Adaptive Method for Following Dynamic Topics on Twitter

Walid Magdy
Qatar Computing Research Institute
Qatar Foundation
Doha, Qatar
wmagdy@qf.org.qa

Tamer Elsayed
Department of Computer Science and Engineering
College of Engineering, Qatar University
Doha, Qatar
telsayed@qu.edu.qa

Abstract

Many research social studies of public response on social media require following (i.e., tracking) topics on Twitter for long periods of time. The current approaches rely on streaming tweets based on some hashtags or keywords, or following some Twitter accounts. Such approaches lead to limited coverage of on-topic tweets. In this paper, we introduce a novel technique for following such topics in a more effective way. A topic is defined as a set of well-prepared queries that cover the static side of the topic. We propose an automatic approach that adapts to emerging aspects of a tracked broad topic over time. We tested our tracking approach on three broad dynamic topics that are hot in different categories: Egyptian politics, Syrian conflict, and international sports. We measured the effectiveness of our approach over four full days spanning a period of four months to ensure consistency in effectiveness. Experimental results showed that, on average, our approach achieved over 100% increase in recall relative to the baseline Boolean approach, while maintaining an acceptable precision of 83%.

Introduction

Social media have been of interest for many researchers in the last years, since they can be used to measure the public response or interest towards given topics happening in real world. Twitter is one of the most studied social media platforms due to the large amount of real-time exchanged information by users in the form of short messages (tweets) that are publically available to everyone. This large amount of communicated information motivated many social scientists to study the public response towards different events and entities over tweets. However, retrieving relevant tweets on a given topic of study requires a scalable and adaptive topic tracking techniques, since the topics to be tracked are usually of broad and dynamic nature. For example, following tweets related to "Presidential Elections" in a given country requires tracking tweets about several sub-topics including candidates, campaigns, political views, election process, etc. Moreover, the sub-topics are also dynamic, e.g., "debates" is an important one before the elections, while "election results" is the most important one during voting; and sometimes span a very short period of time, e.g., press statements by candidates. Similarly, in a study about public response to a long-term event such as "the Syrian conflict", tracking relevant tweets is not a straightforward task, since it requires the coverage of as much of the related posted content as possible to cope with the developing sub-events. Such type of topics, which last for a long period of time and consist of sub-events that change dramatically over time, requires a *set* of queries, rather than just a specific one, to be updated periodically to effectively capture relevant tweets.

Two common features are provided by Twitter and social media in general for tracking tweets. The first is the "follow" feature that allows a user to follow other accounts of entities, persons, or events to get their tweets into the user's timeline. The other method for following specific tweets is searching for given hashtags, which is a common way for users to get updates on topics that are indicated by those hashtags. This method is less strict in tracking information, where more tweets are generally presented to user. However, many off-topic tweets would be retrieved because of the misuse of hashtags by some users. Moreover, many tweets that are relevant to the topic may not include the hashtag itself, and thus will be missed.

In this paper, we present an unsupervised approach for tracking short messages from Twitter that are relevant to broad and dynamic topics. Our main objective is to achieve high *recall* by retrieving a large number of relevant tweets, while preserving high *precision* to avoid bothering users with irrelevant feeds. The main challenge lies in capturing relevant tweets to temporal short-term sub-topics that might only appear for a short period of time. Our approach initially gets a set of user-defined fixed accurate (Boolean) queries that cover the most static part of the topic. These queries are selected carefully to retrieve an initial set of potentially relevant tweets with high precision. The initial retrieved set of tweets is used in a novel, *unsupervised*, and *adaptive* manner to train a binary classifier that is used for detecting other relevant content within a stream of tweets. This classifier is trained and updated automatically and

regularly to adapt to the dynamic nature of the tracked topic without any user intervention.

We tested the approach on three hot topics from different domains: Politics ("Egyptian Politics", defined by 84 queries), War ("Syrian Conflict", defined by 42 queries), and Sports ("International Sports", defined by 96 queries). A stream of 3 to 4 million Arabic tweets per day was used in our experiments with the goal of identifying tweets that are relevant to each of the three broad topics. We evaluated our tracking approach on four days of the stream selected from four different months, to measure the consistency of the approach over time. Our approach achieved a significant increase in the number of identified relevant content without degrading precision so much, compared to a classical approach that applies Boolean filtering using a fixed set of queries. Experimental results showed high effectiveness of our approach, where we managed to increase the number of relevant tweets by more than 100%, while precision is decreased by less than 14%.

Our contributions in this work are as follows:

1. Proposing a novel unsupervised adaptive method for following broad dynamic topics.
2. Introducing a practical evaluation methodology for topic tracking in large data streams.
3. Developing 3 broad topics, all in Arabic, with about 6000 relevance judgments and making the set publicly-available for related research studies.

Related Work

With the recent wide spread of online social network platforms and the huge amount of tweets that are communicated instantly, the users have to cope efficiently with a huge number of tweets on several topics in her timeline. Therefore the problem of topic tracking has emerged once again, but in a different context that has different characteristics than the classical one.

Among the studies that have been conducted recently, two basic categories are identified based on the type of the tracked topics: broad-topics and focused-topics tracking. The former is concerned with more general and long-standing topics that are composed of several sub-topics (e.g., "soccer"), while the latter deals with topics that are related to narrower scope (e.g., "iphone5c").

Tracking Broad Topics

Different approaches have been attempted to tackle the problem of tracking broad topics in Twitter. Sriram et al. (2010) used manually-labelled tweets to train a naïve Bayes classifier to classify tweet stream into general broad topics such as news, opinions, events, deals, and private messages. Phuvipadawat and Murata (2010) focused on tracking and detection of breaking news using predefined search queries, e.g., #breakingnews and "breaking news". They adopted an unsupervised filtering based on clustering similar incoming tweets with emphasis on proper nouns and significant nouns and verbs to track and adapt to developing stories. Duan et al. (2012) proposed a graph-based approach in a graph optimization framework for classification of tweets into six broad topics: entertainment, politics, science and technology, business, lifestyle, and sports. Related tweets, the ones that share either the same hashtag or URL, were used to enrich the representation of each tweet and adaptively update the trained model using the hashtags as a surrogate for user feedback. Adaptive language modelling techniques were used by Lin et al. (2011) to track broad topics represented by most common Hashtags, e.g., "Apple", "Fashion", and "American football". Compared to our proposed approach where the broad topics are defined by a list of predefined queries, hashtags were used in their experiments to identify the topic; and tweets that have the hashtag were used as training data. Several smoothing techniques were tried to integrate both the background model (leveraged to overcome sparsity) and the foreground model (leveraged to capture recency). In a recent work (Magdy et al., 2012; Magdy, 2013), *TweetMogaz* was presented as a news portal of tweets that are relevant to broad topics defined by predefined queries. It leverages external news source to expand the topic profile with more timely phrases, however, no evaluation has been reported on the effectiveness of the application.

The purpose of collecting the tweets for the targeted broad topics extends beyond reading them; it actually includes analysis, summarization, or classification. Some of the applications for such task are news monitoring (Magdy, 2013), crisis management (Vieweg et al., 2010), customer feedback detection (Chen at al., 2013), public political polarization analysis (Conover, 2011), election results prediction (Tumasjan et al., 2010), and public healthcare monitoring tools (Ali et al., 2013). All of these applications require the analysis of a large number of social posts, mainly tweets. To get that amount of relevant tweets with minimum noise, a relatively high-recall high-precision information filtering system is needed, and achieving that level of performance is a key objective of our task.

Tracking Focused Topics

Studies on tracking focused-topics spanned several contexts. Dan et al. (2011) aimed at following tweets that are related to specific TV shows. They proposed a bootstrapping approach that uses domain-knowledge and a 2-stage semi-supervised training of a classifier. Along similar line, Chen et al. (2013) realized that keyword-based Boolean filtering is not so effective for tracking tweets that express customer opinions about commercial brands (e.g., Delta Airlines). As we enriched our tracking approach by using expansion terms in training an unsupervised binary classifier, alternatively they leveraged crowd-sourcing

resources to label tweets that satisfy predefined queries to train a supervised binary classifier. Tackling the problem of "concept drift", Nishida et al. (2012) detected changes in word probabilities over time in a probabilistic classification approach. Hashtags on baseball teams or television networks were used as labels for training the classifier. In a more recent study, Albakour et al. (2013) proposed an unsupervised approach that exploits traditional query expansion to combat sparsity and a decay smoothing technique besides event detection through either tweet or newswire stream to handle topic drift. More work on tracking tweets on focused topics has been reported for the TREC microblog track that has introduced a filtering task for the first time in 2012 (Soboroff et al., 2012). Most of the experiments adopted either a supervised learning approach using manually labelled tweets for training (Zhang et al., 2012), or a Rocchio-based approach that is updated using the available relevance judgments (Soboroff et al., 2012).

Following Dynamic Topics

Dynamic Topics Definition

In this paper, we focus mainly on following broad dynamic topics. This task is of interest to many social scientists and institutes that strive to track public posts on social media, such as governments, news agencies, and crisis managements organizations. The monitored topics on social media are typically: *broad* requiring several queries for tracking; *dynamic* requiring frequent updates to the queries; and lasting for long periods of time requiring a robust approach with minimal user intervention.

A broad dynamic topic t_{BD} in Twitter is represented by a *static set* Q_0 of queries $\{q_1, q_2, \ldots, q_n\}$, each expressed in keywords, phrases, microblog-user accounts, or hashtags. The set of queries is expected to cover a wide space of the broad topic. An upcoming stream of tweets is processed and only relevant tweets to t_{BD} are presented based on the given Q_0, without the need to manually update the underlying set of queries $\{q_1, q_2, \ldots, q_n\}$. No user feedback is given during the tracking process.

The main objective of the tracking approach is to achieve higher recall levels than standard Boolean search methods, to cover more aspects of these topics while maintaining high precision to minimize the amount of noise within the collected tweets. In addition, the approach should be self-adapting to cope with changes occurring to the topics over time, since these topics are highly dynamic.

Tweets Tracking Approach

The simplest technique for tracking tweets relevant to a topic t_{BD} is using Q_0 as a set of *Boolean* queries in a *Boolean* filtering method to track tweets that satisfy any of the queries in Q_0 in the upcoming stream. We call the resulting matched tweets T_B. The quality of T_B would be dependent on the quality of the selected queries in Q_0; if they are selected precisely, it is expected to retrieve results of very high precision; however recall is expected to be low. In addition, emerging subtopics in t_{BD} are expected to be missing from T_B since they are not covered by Q_0.

Our tracking approach relies on automatically detecting new temporal subtopics in t_{BD} and building a classifier to detect relevant tweets to these subtopics in addition to the ones captured by Q_0.

Since Boolean filters are strict, a classifier-based filter can be a good choice for improving recall. To train a binary classifier, samples of positive and negative examples are required. In our problem, it is straightforward to use the set of tweet T_B that match the carefully defined Q_0 as the positive sample, since it is expected to be of high precision. A *random* sample T_{rand} of the tweets that do not match Q_0 can then act as the negative sample. The trained classifier is then used to classify the stream of tweets into relevant or irrelevant. We refer to the resulting classifier as f_C. However, a main concern about f_C is the potential risk of having relevant tweets within the randomly-selected negative tweets sample, since many relevant tweets still do not match Q_0, especially the temporal subtopics. This would lead to a trained classifier that is confused over good features.

To overcome the risk of having possible relevant tweets within T_{rand}, an exclusion process to potentially relevant tweets in T_{rand} is applied using the following steps:

1. Potentially relevant terms that are not included in Q_0 are extracted from T_B using equation (1)

$$TF\_IDF_w(t) = tf_B(t).\log\frac{N_w}{df_w(t)} \quad (1)$$

   where $TF\_IDF_w(t)$ is the TFIDF of term t in a given window of time w. In our context, w should precede the beginning of the online tracking process, resembling a training-like period. $tf_B(t)$ is the total term frequency of term t within T_B; $df_w(t)$ is the number of tweets in w that contained the term t; and N_w is the total number of tweets in w.
2. The top k terms achieving the highest score using equation (1) and are not included in Q_0 are selected to form the potentially relevant terms to topic t_{BD} at the time windows w. These terms are used as *exclusion* terms E, to clean T_{rand} used as the negative sample to train the classifier. It is expected that for different time windows w for the same topic t_{BD}, the set of E terms will be different, since the topic is broad and highly dynamic with subtopics changing over time.
3. After filtering out all tweets matching any of the terms in E from T_{rand}, the resulting set is denoted by T_N, which is used as the "clean" negative sample, along with T_B as the positive sample, to train the

classifier. We refer to the final trained classifier as f_{CE}.

The previous process tries to minimize the chance that the negative sample might contain tweets that are possibly relevant to the topic. The rationale is that E can match many potentially-relevant tweets, and thus can "clean" the negative sample from those noisy examples that would negatively affect the performance of the classifier. The fact that it might also match irrelevant tweets as well, and thus exclude them from the negative sample, should not have negative effect since T_{rand} is naturally of no shortage of non-relevant tweets.

In practical use of f_{CE}, the processes of term selection and classifier training are applied periodically to frequently adapt to the expected changes and drifts in the targeted topic. Both the window of time w for collecting the training samples and the frequency of updating the classifier depend on the dynamicity and broadness of the topic. For example, in our experimentation, we set w to 20 hours of tweets stream, and we update the classifier every 4 hours by retraining it on the past w hours. Tweets in the time window w used for training samples extraction is referred to as T_w, and tweets stream that is processed to track relevant tweets is referred to as T_s. Figure 1 sketches the algorithm used in our approach.

In our experiments, we have used support vector machines (SVM) classifier (Joachims, 2002). Each tweet is represented as a feature vector. Terms, including hashtags and user accounts, are used as the features and feature values are all binary, based on the existence of the terms in the tweet. Since the classifier is trained periodically, it is expected that the set of terms used as features change over time as the training samples change. For an efficient process, we reduce the feature space by selecting only the terms that appear more than 10 times in T_B as the features, after removing stop words[1]. Terms that appear in a tweet but not in the feature space are represented by an additional special feature, denoted by *miss*, which is defined as the percentage of terms in the tweet that do not exist as features in the feature space. For example, if a tweet has 20 terms after stop-word removal, and only 5 terms exist in the feature space, then the corresponding features of the 5 existing terms will be set to one, and *miss* will be set to 0.75 (i.e., 15/20).

We note that the set of features are different from one topic to the other and from one training instance to another, even for the same topic, since it depends on the terms appearing in the set T_B, which changes periodically. We also elect to set the size of the negative samples to be ten times the size of the positive sample, to better cover the wide space of the non-relevant tweets.

[1] http://members.unine.ch/jacques.savoy/clef/index.html

Input: Predefined Queries Q_0, training tweet stream T_w, testing tweet stream T_s
Output: Filtered tweets T_f, Classifier f_{CE}

```
Pos ← booleanFilter(Tw, Q0)      ► get positive sample
E ← termSelection(Pos)           ► select exclusion terms
TE ← booleanFilter(Tw, E)        ► get excluded tweets
Trand ← randomSample(Tw)         ► get a random sample
Neg ← Trand - TE                 ► exclude some tweets
Classifer fCE ← trainClassifier(Pos, Neg)
                                 ► train classifier
Tf ← testClassifier(fCE, Ts)     ► track tweets
```

Figure 1. Our tweets tracking algorithm.

Experimental Setup

Preparing experiments for evaluating a real-life tweets tracking task for following broad dynamic topics was not straightforward. We put some criteria for our experimental setup to ensure an adequate evaluation for the tracking technique. These criteria were:

1. Using real-life microblog stream. To correctly model our tracking task, we have to use a live stream of microblogs that provides large number of microblogs per second as in practical applications.

2. Testing on multiple broad topics from different domains. Although the preparation of a broad dynamic topic requires some effort (to create tens of queries), we had to create more than one topic to measure the performance consistency of our topic tracking technique over multiple topics.

3. Monitoring performance over long period of time. Since our approach relies on preparing the set Q_0 that represents the most static part of a dynamic topic, we were keen to measure the effectiveness of our tracking approach over long period of time using the same Q_0 without manually-updating it. The objective of this criterion is to measure the adaptability of our tracking technique to the dramatic changes that occur to the targeted topics.

Due to the lack of existing test collections for the task, the preparation of the test data, based on the above criteria, required a considerable effort. A large stream produced a large amount of tweets to be classified for each test topic, and each single run in our experiments, resulted in a different filtered set of tweets which required separate relevance assessments. There were three dimensions of experimentation that led to several runs to be evaluated: different topics, different time snippets, and different configurations (parameters) of the tracking approach itself. In this section, we discuss the used data stream, the created topics, the experimented runs, and the adopted evaluation methodology.

Topics and Data Stream

In our experiments, we selected three broad topics of different domains to test our tracking approach. We selected two topics that are considered among the hottest worldwide in 2013: "Egyptian Politics" (denoted by *EGY*) and "Syrian Conflict" (denoted by *SYR*). Additionally, we selected a more general topic, "International Sports" (denoted by *SPO*) as our third topic.

For the microblog data stream, we used the Twitter search API with the general query "lang:ar" to stream tweets of Arabic content. We selected the Arabic stream for two reasons:

1. Two of the selected topics are related to the Arabic region; therefore it is expected to find large number of relevant tweets about the topics in Arabic language.
2. The limitation of the freely available Twitter API allows streaming of no more than 1% of the publicly-available tweets. Getting the 1% sample in English or without specifying the language would lead to a highly sparse stream. It was important to restrict the stream to be focused, general, and unbiased. Restricting the 1% sample to Arabic tweets can get up to 50% of the Arabic posts on Twitter (Arab Social Media Report, 2013). In this case, the stream is general enough, since no specific terms are used for streaming, and it potentially has decent number of relevant content to our targeted topics.

Twitter API was used for streaming the Arabic tweets. The average number of tweets received per second was 37 tweets; making an average of 3.2 million Arabic tweets collected per day. We collected tweets of four full days, each from a different month in February, March, April, and May of 2013. The day is counted between 12:00:00am GMT to 11:59:59pm GMT.

Arabic text of the collected tweets stream is pre-processed using state-of-the-art normalization technique for social Arabic text (Darwish et al., 2012) to facilitate the matching process between queries and tweets. This normalization technique includes letter normalization, diacritics removal, decorative text replacement, and word elongation resolving (Darwish et al., 2012).

We asked three Arabic-speaking volunteers in our institute, each interested in one of the three selected topics, to prepare a set of queries that represent them. We refer to the volunteers as *topic developers*. An Egyptian researcher, Syrian program manager, and a professor prepared the set of queries of the *EGY*, *SYR*, and *SPO* topics respectively. Each set of the prepared queries represents the most static parts of the corresponding topic. For example, entities such as persons (e.g. politicians and players), institutes (e.g. political parties, fighting groups, and soccer teams), and twitter accounts (e.g., accounts of politicians or dedicated news providers). Queries representing temporal events were not used. We asked our topic developers to use precise queries that should retrieve relevant tweets with very high probability. *Boolean* queries were allowed, where *AND* and *OR* operators were used in some of the queries to make it more precise. We asked them to use each of the queries to search Twitter and discard those that return any non-relevant tweet among the top 10 tweets to ensure the precision of the individual queries. The number of queries prepared to cover each topic and some sample queries are presented in Table 1.

Table 1 shows a large number of queries prepared for each test topic. Queries were prepared carefully, similar to the situation when preparing a set of terms to find posts relevant to a topic for a social study or analysis. Despite this large number of queries, we believe that the coverage is not very high for the topic, since only non-temporal static part of the topic was covered. Other potential queries that are relevant, but can also match non-relevant results were excluded. An example to this for *EGY* topic is the word or hashtag "#Egypt" in both English and Arabic. This hashtag exists in many posts related to the Egyptian politics, but also in posts that do not relate to politics or even do not relate to Egypt altogether. Another example, the query "Muslim brotherhood Egypt" shown in the sample is too restrictive and thus has a low coverage of the microblogs talking about the Muslim brotherhood group in Egypt, since people usually mention "brotherhood" only in tweets. However, the word "brotherhood" itself is very ambiguous, and can refer to the brotherhood between friends and brothers, or it can even refer to the Muslim brotherhood group but in different countries other than Egypt. This is why the topic developer restricted the matching to the full description. In fact, the previous examples highlight the challenges of tweets tracking on broad dynamic topics, and show the need for an effective approach that captures different terms used to describe these topics over time.

Experimental Runs

After preparing the topics and the collection, we applied the tracking approach to the collected data streams. For each test day, the tweets in the first 20 hours were used as a static collection for training (i.e., w = 20 hours) and the tweet stream in the last 4 hours was used for testing. In addition of using our tracking approach $f_{CE}(k)$, we applied three baseline runs:

- f_B: which is the Boolean filtering approach used in most of the social studies.
- f_C: which is training the classifier as in f_{CE}, but with using T_{rand} directly as the negative sample without cleaning tweets matching E terms.
- $f_{BE}(k)$: is applying Boolean filtering, but after expanding Q_0 with the potentially relevant terms E.

The parameter k indicates the number of expansion terms in f_{BE} and number of exclusion terms in f_{CE}. The values of k tested in our experiments were 50, 100, and 200 terms.

The identified tweets in the 4 hours test stream by the f_B tracking method is referred to as $\bar{T}_B$ to distinguish it from

T_B that refers to tweets matching f_B in the 20 hours train stream and used as positive samples for training the f_C and $f_{CE}(k)$ classifiers.

Building Relevance Assessments

Building the relevance assessments for this task was a real challenge since the number of tweets classified as relevant to each topic in each day was considerably large. The number was always in thousands for each topic in each day for every tested run. It was almost impossible to exhaustively assess this large number of tweets for relevance.

We initially attempted to utilize crowdsourcing platforms for assessing the resulting tweets of our runs. Considering that the tweets to be assessed are in Arabic, we submitted a pilot run on crowdflower[2]. A set of 500 tweets randomly selected from different runs for *EGY* topic was submitted and we asked for at least three people to assess the relevance of each tweet. Although the task was finished quickly, the average agreement among assessors was 75%, which we considered low. We then asked to topic developer to assess the 500 tweets carefully himself. We found that 19% of the total relevance decisions by crowdsourcing were incorrect. For this kind of task, that was unacceptable, and thus we had to resort to a more efficient sampling strategy and more reliable relevance assessments.

To evaluate our runs, we sampled the classified-as-relevant tweets by selecting 100 tweets from each run for each topic and in each streaming period (i.e., the last 4 hours in each test day). If we *randomly* sampled the tweets resulting from each run, we would produce a large assessment job since there would be a very low chance of sampling common tweets across different runs. To alleviate this problem, we adopted a *guided* sampling strategy; we ranked the tweets resulting from each run based on the number of their retweets (counted as the number of repetitions in the classified-as-relevant set), and then selected the top 100 (of each run) for relevance assessment. This would increase the probability of having common tweets among different runs in the assessment sample, hence reducing the number of tweets to be assessed. Moreover, the selected 100 tweets represent more tweets in the stream, since they are highly frequent. We argue that this guided sampling strategy would not lead to a biased evaluation; the reason is that all of our baseline runs and the tracking approach are completely independent of the retweets information. To experimentally prove it, we measured the Kendall-Tau correlation (Kendall, 1938) between the ranking of the sampled tweets based on the classification score and their ranking based on their retweets, and we repeated that for each run. The correlation ranged from -0.1 to 0.1 for all the runs. This indicates that the retweets were uncorrelated with the classification scores and hence the selected sample was not biased towards the tweets of high classification score. In addition to the testing sample, we also selected the top 100 retweeted tweets in the training set T_B to estimate the accuracy of the positive sample used to train the classifier-based approaches.

Finally, all sampled tweets from all runs on the same topic and day are merged together and randomly shuffled before being handed to the assessors. To ensure more reliable relevance assessments, we asked the topic developers to act as topic assessors too and assess the relevance of their corresponding merged samples. Before performing the assessments, we asked each assessor to take a quick look on the events related to the topic that happened on the date of the test day. That would ensure they are aware of the different aspects of each topic. The total number of tweets to be assessed for each topic was 1922, 1923, and 2051 tweets for the *EGY*, *SYR*, and *SPO* topics respectively. The average number to be assessed for each topic for each test day was around 500 tweets. It took each assessor 15 to 20 hours to finish the assessments.

Table 1. Samples of the predefined queries for the test topics. Arabic queries used are shown with English translation

EGY (*Egyptian Politics*): 84 queries + 16 Twitter accounts ***Persons***: "مرسي" (Morsi), "رئيس مصر" (Egypt President), "شفيق" (Shafiq), "السيسي" (El-Sisi) ***Parties***: "الحرية والعدالة" (Freedom and Justice Party), "حزب النور" (Noor Party), "حزب الدستور" (Dostour Party) ***Groups***: "الاخوان المسلمين" (Muslim Brotherhood Egypt), "حركة ٦ ابريل" (6 April Movement), "حازمون" (Hazemon)
SYR (*Syrian Conflict*): 42 queries + 26 Twitter accounts ***Persons***: "بشار" (Bashar), "البوطي" (Al-Boti), "احمد معاذ الخطيب" (Ahmed Moaz Al-Khateeb) ***Places***: "دمشق" (Damascus), "حماة" (Hamah), "إدلب" (Idlib) ***Fighting Groups***: "الجيش الحر" (Free Army), "جيش النصرة" (Nusra Front), "انصار الشام" (Sham supporters)
SPO (*International Sports*): 96 queries + 5 Twitter accounts ***Teams***: "فريق برشلونة" (Barcelona team), "البرسا" (Barca), "ريال مدريد" (Real Madrid), "المنتخب الانجليزي" (England team) ***Players***: "مسي" (Messi), "روني" (Rooney), "مورينيو" (Mourinho) ***Sports Events***: "دوري ابطال اوروبا" (European champions league), "كأس العالم" (World Cup)

Evaluation Measures

The effectiveness of a topic tracking system can be generally measured by the precision and recall of the resulting tweets. The precision of a run is estimated by calculating the number of relevant tweets among the selected sample of 100 tweets. However, it was infeasible to calculate recall, since there is no estimation about the number of relevant tweets in a test stream. Instead, we computed the *relative* recall, denoted by $Recall_{relative}$, to measure the effectiveness of our tracking approach in retrieving additional relevant results compared to the baseline Boolean filtering method f_B.

[2] http://crowdflower.com/

Precision and relative recall of a given run x is computed as follows:

$$Precision(x) = \frac{|relevant\ tweets\ in\ sample(x)|}{100}$$

$$Recall_{relative}(x) = \frac{Classified\_Rel(x) * Precision(x)}{Classified\_Rel(f_B) * Precision(f_B)}$$

where, $Classified\_Rel(f_B) = |\bar{T}_B|$.

The overall system performance is evaluated using the mean of the calculated scores over all topics and test dates.

Unfortunately, we were not able to measure the performance in a single figure of merit that combines both recall and precision, such as F-measure or T11SU, due to the absence of the estimation of total number of relevant results in the data stream. However, we believe that considering both precision and relative recall is sufficient to measure the performance of the filtering technique.

Topic Tracking Performance

Precision of Training Data

In this subsection, we measure the quality of the prepared set of queries Q_0 for each of the topics by analysing the precision of T_B, which is used as a positive sample for training the classifiers. Precision was found to be generally high, ranging from 88% to 97% except in one case; this demonstrates how well Q_0 was generally prepared. The exception case was in one test day in April 2013 for *SYR* topic that exhibited relatively low precision. With further investigation, we realized that one of queries in Q_0, was "*@Yathalema*" representing a Twitter account of"*The Syria Relief Organisation*", which normally posts news about the Syrian conflict, however, on that day, the account was posting many tweets related to a cyber-attack by hackers on Israeli organizations. These posts were reasonably assessed as non-relevant to *SYR* by the assessor. This sample test day shows that even with carefully-prepared Q_0, there can be some cases where non-relevant tweets could be matched. The rest of mistakes occurring in T_B originated from tweets matching one of the queries in Q_0, but referring to non-relevant subtopics. Later in this section, we analyse the effect of the training data quality on the final performance of the classifier.

We also calculated the number of tweets that contained hashtags within T_B to check if hashtags can be sufficient for tracking broad topics as previously used in several related studies (Lin et al., 2011; Vieweg et al., 2010). Surprisingly, we found that only 37.8% of tweets in T_B contained hashtags on average. This shows that solely relying on hashtags would lead to a significant drop in recall.

Tweets Tracking Approach Results

Figure 2 presents the average relative recall and average precision of the our tweets tracking technique compared to

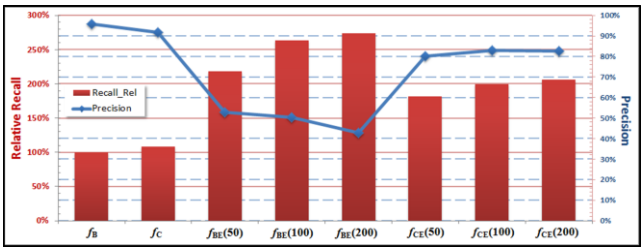

Figure 2 Average precision (on the right Y-axis) and relative recall (on the left Y-axis) achieved for tweets tracking using different approaches, tested on 3 broad topics over 4 test days.

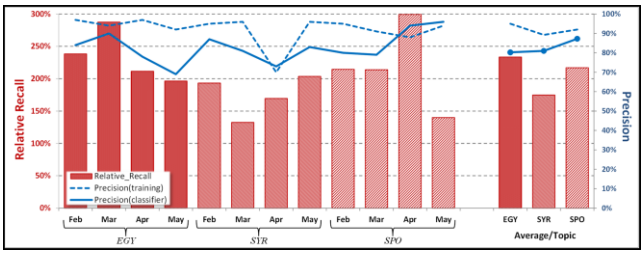

Figure 3. Topic tracking performance across different topics and over time for $f_{CE}(200)$

the three baseline techniques tested on the three topics over four different test days. Both measures are estimated from the 100-tweet sample of each run as described earlier.

As shown in Figure 2, relying only on Q_0 for tracking broad dynamic topics leads to highly precise results in most cases, but limited recall in general, compared to what could be achieved by the other approaches. This is clearly shown in the results of f_B, which achieved the highest precision, but the lowest recall among all other techniques.

f_C led to only 9% increase in the recall compared to f_B, while precision dropped from 96% to 92%. Although changes were found to be statistically significant, this change is still not major to claim a superior improvement in performance in real-life applications. Having a deeper look on the classified tweets, it was found that 94% of the tweets detected by f_B ($\bar{T}_B$) were already identified by this tweets tracking approach, which shows a strong bias of f_C to T_B used as positives samples in training, since it nearly assumes everything else as irrelevant.

Applying query expansion using different number of expansion terms k = 50, 100, and 200 without applying any classification to the tweets led to a superior increase in the number of relevant tweets retrieved, but a dramatic drop in precision. This is shown in the $f_{BE}(k)$ runs in Figure 2. Around half of the identified-as-relevant tweets by this approach were incorrect. Figure 2 shows that expanding Q_0 with additional expansion terms from E led to an estimated increase in recall of more than double. However, precision ranged from 53% to 43% when using 50 to 200 expansion terms respectively. This result indicates two important themes; firstly, there is a large number of relevant tweets in the stream that standard filtering method f_B misses the majority of it. Secondly, achieving a large increase in recall

requires more advanced techniques than simple Boolean filtering to maintain acceptable precision levels.

The last set of runs in Figure 2 presents the performance of our main proposed tracking approach for broad dynamic topics f_{CE}. As shown, a large increase in recall was achieved with a slight decrease in precision compared to f_B and f_C tracking techniques. The best run was achieved when using 200 terms for cleaning the negative samples of the trained classifier, i.e. using k=200. The achieved relative recall was 208%, indicating the retrieval of more than double of relevant tweets retrieved by f_B, and precision of 83%, which is seen acceptable for real-life applications for tracking tweets relevant to broad topics. Regarding the overlap between the classified tweets by this tracking method and $\bar{T}_B$ identified by f_B, we noticed that 95%, 97%, and 97% of tweets of $\bar{T}_B$ were also identified by $f_{CE}(50)$, $f_{CE}(100)$, and $f_{CE}(200)$ respectively. This means that most identified tweets by f_B are also identified by f_{CE}.

In addition to the previous runs, we tested applying the classifier f_{CE} to the identified tweets from the f_{BE} only instead of the full stream, i.e., classifying T_{BE} instead of the full 4 hours stream; then we add $\bar{T}_B$ directly to the classified tweets. We refer to this run $f_{CE}*(k)$, where tracking of tweets is done on two stages instead of full direct classification to the stream. The purpose of this run was to check if classifying a much smaller stream of tweets T_{BE} of relevance potential instead of classifying the full stream can lead to better performance or not. In addition, the setup of this run is more practical to the tweets streaming limitation by Twitter, which allows streaming 1% only of the full tweets stream. Therefore, tweets could be streamed based on Q_0, in addition to streaming tweets based on E terms that are updated periodically according to our approach. This run achieved almost the same exact results as $f_{CE}(k)$. Regarding processing time, $f_{CE}*(k)$ classifies much smaller number of tweets than $f_{CE}(k)$. However, the time taken for checking the presence of one of the keywords in the tweets is actually larger than classifying the tweets directly using our linear SVM classifier. In all cases, the time required for both techniques is relatively negligible compared to the time between tweets in a stream. Therefore, we consider both techniques effective and fast enough to keep up with a real-life stream.

Results in Depth

To measure the performance of the tracking technique across different topics and over time, Figure 3 presents the precision and relative recall for each of the topics in each test date individually and the average across each topic for the $f_{CE}(200)$. Furthermore, the precision of the positive samples T_B used in training the classifier is plotted to measure its effect on the classification precision. As shown, the classification always achieves a gain in the recall in all our test samples. The lowest achieved gain in recall was 32% and the largest reached 200% for some test dates, such as *EGY*-Mar and *SPO*-Apr, where relative recall was 288% and 299% respectively. Regarding precision, the classifier usually achieves lower precision than training data precision. However, it was surprisingly noticed that for the two test samples with the least training data precision, the classification precision was higher (*SYR*-Apr and *SPO*-Apr). We described before the reason of the low precision of training samples for *SYR*-Apr, this definitely affected the classifier performance. Fortunately, that the posts talking about the "Israel cyber-attack" was less in the 4 hours test stream than in the 20 hours training stream. This may be the reason of achieving higher precision. Similar behaviour was noticed for the *SPO*-Apr run.

Table 2 reports the actual number of tweets classified as relevant by the classifier of $f_{CE}(200)$, with an estimation for the number of correct and incorrect decisions based on the 100 samples of each run. The table gives insight about the numbers of classified tweets using f_B vs. $f_{CE}(200)$. As shown, there is a large difference in the amount of relevant tweets detected by each approach. The significant increase in the recall of relevant content comes at the cost of retrieving additional irrelevant content. Nevertheless, the percentage of irrelevant content is seen to be within an acceptable range, even for practical applications. Table 2 shows that the number of relevant content to a given topic varies a lot over time, depending on the events happening at the time. For example, the last test day for the *SPO* topic received relevant posts more than ten times the average. It was noticed that on that day, the 1st of May 2013, was a *Champions League* match between *Barcelona* and *Bayern Munich*, and large amount of posts were talking about the match.

Discussion

Reported results showed the effectiveness of the proposed tweets tracking approach in retrieving additional large number of relevant results compared to Boolean filter, while preserving decent value of precision. Identifying these additional tweets was a real challenge, since these tweets use ambiguous terms and/or describe temporal events that relate to the broad topic and expire quickly.

The key idea in our tracking technique is the methodology used for selecting effective training data for the classifier in an unsupervised manner, and updating it frequently to maintain high performance with dynamic events related to broad topics. We have demonstrated the impact of cleaning the negative training samples by removing tweets containing terms potentially related to the topic. This was clear in the difference in performance between f_C and f_{CE}. We calculated the amount of tweets that were excluded from the negative sample T_{rand} to produce the clean negative sample T_N. It was found that the excluded tweets ($T_{rand} - T_N$) represented only 2.1%, 2.9%, and 3.7% of negative training samples T_{rand} for the $f_{CE}(50)$, $f_{CE}(100)$, and $f_{CE}(200)$ respectively. Surprisingly, this very small

difference in the negative training samples of classifiers of f_{CE} compared to f_C led to a superior change in performance. An explanation to this behaviour is the large confusion created to the classifier because of these samples if not excluded, since large portion of them are potentially relevant tweets. Cleaning them out of the negative sample makes the decision to the classifier unbiased. This illustrates the effectiveness of our unsupervised method for training the classifier for creating an effective and powerful tweets tracking system.

The results in Figure 2, Figure 3, and Table 2 confirm that the performance of our tracking technique is consistently effective with different topics, and over long period of time. In our experiments, the same predefined queries Q_0 of each topic were used over the 4 months without any modification. In real-life, huge changes were occurring to these topics, especially *SYR* and *EGY*. To analyse the way the tracking technique was working over time, we computed the overlap between the top 200 extracted terms in *E* used for cleaning the negative samples for all the topics across the 4 test days.

We computed the average overlap between the 200 terms of the 3 topics over time that was found to be 0.25. This indicates a large drift in the topics, and shows that tracking broad dynamic topics using fixed queries or hashtags would be insufficient, since these topics are highly dynamic. This demonstrates the need to our tracking approach.

Table 3 presents samples of the exclusion terms *E* extracted from the T_B of the training stream, which were used in the cleaning process of the negative samples. The first samples of terms are those that were extracted in all four test dates. As shown, these terms are usually relevant terms to the topics, but are general enough to match irrelevant topics also. For example, the terms "Egypt" and "Syria" for the *EGY* and *SYR* topics respectively are normally relevant. However, they still can be used in posts related to non-political issues, and hence become not relevant to the search topics. Similarly for the term "Madrid" for the *SPO* topic, it can be referring to "Real Madrid" team, and hence become relevant, or referring to the town, and hence become irrelevant. Therefore, our tracking approach excludes tweets containing these terms from the negative training samples to prevent any confusion in the classification model.

The right column of Table 3 presents samples of the terms that appeared for each of the topics once or twice in our test samples over the 4 months period. These are the terms that represent 75% of the top 200 terms in *E*. These terms typically represent events related to the topics for a short period of time. For examples, the three terms of the *EGY* topic were found related to three different clash events happened in Egypt between government and opposition parties during February, March, and April. The example terms for *SYR* topic; "Hizb" and "Allat" refer to "Hizb Allah" the Lebanon armed group that started to fight in Syria starting from April 2013; therefore, these two terms appeared in the test days of April and May. Also the term "Zineb" is a Holy place in Syria that had some battles close to it in April 2013. Similarly for the *SPO* topic; for example, "Camp-Nou" was the name of the stadium that hosted a semi-final match in the champions' league in May 2013. Another interesting feature that was noticed in the *E* terms was the example {**"مورينهو"**} in the *SPO* topic; this is an Arabic translitration for the name "Mourinho" that is different from the one used in Q_0 shown in Table 1. This shows that the *E* terms capture general related terms, temporally relevant terms, and also different spellings of relevant terms.

The previous illustrative samples of the *E* terms demonstrate the high adaptability of our tracking approach, which works in a full unsupervised mode. However, we do not claim that the initial predefined set of queries Q_0 does not require manual update forever. We mentioned earlier that Q_0 is formed of the *most* static part of the dynamic topic. Thus it is expected that these static parts will change eventually. For example in a political topic, the president and governments get changed, which would require updating to the set of queries. However, our main objective was to minimize the manual updates to the query, which we showed in our experiments.

Finally, we can claim that our main objectives for the task listed were achieved. The tweets tracking technique f_{CE} successfully achieved significantly higher recall than Boolean filtering methods while maintaining a high

Table 2. Number of classified tweets using f_B vs. $f_{CE}(200)$.

Topic	Test Date	f_B		$f_{CE}(200)$	
		Rel	Irrel	Rel	Irrel
EGY	Feb	4,386	280	10,449	1,990
	Mar	4,572	141	13,144	1,460
	Apr	4,993	50	10,557	2,978
	May	2,429	155	4,769	2,143
SYR	Feb	2,375	48	4,589	686
	Mar	6,048	252	8,007	1,878
	Apr	2,342	290	3,969	1,468
	May	3,227	206	6,569	1,346
SPO	Feb	1,488	30	3,191	798
	Mar	1,659	87	3,548	943
	Apr	1,329	70	3,980	254
	May	29,929	0	41,901	1,746

Table 3. Sample of Expansion Terms

Topic	Appeared in all test dates	Appeared once/twice
EGY	"الاخوان" (brotherhood) "مصر" (Egypt) "الرئيس" (president)	"الجندي" (Al-Gendi) "المقطم" (Moqatam) "الكاتدرائية" (Cathedral)
SYR	"الأسد" (Al-Asad = "lion") "سوريا" (Syria) "الحر" (free)	"حزب" (Hizb) "اللات" (Allat) "زينب" (Zineb)
SPO	"مدريد" (Madrid) "مباراة" (match) "دوري" (league)	"مورينهو" (Mourinho) "نهائي" (Final) "كامبو-نو" (Camp-Nou)

precision level, and the approach proved to be self-adapting to the changes occurred to the tested topics.

Conclusion and Future Work

In this paper, we studied a novel topic tracking approach for following broad dynamic topics on Twitter. We demonstrated the need of an effective approach for following broad topics on social media with wide applications. We compared our tracking technique with other simpler techniques. We created a test set for the task that contained three different topics of different domains, and we used a stream of Arabic tweets as our data stream. We conducted our experiments over four test days of stream from four different months using a fixed predefined queries set Q_0 for each topic. Our results showed high quality performance for our tweets tracking approach, which managed to increase the recall by more than the double compared to baseline, while achieving a decent precision value of 83%. Our extensive analysis to the results showed high consistency over long periods of time and in different topic domains. We analysed the key idea of our approach, which relies on the ability of collecting clean training data in an unsupervised and adaptive manner to train the classifier frequently. Although our approach is tested on Arabic test data, it is language-independent and general enough to apply on topics and tweets of any other language.

For future work, it would be interesting to test our dynamic topic tracking approach on Twitter when topics are defined by much simpler queries, such as one or two keywords (e.g. "#Egypt" and "#Syria"). It is interesting to know how the performance would be affected when less coverage of the topic is presented in Q_0, and more irrelevant tweets are presented. In addition, applying the algorithm for additional domains in real-time applications create a direct practical evaluation for the approach. Different domains can include: healthcare, crisis management, and customer-care applications.

Acknowledgements

The authors would like to thank the topic developers for their effort for preparing a well-defined sets of queries that cover the static aspects of our topics. Also, a special thanks to Maram Hasanain for her valuable comments on an earlier version of this paper.

References

M. D. Albakour, C. Macdonald, and I. Ounis. (2013). On Sparsity and Drift for Effective Real-time Filtering in Microblogs. In *CIKM 2013*.

A. Ali, W. Magdy, and S.Vogel. (2013). A Tool for Monitoring and Analyzing HealthCare Tweets. In *HSD workshop, SIGIR 2013*.

J. Chen, A. Cypher, C. Drews, and J. Nichols. (2013). CrowdE: Filtering Tweets for Direct Customer Engagements. In *ICWSM 2013*.

M. Conover, J. Ratkiewicz, M. Francisco, and B. Gonçalves. (2011). Political Polarization on Twitter. In *ICWSM 2011*

O. Dan, J. Feng, and B. D. Davison. (2011). A Bootstrapping Approach to Identifying Relevant Tweets for Social TV. In *ICWSM 2011*.

K. Darwish, W. Magdy, A. Mourad (2012). Language Processing for Arabic Microblog Retrieval. In *CIKM 2012.*

Y. Duan, F. Wei, M. Zhou, and H.-Y. Shum. (2012). Graph-based collective classification for tweets. In *CIKM 2012*.

T. Joachims. (2002). Learning to Classify Text Using Support Vector Machines. *Dissertation, Kluwer, 2002*

M. Kendall. (1938). A new measure of rank correlation. *Biometrika*, 30(1/2):81-93, 1938.

J. Lin, R. Snow, and W. Morgan. (2011). Smoothing techniques for adaptive online language models: topic tracking in tweet streams. In *SIGKDD 2011.*

W. Magdy, A. Ali, and K. Darwish. (2012). A summarization tool for time-sensitive social media. In *CIKM 2012*

W. Magdy. (2013). TweetMogaz: a news portal of tweets. In *SIGIR 2013*.

K. Nishida, T. Hoshide, and K. Fujimura. (2012). Improving tweet stream classification by detecting changes in word probability. *SIGIR 2012*

S. Phuvipadawat and T. Murata. (2010). Breaking News Detection and Tracking in Twitter. In *WI-IAT 2010*

I. Soboroff, I. Ounis, and J. Lin. (2012). Overview of the TREC-2012 Microblog Track. In *TREC-2012*

B. Sriram, D. Fuhry, E. Demir, H. Ferhatosmanoglu, and M. Demirbas. (2010). Short text classification in twitter to improve information filtering. In *SIGIR 2010.*

A. Tumasjan, T. O. Sprenger, P. G. Sandner, I. M. Welpe. (2010). Predicting Elections with Twitter: What 140 Characters Reveal about Political Sentiment. In *ICWSM 2010*

S. Vieweg, AL Hughes, K Starbird, L Palen. (2010). Microblogging during two natural hazards events: what twitter may contribute to situational awareness. In *SIGCHI 2010*.

J. Zhang, S. Chen, Y. Liu, J. Yin, Q. Wang, W. Xu, and J. Guo. (2012). PRIS at 2012 Microblog Track. In *TREC 2012*.

Arab Social Media Report (2013). Twitter in the Arab Region. *http://www.arabsocialmediareport.com/Twitter/ LineChart.aspx, March 2013*

.