopt a process that manually crafts a list of keyword and phrase queries for each topic by performing *interactive search* on Twitter's website. That (compared to searching our collection) would ensure a wider coverage of the topic aspects. The crafted queries can then be used to search the collection using an off-the-shelf retrieval engine to retrieve the potentially-relevant documents that constitute the judgment pool.

3.5 Reliable Judgments

After identifying the potentially-relevant tweets, we need to collect relevance and novelty judgments. There are two issues to consider:

1. For relevancy, the pool for each topic can be annotated by several annotators, since the decision on whether a tweet is relevant is independent of the others. However, that is not the case for novelty judgments, where the decision on whether a tweet is novel depends on the other relevant tweets.
2. Since the topics are significant events, we expect a relatively large pool of tweets to judge for relevancy, which calls for a large-scale annotation effort. However, the pool for novelty is expected be much smaller as it includes the relevant tweets only.

Based on that, crowd-sourcing is a suitable option for relevancy annotations in terms of scale; however, we need to ensure the reliability and quality of the judgments. We achieve that through three steps: (1) we require multiple judgments per tweet, (2) We require a minimum quality level of annotators based on their work history, and (3) we employ pre-qualification test for potential crowd-sourcing annotators who are also required to maintain a minimum accuracy throughout the process. To further improve the overall quality of the annotations, after collecting the relevance judgments of all topics, we exclude the topics of low Kappa inter-annotator agreement (Fleiss 1971), keeping only the topics of high agreement.

The novelty annotations require the annotator to go through the entire stream of relevant tweets for a given topic to decide which of them are redundant and which are novel. We therefore opt to hire in-house annotators to ensure consistency and integrity of the annotations. This allows us to closely train the annotators and monitor their work.

3.6 Easily-Accessible Highly-Available Collection

Due to Twitter's tweet re-distribution policy, only tweet IDs (not content) can be distributed. Therefore, researchers interested in using the collection have to re-crawl the tweets from Twitter using their IDs. That introduces two major challenges. First, many tweets might become inaccessible over time, e.g., deleted tweets or tweets of deleted or suspended accounts. This will make performance comparisons over the collection less meaningful due to the use of probably different crawls performed at different times. Second, the estimated crawl time of a large set of millions of tweets over a single machine is in the order of tens of days based on the rate limit enforced by Twitter's API. This can be challenging for research teams that are interested in using the collection but do not have the time or computational power to crawl it.

To tackle the first challenge, we mitigate the effect of inaccessible tweets by removing (most of) them from the collection (i.e., not including them in the released version) after a reasonable period of time has been spent since the original crawl. That would make the collection more stable by reducing the chance of including deleted tweets for example, as we believe it decays over time.

To tackle the second one, we make multiple smaller subsets available; all include the full judged set of tweets. The subsets have to be small enough to be easily crawled in a timely manner, while being representative of the complete collection. They also allow much faster evaluation. Moreover, they make it possible to conduct experiments on different scales of data, which might lead to different insights.

4 Constructing EveTAR

Following the approach we presented in Section 3, we have constructed *EveTAR*, a multi-task test collection over Arabic tweets. We have applied our approach over two phases. In the first phase, we collected a set of Arabic tweets, developed topics using Wikipedia's Current Events Portals, and collected relevance judgments via

crowdsourcing. The collection at the end of this phase was released earlier (Almerekhi et al. 2016). In the second phase, we aimed to improve the quality of the judgments, extend the collection to support more retrieval tasks, and make it more accessible. We finally collected novelty labels to support TTG and RTS tasks. We also created four subsets of the collection to serve different purposes. In the following sub-sections, we describe in detail the pipeline of these steps illustrated in Figure 2.
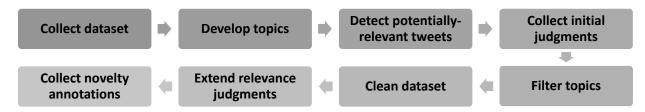

Fig. 2 Pipeline of steps followed to create *EveTAR*

4.1 Collecting the Dataset

To construct a test collection, a dataset has to be acquired first. We used the Twitter's streaming API[8] to collect Arabic tweets spanning the period between December 30th, 2014 and February 2nd, 2015. To avoid sparseness, we used Twitter's streaming API to track[9] 400 frequent Arabic words so that we can collect Arabic tweets that match those words. The words were identified as the most frequent in a set of 2M Arabic tweets that were crawled via the 1% random sample of tweets collected over 10 days starting from April 10th, 2014. Using this methodology, we eventually collected a set of 590M Arabic tweets within the time period mentioned above.

4.2 Developing Topics

The second major component of the test collection is the set of topics representing user needs. As mentioned earlier, we elected to select significant events to be the core topic set of *EveTAR*. Guided by the work of McMinn et al. (2013), we collected a set of 357 events that took place in January 2015 according to the English[10] and Arabic[11] Wikipedia's Current Events Portals (WCEP). Each event in this list is represented by a date and a title (or an English-to-Arabic translated title). The event set went through two filtering phases to detect events meeting our significance criteria. In the first phase, we kept only the events that have been discussed by at least one online *Arabic* news article, which resulted in a set of 71 events only. This process was done by manually searching for articles in popular Arabic news websites such as Aljazeera[12] and CNN Arabic[13] using event titles as queries. In the second phase, Twitter's advanced search service[14] was used to manually find at least 20 relevant Arabic tweets for each event using several manually-crafted Arabic queries per event. We limited the search for each event to a 5-day period starting 2 days prior to the event date (since some discussions about the event might start prior to it, while most of the relevant tweets can be found on the day of the event or within two days of its occurrence). Following these two filtering steps, we acquired a final set of 66 significant events. Each event in *EveTAR* is represented by a rich representation including event title, date, location, and category extracted from WCEP. Figure 3 shows a translated example of one event in *EveTAR*. Other examples of events in *EveTAR* include: "North Korea hacks Sony accounts", "Suicide bombing in Ibb", and "Match of Australia vs Kuwait at the opening of the Asian Cup".

---

[8] https://dev.twitter.com/streaming/reference/get/statuses/sample
[9] https://dev.twitter.com/streaming/overview/request-parameters#track
[10] https://en.wikipedia.org/wiki/Portal:Current_events
[11] http://bit.ly/2n5TYhY
[12] http://www.aljazeera.net
[13] arabic.cnn.com
[14] https://twitter.com/search-advanced

ID	E12
Title	Discovery of tomb of Egyptian queen Khentakawess III
Date	January 04, 2015
Location	Abusir, Egypt
Category	Arts and Culture
Reference	`http://cnn.it/1O6grQK`
Keywords	Khentakawess, Egyptian queen, archaeologist
Description	A Czech archaeological team discovered the tomb of an Egyptian queen named Khentakawess III who lived during the 5th dynasty.

Fig. 3 A translated example of an event as represented in *EveTAR*

4.3 Detecting Potentially-Relevant Tweets

Having acquired the dataset and developed the topics, the next natural step is to collect judgments per topic. In *EveTAR*, we have two types of judgments, relevance and novelty; however, relevance judgments have to be collected first before determining which of the relevant ones are also novel. To construct the judgment pool, we adopted an interactive search approach by manually crafting a list of keyword and phrase queries for each topic using the search service on Twitter's Website. For example, some of the (translated) queries for the event described in Figure 3 are: "Abusir", "Khentakawess Third" and "Egyptian queen". Given an average of 6 queries per topic, we created one long "OR" query comprising all queries per topic, and retrieved 10K tweets per topic using Lucene open-source search library[15]. We used Lucene's default retrieval model for ranking. We also limited the search to the event specific time period. After removing exact tweet-text duplicates, we obtained a pool of 134K tweets to be judged.

4.4 Collecting Initial Relevance Judgments

To collect relevance judgments for the pool, we used CrowdFlower crowdsourcing platform[16]. We conducted several pilot studies to design our tasks. We ran a separate judging task for each event/topic. In each task, the corresponding title, description, and date of the event have been shown to annotators. We also provided the content of an Arabic news article discussing the event to help annotators better understand the topic and make better-informed judgments. We decided to display the content of the article rather than just providing a hyperlink to it, to increase the chance of annotators reading the article.

Annotators were required to be Arabic speakers with an intermediate level based on CrowdFlower's ranking of annotators[17]. Before starting the judging process, we conducted a qualification test in which annotators were required to correctly label 8 out of 10 gold tweets. Gold tweets are randomly sampled from the set of candidate tweets per event and labeled by one of the authors. Annotators were also required to maintain a minimum accuracy of 80% over gold tweets embedded within the actual task. Eight annotators on average per task have failed to maintain that judging accuracy and were eliminated (along with all of their judgments).

Each tweet-event pair was judged by 3 annotators to ensure a majority label. Annotators spent 59 judging hours per event on average. Given the 3 judgments per tweet, we used majority voting to select the final label. Eventually, our relevance judgment set consisted of 134K labeled tweets, out of which 51K tweets were judged as relevant. We measured inter-annotator agreement per event using Fleiss' Kappa (Fleiss 1971), which allows measuring agreement when a data item is annotated by more than 2 annotators. The average Kappa value over all topics is 0.60. The dataset, 66 topics, and collected judgments (at that point) were made available in an early version of *EveTAR* (Almerekhi et al. 2016).

4.5 Filtering Topics

To improve the average judgment quality of *EveTAR*, we only kept the 50 topics with the highest agreement levels, dropping the lowest 16. The distribution of categories of the 50 events is shown in Table 2. As the table

---

[15] `https://lucene.apache.org`

[16] `http://www.crowdflower.com`

[17] `https://success.crowdflower.com/hc/en-us/articles/115000832063-Frequently-Asked-Questions`

shows, events in the "Armed conflicts and attacks" category constitutes 64% of the entire set; this is due to the distribution of events in WCEP itself.

Table 2 Distribution of events in *EveTAR*

Category (Events)	Category (Events)
Armed Conflicts & Attacks (32)	Sports (5)
Business & Economy (1)	Arts & Culture (2)
International Relations (3)	Law & Crime (1)
Disasters & Accidents (2)	Politics & Elections (4)

We also investigate whether the 50 events are temporally biased. Figure 4 shows the number of *EveTAR* events started on each day of January 2015. We can see that the events are generally well-distributed over the period of the collection indicating that *EveTAR* events are not biased to specific dates.
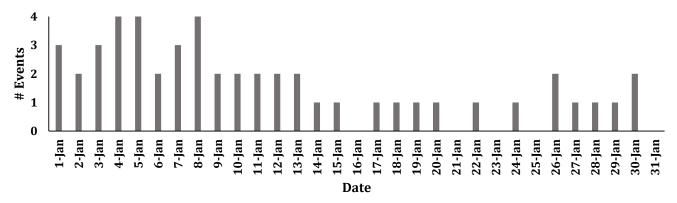

Fig. 4 Daily distribution of *EveTAR* topics over the month of January, 2015

4.6 Cleaning the Dataset

Tweets might become inaccessible over time due to various reasons (Bagdouri and Oard 2015). Therefore, removing inaccessible tweets from the test collection is essential to evaluate IR systems reliably over a re-crawled version of *EveTAR*, and also reduce re-crawling time.

In December 2016, we examined the judged tweets released with the early version of *EveTAR* (Almerekhi et al. 2016) and found that around 40% of them have been inaccessible. The two major reasons behind that were (1) due to users deleting their own tweets, or (2) the whole user account is no longer accessible because either it was suspended[18], deleted, or made private by the user. The majority resulted from the inaccessibility of accounts (which is inline with findings of Liu et al.'s study (Liu et al. 2014) conducted at a very large scale).

Due to the huge size of the dataset (590M tweets), it was infeasible to automatically check each tweet for deletion. As a compromise, we opted to apply an efficient solution; we observed that the majority of the inaccessible tweets belong to inaccessible accounts. Therefore, we have automatically checked[19] each of 7.1M unique accounts in our original dataset and only kept the accessible accounts along with their tweets, resulting in around 5.8M accounts and a final dataset of 355M tweets.

4.7 Extending Relevance Judgments

Cleaning the dataset has encouraged us to improve the relevance judgments as well. We first removed retweets (since they express no new information) and filtered out all inaccessible tweets. That resulted in a significant drop

---

[18] Shut-down by Twitter; check https://support.twitter.com/articles/15790.
[19] on December 25th 2016.

in the number of relevance judgments. We also observed that many of the labeled tweets were written in modern standard Arabic (MSA); however, Twitter Arab users tend to use dialectal Arabic in their tweets (Darwish and Magdy 2014). Therefore, we decided to extend the relevance judgments and increase linguistic diversity among labeled tweets.

To achieve that, we developed 3 to 9 new queries per topic following the same earlier guidelines while attempting to include dialectal queries whenever possible. In a similar approach to that explained in Section 4.3, we used the developed queries in searching using Lucene over an index of the cleaned dataset. For each topic, the top 1K tweets were considered for labeling. Before passing those tweets to the relevance labeling task, we applied exact-text deduplication and retweet removal, and also excluded the tweets that have been judged before. Eventually, we collected 180 additional judgments on average per topic using the same crowdsourcing tasks we employed previously.

Since some of the labeled tweets have exact duplicates in the collection, a system that returns an exact duplicate rather than the labeled tweet would not be scored on that tweet. Moreover, a recent study (Baruah et al. 2014) has found that adding exact duplicates to the judgment set significantly affects system performance over different queries for 13 out of 28 systems participated in the temporal summarization track in TREC 2013 (Aslam et al. 2013). Therefore, we also propagated the relevance judgments by adding 3,947 exact duplicates of judged tweets fetched from the dataset. Note that we only considered exact duplicates that appear within active period of each event.

Eventually, we have a final relevance judgment set of 61,946 tweets, out of which 24,086 (39%) are relevant. Computing Fleiss' Kappa over the final qrels resulted in an average Kappa of 0.71, which is considered a *substantial* agreement according to a widely-adopted Kappa categorization (Sim and Wright 2005), showing how strong the agreement is. Figure 5 illustrates the distribution of the 50 events according to that categorization. 78% of the events has an inter-annotator agreement that is above moderate which indicates a strong inter-annotator agreement for most of the topics in *EveTAR*. We also observe that 2 topics only got an agreement as low as fair. Analyzing labels for the topic with the least agreement, we observe that conflicts in labels might have happened because of incompleteness in the title of the event. Although the topic is about a delay of a specific Etihad Airways flight to San Fransisco, the title did not reflect the important piece of information about the destination of the flight. At the same time, the article displayed to the annotators included that piece of information. For further studies on crowdsourcing, we released the complete set of judgments and inter-annotator agreement per event.
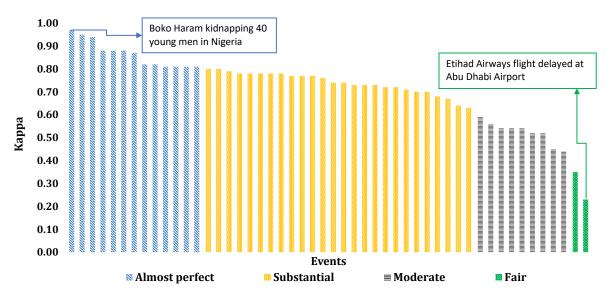

Fig. 5 Distribution of *EveTAR* events over categories of Kappa values

Figure 6 shows the number of relevant tweets per event. The figure demonstrates that the size of the relevant sets spans a large range and that we have events at all difficulty levels, assuming that events with more relevant judgments are easier to handle.
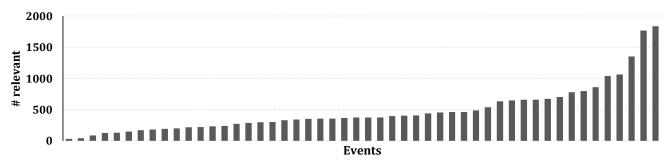
Fig. 6 Distribution of relevant tweets over *EveTAR* topics

4.8 Collecting Novelty Annotations

Novelty judgments are considered a second layer of labels on top of the relevance judgments. To collect the novelty judgments, we adopted the following definition of semantic-similarity between on-topic tweets, introduced in TREC 2015 Microblog (Lin et al. 2015) and TREC 2016 RTS (Lin et al. 2016) tracks. Relevant tweets for each topic are distributed into clusters of "semantically-similar" tweets; each cluster has a set of relevant tweets considered to carry the same information. Particularly, while receiving a chronologically-ordered stream of relevant tweets for a given topic, any incoming tweet that conveys the same information as previously-seen tweets, or it contains additional information that the user will not be interested to see, is considered redundant and is added to the cluster of previously-seen similar tweets. Otherwise, the tweet is judged as novel and it initiates a new cluster.

Following that definition, we collected novelty judgments using a web-based annotation interface (Wang et al. 2015). We recruited 12 in-house annotators (a mix of undergraduate students and alumni of Qatar University) and conducted two 3-hour training sessions, each covered motivation, task definition, and a demo of the annotation interface. Each annotator had to annotate two training topics before being eligible to work on the task. After checking the quality of annotations on the training topics, only 9 annotators were finally selected to work on the task. We made the instructions of the task available after the training session [20] for annotators to refer to during the actual task.

After excluding the exact duplicates of relevant tweets, only 22K relevant tweets had to be annotated for novelty over 50 topics[21]. Each topic was annotated by one annotator. The number of topics each annotator worked on varied from 1 to 9 depending on annotators' speed and availability, and also topic pool size. The whole annotation task finished in approximately 100 hours over 3 weeks and produced 66 clusters per topic on average.

4.9 Subsets

The estimated crawl time of 355M tweets over a single machine is 41 days based on the rate limit enforced by Twitter's API on tweets re-crawl[22]. Therefore, it can be challenging to re-crawl *EveTAR* by research groups with limited computational resources. In order to make *EveTAR* more accessible, we created a subset that can be re-crawled in a shorter time period while supporting the evaluation on different scales of datasets. Furthermore, the Arabic language has many dialects that are also commonly used on Twitter (Darwish and Magdy 2014). To support the evaluation and performance comparisons of IR systems over dialectal vs. MSA tweets, we also provided the MSA and dialectal subsets of the sample. Finally, we provided another subset that includes only qrels to support even faster evaluation and facilitate quality analysis of crowdsourcing annotations.

Table 3 summarizes the different versions of *EveTAR* we made publicly available. In addition to *EveTAR*-F, which is the full collection of 355M tweets, we also release the following four subsets:

1. ***EveTAR*-S**: A random sample of 15M tweets of *EveTAR* (i.e., 4% of the full dataset) that includes all judged tweets and can be recrawled in 1.5 days based on our estimation.
2. ***EveTAR*-S.m**: The subset of *EveTAR*-S that includes only MSA tweets.
3. ***EveTAR*-S.d**: Similar to *EveTAR*-S.m but includes only dialectal tweets.

---

[20] `https://reemsuwaileh.github.io/EveTARNovelty/training.html`

[21] Cluster labels were eventually propagated back to duplicate tweets.

[22] `https://dev.twitter.com/rest/reference/get/statuses/lookup`

Table 3 EveTAR Test Collection Versions

Version	Description	Size	#qrels	#rels
EveTAR-**F**	Full collection	355,821,033	61,946	24,086
EveTAR-**S**	Random sample	15,057,208	61,946	24,086
EveTAR-**S.m**	MSA tweets of the sample	8,066,583	47,369	21,233
EveTAR-**S.d**	Dialectal tweets of the sample	6,990,533	14,577	2,853
EveTAR-**Q**	Qrels only	59,732	61,946	24,086

4. ***EveTAR*-Q**: The subset of the full collection that includes only the judged tweets (i.e., qrels). Notice that the size of that subset is smaller than the size of qrels because one tweet might be judged for more than one event.

We classified tweets of *EveTAR*-S into MSA and dialectal subsets using a pre-trained classifier (Eldesouki et al. 2016) which achieved the best accuracy (83.9%) in DSL shared-task[23]. The classifier predicted 46.4% of the random sample as dialectal tweets. Among tweets of *EveTAR*-S, there were 92 tweets that contain no text, and thus were filtered out and not included in neither of the two subsets.

5 Evaluation

In this section, we evaluate *EveTAR* from three perspectives. First, we demonstrate its usability for evaluating the tasks it supports (Section 5.1). This also produces reference performance results for future advancements in those tasks. Second, we evaluate the reliability of the test collection in ranking IR systems in different tasks and investigate the effect of sampling in terms of evaluation reliability (Section 5.2). Finally, we compare *EveTAR* with other widely-used TREC microblog test collections (Section 5.3).

5.1 Usability

We first evaluate strong existing IR systems for each supported task using *EveTAR*. Experiments are conducted over each version of *EveTAR* (whenever feasible) to provide baseline performance results at different scales.

5.1.1 Event Detection

As events are at the heart of *EveTAR*, event detection is a primary supported task. We ran two off-the-shelf event detection algorithms on *EveTAR* to demonstrate its usability for this task. We used open-source implementation of Trending Score (Benhardus and Kalita 2013) and Peaky Topics (Shamma et al. 2011) algorithms provided by SONDY platform[24]. We used SONDY's default parameter settings and ran both over all versions of *EveTAR* except the full dataset as we did not have enough computational power to apply the algorithms at that scale.

To automatically evaluate the algorithms, we adopted Petrovic's (Petrovic̀ 2013) approach which measures effectiveness using recall:

$$recall = \frac{\#covered\ events}{\#reference\ events} \tag{1}$$

where *#reference events* is the number of events in *EveTAR* and *#covered events* is the number of reference events covered by the algorithm's detected events (where an event is represented by a list of tweets). A reference event is assumed to be covered by a detected event if at least 50% of the tweets in the detected event belong to that reference event. It should be noted that it is not feasible to automatically compute precision under this setup, since the "complete" list of possible significant events in the collection is not available. As it is difficult to manually compute precision given the huge number of events detected by each algorithm, we only report recall results in Table 4.

The table shows that Trending Score is outperforming Peaky Topics over all versions of *EveTAR*. We also observe that both algorithms managed to achieve perfect recall over *EveTAR*-Q due to its small size and the very

---

[23] http://ttg.uni-saarland.de/vardial2016/dsl2016.html

[24] mediamining.univ-lyon2.fr/people/guille/sondy.php

Table 4 Event detection over *EveTAR*. Best recall per version is **boldfaced**

Version	Peaky Topics	Trending Score
EveTAR-S	0.42	**0.52**
EveTAR-S.m	0.74	**0.92**
EveTAR-S.d	0.14	**0.48**
EveTAR-Q	**1.00**	**1.00**

high prevalence of relevant documents. Similarly, event detection over *EveTAR*-S.m is noticeably better than that over *EveTAR*-S and *EveTAR*-S.d due to the higher prevalence of relevant documents in *EveTAR*-S.m (0.26%) as opposed to the other two subsets (0.16% in *EveTAR*-S and 0.041% in *EveTAR*-S.d).

5.1.2 Ad-hoc Search

Using *EveTAR*, we evaluated four AS systems, denoted by query likelihood (QL), query expansion (QE), temporal decay (TD) and temporal relevance modeling (TRM), which were adopted by one of the top teams (Hasanain and Elsayed 2014) in the ad-hoc search task in TREC 2014 microblog track (Lin et al. 2014). Moreover, these models represent a diverse set of retrieval models usually used in tweet search. To be able to run those systems on Arabic tweets, we added Arabic-specific preprocessing modules, e.g., stemming and removing stopwords, emoticons, and diacritics, in addition to URL and mentions removal. We ran the systems (using their reported parameter values) on all 5 versions of *EveTAR*, as shown in Table 5, and limited the search to the active period of each event. We used mean average precision (MAP) and precision at rank 30 (P@30) averaged over all 50 topics to evaluate system performance, as typically used (Lin et al. 2014).

Table 5 Ad-hoc search over *EveTAR*. Best result per version per evaluation measure is **boldfaced**

	MAP				P@30			
Version	QL	QE	TD	TRM	QL	QE	TD	TRM
EveTAR-F	0.3951	0.4494	0.3926	**0.4534**	0.7340	0.7180	0.7313	**0.7707**
EveTAR-S	0.5986	**0.6642**	0.5961	0.6630	0.8373	0.8520	0.8360	**0.8747**
EveTAR-S.m	0.6202	0.6848	0.6184	**0.6946**	0.8473	0.8480	0.8453	**0.8680**
EveTAR-S.d	0.5529	0.5817	0.5505	**0.6006**	0.5007	**0.5313**	0.4927	0.5233
EveTAR-Q	0.6340	0.6882	0.6317	**0.6951**	0.8133	**0.8600**	0.8140	0.8373

Table 5 shows that, in both measures, TRM is generally outperforming all other models in all versions of *EveTAR*. As expected, QE is almost as effective as TRM since both systems are query expansion retrieval models that utilize pseudo-relevant tweets to expand the query. Results also showed that the systems are not as effective over *EveTAR*-F and *EveTAR*-S.d as they are over the other subsets. This is due to the low prevalence of relevant documents in these two subsets (0.007% in the full version and 0.041% in the dialectal sample) as opposed to the other subsets, making it a more difficult retrieval task.

5.1.3 Tweet Timeline Generation

We conducted experiments with two TTG systems, cutoff and incremental clustering, adopted by one of the top teams (Hasanain and Elsayed 2014) in the TTG task in TREC 2014 microblog track (Lin et al. 2014). Both systems follow a two-step approach to generate a timeline. The system first retrieves a set of candidate tweets in response to a topic using a retrieval model, then applies a summarization technique to the retrieved set to generate a timeline. In the first step, we experimented with the same four ad-hoc retrieval models presented earlier, namely QL, QE, TD, and TRM.

We enabled the TTG systems to run on Arabic tweets similar to the AS systems, and configured both using their reported parameter values. We used weighted F_1 (wF_1) measure, which is the official one used in TREC 2014 (Lin et al. 2014), to evaluate the performance of the systems.

Table 6 TTG over *EveTAR*. The best result per version is **boldfaced**

Version	Cutoff				Incremental Clustering			
	QL	QE	TD	TRM	QL	QE	TD	TRM
EveTAR-F	0.2164	0.2009	0.2180	0.2043	0.3649	0.3578	0.3608	**0.3809**
EveTAR-S	0.2574	0.2474	0.2588	0.2450	0.4301	0.4210	0.4296	**0.4329**
EveTAR-S.m	0.2586	0.2474	0.2597	0.2563	0.4252	0.4144	0.4265	**0.4374**
EveTAR-S.d	0.1817	0.1874	0.1793	0.1879	0.2229	**0.2337**	0.2203	0.2325
EveTAR-Q	0.2648	0.2445	0.2686	0.2520	0.4070	0.3819	0.4058	**0.4115**

Table 6 reports the average wF_1 scores over the 50 topics for both TTG systems over *EveTAR*. Similar to AS results, the table shows that TTG systems are less effective with *EveTAR*-F and *EveTAR*-S.d test collections, due to lower percentage of relevant tweets. We also notice that Incremental Clustering is far more effective than Cutoff. Combining Incremental Clustering with the TRM retrieval model achieves the best performance over almost all versions of *EveTAR*. This is consistent with findings of a recent study that showed the performance of TTG systems adopting the 2-step approach is highly dependent on the underlying retrieval model (Magdy et al. 2016).

5.1.4 Real-Time Summarization

We evaluated three RTS systems over *EveTAR*. Two of them, namely Vector Space Model (*VSM*) and *Rocchio* (Suwaileh et al. 2016), constitute the best two automatic systems in the "push notification" scenario in TREC 2016 RTS track (denoted then as QUBaseline and QUExpP). We also adopted a third system, denoted as *Silent*, that was designated as the baseline at the same track (Lin et al. 2016). Basically, this system is always silent, i.e., not pushing anything to the user during the entire evaluation period and just accumulates the gain of silent days in which a topic does not have any relevant tweets.

Table 7 RTS over *EveTAR*. The best result per version per evaluation measure is **boldfaced**

Version	EG-1			nCG-1			#pushed	
	VSM	Rocchio	Silent	VSM	Rocchio	Silent	VSM	Rocchio
EveTAR-F	**0.2469**	0.2333	0.1600	**0.2688**	0.2384	0.1600	466	316
EveTAR-S	**0.2557**	0.2365	0.1600	**0.2799**	0.2455	0.1600	392	251
EveTAR-S.m	**0.2766**	0.2620	0.1800	**0.2975**	0.2620	0.1800	344	227
EveTAR-S.d	**0.4737**	**0.4737**	0.4320	**0.4813**	0.4741	0.4320	128	100
EveTAR-Q	**0.2569**	0.2362	0.1600	**0.2807**	0.2479	0.1600	379	245

Table 7 illustrates the performance of the three systems measured by Expected Gain (EG-1) and Normalized Cumulative Gain (nCG-1) (Lin et al. 2016) over all versions of *EveTAR*. It also shows the total number of pushed tweets in each case. We can see that VSM outperforms other systems in all settings. In addition, EG-1 and nCG-1 scores of all systems in all versions of *EveTAR* are relatively close to each other, except for *EveTAR*-S.d. Upon investigation, we realized that the titles of topics used in summarization, were written in MSA, while tweets of that version are all in dialectal Arabic. Therefore, fewer relevant tweets have been detected as relevant and pushed by both VSM and Rocchio. Since RTS evaluation is precision-oriented, systems pushing fewer but relevant tweets are more likely to achieve higher performance. Silent also achieves higher performance on *EveTAR*-S.d, as there are more silent days in *EveTAR*-S.d due to fewer relevant tweets in that subset.

5.2 Reliability

In the previous section, we demonstrated the usability of *EveTAR* by evaluating the performance of multiple systems over the different supported tasks. In this section, we adopt Generalizability Theory (GT) (Brennan 2001; Bodoff and Li 2007; Bodoff 2008; Urbano et al. 2013) to measure the reliability of the reported results.

GT is based on the analysis of variance components (i.e., variance in effectiveness scores due to systems, topics and system-topic interactions). We computed the *generalizability coefficient* (GC) (using the implementation of Urbano et al. (2013)) which measures the reliability of ranking of IR systems. For a reliable ranking, GC score approaches 1. Note that the GC measure is sensitive to systems used in the computation, and thus is used to measure the reliability of a specific ranking, not the overall quality of a test collection.

In computing GC, we used MAP for the AS task, wF_1 for the TTG task, and EG-1 for the RTS task[25]. We considered both *EveTAR*-F and *EveTAR*-S to investigate the effect of sampling in terms of evaluation reliability. To have a reference point, we computed GC scores of the rankings of same systems but over the test collections of the most recent TREC track of each task (TREC 2014 Microblog Track for AS and TTG, and TREC 2016 RTS Track for RTS). Results presented in Table 8 show that *EveTAR* achieves higher GC scores than TREC collections, which indicates that the reliability of *EveTAR* in ranking the used systems is at least comparable to TREC test collections.

Table 8 Generalizability Coefficient Scores of system rankings

Task	*EveTAR*-F	*EveTAR*-S	TREC
AS	0.937	0.966	0.901
TTG	0.974	0.979	0.872
RTS	0.974	0.975	0.743

Table 8 also shows very similar GC scores for *EveTAR*-S and *EveTAR*-F, which suggests that sample subset is as reliable as the full collection in ranking systems. To further investigate the representativeness quality of the sample, we computed Pearson's correlation between performance scores of each system per topic over both *EveTAR*-F and *EveTAR*-S, and then averaged it over all systems per task[26]. Table 9 shows very high correlation between scores over both collections with very small standard deviation. Moreover, correlations for all systems are statistically significant at a 95% confidence level. This suggests that, overall, the sampled version of *EveTAR* (i.e. *EveTAR*-S) is as reliable as the full version in evaluating the tasks when it is not feasible to acquire the full collection.

Table 9 Pearson's correlation of performance over *EveTAR*-F and *EveTAR*-S

Task	Average Pearson's Correlation (Standard Deviation)
AS	0.866 (0.009)
TTG	0.892 (0.033)
RTS	0.973 (0.026)

5.3 In Context of TREC Collections

TREC test collections are among the most widely used test collections, covering a large variety of IR tasks. In this section, we compare *EveTAR* (mainly the full version) against TREC microblog test collections used in TREC 2011-2015 Microblog tracks (Ounis et al. 2011; Soboroff et al. 2012; Lin and Efron 2013; Lin et al. 2014, 2015) and TREC 2016 RTS track (Lin et al. 2016). Table 10 summarizes the comparison in several aspects.

- **Language:** *EveTAR* has the largest monolingual tweet collection compared to TREC collections which target a smaller subset of English tweets to match the language of topics. Moreover, *EveTAR* has MSA and dialectal subsets that can be useful in IR studies investigating linguistic diversity within the same language (e.g., cross-dialect IR).

---

[25] Evaluation reliability of ED task is not reported because we need evaluation scores for each topic in order to apply GT. However, in evaluation of ED systems, we calculate a single score for all topics.

[26] We are not able to report results for event detection since we could not run event detection on the full dataset due to its large size.

- **Tasks:** *EveTAR* supports four different IR tasks, one of them (ED) has never been supported by any of the TREC tweet test collections.
- **Topics:** The topic set size of *EveTAR* is comparable to TREC collections. *EveTAR* topics are defined in the TREC style (title, description, and narrative) in addition to having multiple queries per topic. Those queries can be useful in several research studies, e.g., automatic evaluation (e.g., (Efron and Winget 2010)) and development of new evaluation measures (e.g., (Zuccon et al. 2016)).
- **Qrels:** To create a pool of tweets to be judged, we used interactive search to generate multiple queries representing the same topic, which were then employed to search the collection and retrieve potentially-relevant tweets. TREC test collections employed the traditional pooling method via the shared-task campaign. While the pool size in *EveTAR* is comparable to most of the TREC collections, the total number of relevant tweets (and also per topic) in *EveTAR* is much higher than others. We expect this is due to two factors. First, the topics represent "significant events", which are naturally popular and more widely discussed. Second, *EveTAR* is a much larger collection than the others. However, percentage of documents in the collection that are relevant is within the range observed in TREC collections. Finally, the high percentage of relevant tweets in the pool (39%) might indicate the effectiveness of the multiple-queries in capturing relevant tweets.

The above comparison suggests that, while *EveTAR* is much larger, supports more IR tasks, and uniquely focuses on Arabic, it is comparable in multiple aspects to the TREC tweet test collections. This might shed some light on the expected quality of *EveTAR*.

Table 10 Comparison of *EveTAR* against six TREC microblog test collections

	EveTAR	TREC'11	TREC'12	TREC'13	TREC'14	TREC'15	TREC'16
Tweets	355M	16M	16M	243M	243M	43M[†]	45M[†]
Data Lang.	Arabic	Multiple	Multiple	Multiple	Multiple	Multiple	Multiple
Topic Lang.	Arabic	English	English	English	English	English	English
Tasks	AS, ED, TTG, RTS	AS	AS, Filtering	AS	AS, TTG, RTS*	TTG*, RTS	TTG*, RTS
Topics	50	50	60	60	55	51	56
Qrels Pool	Interactive	Shared Task	Shared Task	Shared Task	Shared Task	Shared Task	Shared Task
Pool Size	61,946	60,129	73,073	71,279	57,985	94,066	67,525
Relevant	24,086	2,965	6,286	9,011	10,645	8,233	3,339
Rel/Topic	482	61	107	150	194	161	60
%Rel/Coll.	0.007%	0.019%	0.039%	0.004%	0.004%	0.019%	0.007%

* Task was not officially run in the track but can be supported by the test collection.
† Dataset was crawled over 10 days at each participant end. Number is based on our local crawl.

6 Conclusion and Future Work

Test collections are evaluation tools that are essential for advancing the state-of-the-art in the field of information retrieval; however, the scarcity of Arabic test collections is evident. In this article, we addressed the problem of creating a high-quality multi-task tweet test collection without the need of a shared-task campaign. We proposed a new language-independent approach, applied on Arabic tweets, that leverages significant events as the core of the collection design. Our study introduced *EveTAR*, the first publicly-available multi-task Arabic tweet test collection. *EveTAR* supports four different tasks, namely event detection, ad-hoc search, timeline generation, and real-time summarization, over Arabic tweets. The collection consists of 355M Arabic tweets (crawled in January 2015), 50 topics (representing significant events), 62K relevance judgments (collected via carefully-designed crowd-sourcing tasks), and novelty annotations (gathered by in-house annotators). In addition to the full collection, we provided 4 different subsets in order to make it more accessible and further help research on dialectal Arabic IR.

We demonstrated the quality of *EveTAR* from different aspects. First, the relevance judgments among assessors have substantial agreement (average Kappa score of 0.71), suggesting that we managed to collect high-quality judgments via crowdsourcing. Second, we evaluated a number of IR systems for each supported task and found a high correlation between the performance of systems over the full and sampled versions. Therefore, researchers who do not have resources to experiment with the full collection can also use the sampled version. Moreover,

the presented performance results exhibit reference points for future studies. Finally, *EveTAR* is comparable to TREC tweet test collections in many aspects including reliability of ranking the tested systems.

There are several possible directions for future work. First, the introduction of *EveTAR* opens the door for building Arabic-specific retrieval systems on Twitter. To further encourage the employment of *EveTAR* and advance the research on Arabic IR, we plan to organize a shared-task evaluation campaign around the tasks supported by *EveTAR*, where the performance results presented in this work would serve as strong baselines. Second, with the success of the evaluation-as-a-service paradigm (Paik and Lin 2016), we are planning to develop a search API to allow direct access to the full collection without the need of re-crawling. Finally, we plan to extend the number of labeled dialectal tweets and identify their exact dialects (e.g. Yemeni) in order to enable the evaluation of IR systems working on dialect-related tasks such as cross-dialect search.

References

Almerekhi, H., Hasanain, M., Elsayed, T. (2016). Evetar: A new test collection for event detection in arabic tweets. In *Proceedings of the 39th International ACM SIGIR conference on Research and Development in Information Retrieval*, SIGIR '16, (pp. 689–692).

Alsaedi, N., Burnap, P. (2015). Arabic event detection in social media. In *Computational Linguistics and Intelligent Text Processing*, (pp. 384–401).

Alsaedi, N., Burnap, P., Rana, O. (2016). Sensing Real-World Events using Arabic Twitter Posts. In *Proceedings of the tenth International AAAI Conference on Web and Social Media*, ICWSM '16, (pp. 515–518).

Aslam, J., Diaz, F., Ekstrand-Abueg, M., Pavlu, V., Sakai, T. (2013). Trec 2013 temporal summarization. In *Proceedings of the 22nd Text REtrieval Conference*, TREC '13

Azmi, A. M., Alshenaifi, N. A. (2016). Answering arabic why-questions: Baseline vs. rst-based approach. *ACM Transactions on Information Systems* 35(1):6:1–6:19, DOI 10.1145/2950049, URL http://doi.acm.org/10.1145/2950049

Bagdouri, M., Oard, D. W. (2015). On predicting deletions of microblog posts. In *Proceedings of the 24th ACM International on Conference on Information and Knowledge Management*, CIKM '15, (pp. 1707–1710).

Baruah, G., Roegiest, A., Smucker, M. D. (2014). The effect of expanding relevance judgments with duplicates. In *Proceedings of the 37th International ACM SIGIR Conference on Research & Development in Information Retrieval*, SIGIR '14, (pp. 1159–1162).

Becker, H., Naaman, M., Gravano, L. (2011). Beyond trending topics: Real-world event identification on Twitter. In *Technical Report cucs-012-11*, Columbia University

Benhardus, J., Kalita, J. (2013). Streaming trend detection in twitter. *International Journal of Web Based Communities* 9(1):122–139

Bodoff, D. (2008). Test theory for evaluating reliability of ir test collections. *Information Processing & Management* 44(3):1117 – 1145

Bodoff, D., Li, P. (2007). Test theory for assessing ir test collections. In *Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '07, (pp. 367–374).

Brennan, R. L. (2001). *Generalizability Theory*. Springer New York

Bruns, A., Highfield, T., Burgess, J. (2013). The Arab Spring and Social Media Audiences English and Arabic Twitter Users and Their Networks. *American Behavioral Scientist* 57(7):871–898

Carterette, B., Allan, J., Sitaraman, R. (2006). Minimal test collections for retrieval evaluation. In *Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '06, (pp. 268–275).

Cormack, G. V., Palmer, C. R., Clarke, C. L. A. (1998). Efficient construction of large test collections. In *Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '98, (pp. 282–289).

Darwish, K., Magdy, W. (2014). *Arabic information retrieval*. Now Publishers

Darwish, K., Magdy, W., Mourad, A. (2012). Language processing for arabic microblog retrieval. In *Proceedings of the 21st ACM International Conference on Information and Knowledge Management*, CIKM '12, (pp. 2427–2430).

Efron, M., Winget, M. (2010). Query polyrepresentation for ranking retrieval systems without relevance judgments. *Journal of the American Society for Information Science and Technology* 61(6):1081–1091

Eldesouki, M., Dalvi, F., Sajjad, H., Darwish, K. (2016). Qcri @ dsl 2016: Spoken arabic dialect identification using textual features. In *Proceedings of the Third Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial3)*, The COLING 2016 Organizing Committee, (pp. 221–226).

Fleiss, J. L. (1971). Measuring nominal scale agreement among many raters. *Psychological bulletin* 76(5):378

Gey, F. C., Oard, D. W. (2001). The trec-2001 cross-language information retrieval track: Searching arabic using english, french or arabic queries. In *Proceedings of the Tenth Text REtrieval Conference*, TREC 2001

Ghosh, S., Ghosh, K. (2016). Overview of the fire 2016 microblog track: Information extraction from microblogs posted during disasters. *Working notes of FIRE* (pp. 7–10).

Glenn, M., Strassel, S., Kong, J., Maeda, K. (2006). Tdt5 topics and annotations. *Linguistic Data Consortium (LDC)*

Hasanain, M., Elsayed, T. (2014). QU at TREC-2014: Online Clustering with Temporal and Topical Expansion for Tweet Timeline Generation. In *Proceedings of the 23rd Text REtrieval Conference*, TREC '14

Jones, K. S., Van Rijsbergen, C., Research, B. L., Department, D. (1975). *Report on the Need for and Provision of an Ideal Information Retrieval Test Collection*. British Library Research and Development reports, University Computer Laboratory

Kunneman, F., Bosch, A. V. D. (2014). Event detection in Twitter: A machine-learning approach based on term pivoting. In *Proceedings of the 26th Benelux Conference on Artificial Intelligence*, BNAIC'14, (pp. 65–72).

Lin, J., Efron, M. (2013). Overview of the trec-2013 microblog track. In *Proceedings of the 22nd Text REtrieval Conference*, TREC '13

Lin, J., Efron, M., Wang, Y., Sherman, G. (2014). Overview of the trec-2014 microblog track. In *Proceedings of the 23rd Text REtrieval Conference*, TREC '14

Lin, J., Efron, M., Wang, Y., Sherman, G., Voorhees, E. (2015). Overview of the trec-2015 microblog track. In *Proceedings of the 24th Text REtrieval Conference*, TREC '15

Lin, J., Roegiest, A., Tan, L., McCreadie, R., Voorhees, E., Diaz, F. (2016). Overview of the trec 2016 real-time summarization track. In *Proceedings of the 25th Text REtrieval Conference*, TREC '16

Liu, X., Li, Q., Nourbakhsh, A., Fang, R., Thomas, M., Anderson, K., Kociuba, R., Vedder, M., Pomerville, S., Wudali, R., et al. (2016). Reuters tracer: A large scale system of detecting & verifying real-time news events from twitter. In *Proceedings of the 25th ACM International on Conference on Information and Knowledge Management*, CIKM '16, (pp. 207–216).

Liu, Y., Kliman-Silver, C., Mislove, A. (2014). The tweets they are a-changin: Evolution of twitter users and behavior. In *Proceedings of the Eighth International AAAI Conference on Web and Social Media*, ICWSM '14

Ma, Z., Sun, A., Yuan, Q., Cong, G. (2014). Tagging your tweets: A probabilistic modeling of hashtag annotation in twitter. In *Proceedings of the 23rd ACM International Conference on Conference on Information and Knowledge Management*, CIKM '14, (pp. 999–1008).

Magdy, W., Elsayed, T. (2016). Unsupervised adaptive microblog filtering for broad dynamic topics. *Information Processing & Management* 52(4):513–528

Magdy, W., Elsayed, T., Hasanain, M. (2016). *On the Evaluation of Tweet Timeline Generation Task*, Springer International Publishing, (pp. 648–653).

McMinn, A. J., Moshfeghi, Y., Jose, J. M. (2013). Building a Large-scale Corpus for Evaluating Event Detection on Twitter. In *Proceedings of the 22Nd ACM International Conference on Information & Knowledge Management*, CIKM '13, (pp. 409–418).

Mitra, T., Gilbert, E. (2015). Credbank: A large-scale social media corpus with associated credibility annotations. In *Proceeding of the Ninth International AAAI Conference on Weblogs and Social Media*, ICWSM '15

Moffat, A., Scholer, F., Thomas, P., Bailey, P. (2015). Pooled evaluation over query variations: Users are as diverse as systems. In *Proceedings of the 24th ACM International on Conference on Information and Knowledge Management*, CIKM '15, (pp. 1759–1762).

Nakov, P., Màrquez, L., Magdy, W., Moschitti, A., Glass, J., Randeree, B. (2015). Semeval-2015 task 3: Answer selection in community question answering. In *Proceedings of the 9th International Workshop on Semantic Evaluation*, Association for Computational Linguistics, SemEval '15, (pp. 269–281).

Nakov, P., Màrquez, L., Moschitti, A., Magdy, W., Mubarak, H., Freihat, a. A., Glass, J., Randeree, B. (2016). Semeval-2016 task 3: Community question answering. In *Proceedings of the 10th International Workshop on Semantic Evaluation*, Association for Computational Linguistics, SemEval '16, (pp. 525–545).

Nakov, P., Hoogeveen, D., Màrquez, L., Moschitti, A., Mubarak, H., Baldwin, T., Verspoor, K. (2017). SemEval-2017 task 3: Community question answering. In *Proceedings of the 11th International Workshop on Semantic Evaluation*, Association for Computational Linguistics, SemEval '17

Oard, D. W., Gey, F. C. (2002). The trec 2002 arabic/english clir track. In *Proceedings of the Eleventh Text REtrieval Conference*, TREC 2002

Ounis, I., Macdonald, C., Lin, J., Soboroff, I. (2011). Overview of the trec-2011 microblog track. In *Proceedings of the 20th Text REtrieval Conference*, TREC '11

Paik, J. H., Lin, J. (2016). Retrievability in api-based "evaluation as a service". In *Proceedings of the 2016 ACM International Conference on the Theory of Information Retrieval*, ICTIR '16, (pp. 91–94).

Pavlu, V., Aslam, J. (2007). A practical sampling strategy for efficient retrieval evaluation. Tech. rep., Northeastern University

Peñas, A., Hovy, E., Forner, P., Rodrigo, Á., Sutcliffe, R., Sporleder, C., Forascu, C., Benajiba, Y., Osenova, P. (2012). Overview of qa4mre at clef 2012: Question answering for machine reading evaluation

Petrovìc, S. (2013). Real-time event detection in massive streams. PhD thesis, School of Informatics, University of Edinburgh

Petrovìc, 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*, Association for Computational Linguistics, NAACL HLT'10, (pp. 181–189).

Petrovìc, S., Osborne, M., Lavrenko, V. (2012). Using paraphrases for improving first story detection in news and twitter. In *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, Association for Computational Linguistics, NAACL HLT '12, (pp. 338–346).

Qin, Y., Zhang, Y., Zhang, M., Zheng, D. (2013). Feature-rich segment-based news event detection on twitter. In *Proceedings of the Sixth International Joint Conference on Natural Language Processing*, (pp. 302–310).

Rowe, M., Stankovic, M. (2012). Aligning tweets with events: Automation via semantics. *Semantic Web* 3(2):115–130

Salem, F. (2017). The Arab Social Media Report 2017: Social Media and the Internet of Things: Towards Data-Driven Policymaking in the Arab World (Vol. 7). Tech. rep., Mohammed Bin Rashid School of Government, Dubai, UAE

Sanderson, M. (2010). *Test collection based evaluation of information retrieval systems*. Now Publishers Inc

Shamma, D. A., Kennedy, L., Churchill, E. F. (2011). Peaks and persistence: Modeling the shape of microblog conversations. In *Proceedings of the ACM 2011 Conference on Computer Supported Cooperative Work*, CSCW '11, (pp. 355–358).

Sim, J., Wright, C. C. (2005). The kappa statistic in reliability studies: use, interpretation, and sample size requirements. *Physical therapy* 85(3):257–268

Soboroff, I., Ounis, I., Macdonald, C., Lin, J. (2012). Overview of the trec-2012 microblog track. In *Proceedings of the 21st Text REtrieval Conference*, TREC '12

Strassel, S., Kong, J., Graff, D. (2005). Tdt4 multilingual text and annotations. *Linguistic Data Consortium, Philadelphia*

Suwaileh, R., Hasanain, M., Elsayed, T. (2016). Light-weight, conservative, yet effective: Scalable real-time tweet summarization. In *Proceedings of the 25th Text REtrieval Conference*, TREC '16

Teevan, J., Ramage, D., Morris, M. R. (2011). # twittersearch: a comparison of microblog search and web search. In *Proceedings of the fourth ACM international conference on Web search and data mining*, ACM, (pp. 35–44).

Urbano, J., Marrero, M., Martín, D. (2013). On the measurement of test collection reliability. In *Proceedings of the 36th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '13, (pp. 393–402).

Wang, Y., Sherman, G., Lin, J., Efron, M. (2015). Assessor differences and user preferences in tweet timeline generation. In *Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '15, (pp. 615–624).

Wasike, B. S. (2013). Framing news in 140 characters: How social media editors frame the news and interact with audiences via Twitter. *Global Media Journal* 6(1)

Zhang, Y., Szabo, C., Sheng, Q. Z. (2015). *Sense and Focus: Towards Effective Location Inference and Event Detection on Twitter*, Springer International Publishing, (pp. 463–477).

Zhao, Z., Mei, Q. (2013). Questions about questions: An empirical analysis of information needs on twitter. In *Proceedings of the 22nd international conference on World Wide Web*, ACM, (pp. 1545–1556).

Zuccon, G., Palotti, J., Hanbury, A. (2016). Query variations and their effect on comparing information retrieval systems. In *Proceedings of the 25th ACM International on Conference on Information and Knowledge Management*, CIKM '16, (pp. 691–700).