# Whom Should I Follow? Identifying Relevant Users During Crises

Shamanth Kumar, Fred Morstatter, Reza Zafarani, Huan Liu
Computer Science & Engineering, School of CIDSE, ASU
{shamanth.kumar, fred.morstatter, reza, huan.liu}@asu.edu

## ABSTRACT

Social media is gaining popularity as a medium of communication before, during, and after crises. In several recent disasters, it has become evident that social media sites like Twitter and Facebook are an important source of information, and in cases they have even assisted in relief efforts. We propose a novel approach to identify a subset of active users during a crisis who can be tracked for fast access to information. Using a Twitter dataset that consists of 12.9 million tweets from 5 countries that are part of the "Arab Spring" movement, we show how instant information access can be achieved by user identification along two dimensions: user's location and the user's affinity towards topics of discussion. Through evaluations, we demonstrate that users selected by our approach generate more information and the quality of the information is better than that of users identified using state-of-the-art techniques.

## Categories and Subject Descriptors

H.2.8 [**Database Applications**]: Data Mining; H.3.3 [**Information Search and Retrieval**]: Information Filtering

## General Terms

Experimentation, Human Factors, Measurement

## Keywords

User Identification, Crisis Monitoring, Microblogging, User Relevance Measurement, Twitter

## 1. INTRODUCTION

Natural disasters, riots, and revolutions are inevitable and have made a worldwide impact regardless of where they occur. In March of 2011, a destructive earthquake of magnitude 8.9 struck off the coast of Japan and was followed by a devastating tsunami. The National Police Agency[1] of Japan reports that 15,000 people were killed and more than 128,000 buildings collapsed as a result of the tsunami. Aid agencies from around the world responded to assist in the recovery and provide disaster relief. Hurricane Irene belted the East coast of the United States in August of 2011, causing widespread damage. The property damage in the United States alone was estimated to be around $3 billion and more than 4 million homes experienced loss of electricity[2]. The "Arab Spring" revolutions in the Middle East toppled several regimes in the region. The movement started in Tunisia in late December of 2010, with the self-immolation of Mohammed Bouazizi. The revolution in Tunisia was soon followed by one in Egypt and spread to other countries in the region. A noteworthy record of movements in the Arab Spring countries is being maintained by The Guardian[3]. A common feature of these significant events is that all have impacted the lives of millions locally, as well as globally.

Historically, in covering stories of this magnitude, traditional media such as television and printed news provide a manicured view of the story to their audience backed with vetted, credible resources. While these media often provide a filtered (or edited) view of the story, the overhead incurred in the process results in a slower flow of information. The pervasive use of social media changes the way of communications: the low barrier to publication allows anyone to publish information at any time, making the details of an event instantly available. Instead of providing some edited, exclusive views of an event, social media provides not only timely information in the critical minutes and hours as an event develops, but also many different or inclusive views of the event. Meanwhile, social media generates mountains of data, at times mixed with noise.

With this noisy data in place, *how can we get fast access to relevant and useful information in social media during these events?* An inclusive approach to finding relevant information from inclusive messages is to identify relevant people in social media who are more likely to be the sources publishing useful information (*Information Leaders*) for dynamic events. In general, for a global-scale event, social media users can be naturally categorized into local users who witness the unfolding event and remote users who are connected via social media. Local users have first-hand experience, publishing specifics about the event. To answer this question, we aim to develop an effective way of solving the following problem.

**Problem Statement.** Given a social media site, and

[1] http://tinyurl.com/4lg3ayl

[2] http://tinyurl.com/7b5nags

[3] http://tinyurl.com/68tu9vr


*24th ACM Conference on Hypertext and Social Media*
*1–3 May 2013, Paris, France*

**Table 2: Characteristics of the Arab Spring Dataset**

| | Egypt | Tunisia | Syria | Yemen | Libya |
|---|---|---|---|---|---|
| #users | 514,272 | 19,094 | 146,996 | 43,512 | 375,924 |
| #tweets | 6,184,346 | 86,437 | 2,916,449 | 381,386 | 3,418,485 |
| #geolocated tweets | 84,899 | 5,229 | 16,575 | 849 | 17,814 |
| #retweets | 2,821,864 | 31,392 | 1,253,551 | 142,103 | 1,919,540 |

an event $E$, let $C$ be the content associated with $E$ and $U$ be a set of corresponding users; find "information leaders" $S \subset U$ such that by following $S$, one can effectively obtain information about $E$.

Due to its effectiveness in recent studies [13, 6] and its rapid information dissemination capabilities [17], Twitter is selected as the social media site under study. The content $C$ is therefore represented using tweets (hereafter referred to as $T$) and the event in our case is the Arab Spring revolutions. The information from the few cannot replace the information from all the users posting about the event. However, we aim to develop a method that can quickly access hot or critical information to gain situational awareness and help determine if further information is needed when time permits. To identify these users in an event, we first need to identify specific events for which tweets can be collected to be used for the study. In this paper, we focus on 5 countries swept by the Arab Spring. Before we discuss our approach for geo-topical user identification, we first provide the details about data collection and preprocessing.

## 2. DATA COLLECTION

Below, we discuss our data collection procedure and our data preprocessing steps, which is followed by a discussion on some salient characteristics of our dataset.

### 2.1 Collection Methodology

We systematically collected tweets from various countries within and outside the Middle East, which were related to the Arab Spring. This process involved the usage of certain variables, namely: keywords, hashtags, and geographic regions. We collected 12.9 million tweets which were generated about or from the countries Egypt, Libya, Syria, Tunisia, and Yemen. The tweets were crawled using the system TweetTracker proposed in [9] over the course of 7 months starting from February 1, 2011 to August 31, 2011. A full list of the variables used is presented in Table 1. Column 2 in the table contains the keywords and/or hashtags used. Column 3 contains the geographic boundary box surrounding each country used to crawl all the geolocated tweets from the region. The box is specified as the SW corner (longitude, latitude) of the geographic box followed by the NE corner (longitude, latitude) of the box, separated by a comma. More information on the characteristics of the collected data are presented in Table 2. The data will be shared upon request in accordance with Twitter API Terms of Use.

### 2.2 Data Preprocessing

The Arab Spring movement was not an isolated incident pertaining to a single country. The movement began and subsequently spread across several countries in the Middle East with prominent populations of Arabic, and English speakers. This mixture of language requires special care with respect to processing. As a result, the methods we choose to process the data are language-independent. To preprocess the data we first remove stop words from the dataset (using a comprehensive list of stop words from the English and Arabic languages). In addition to traditional stop words, we also removed Twitter artifacts from the text such as hashtags, user mentions, and URLs. Next we attempted to stem the words. However, this became problematic as we soon discovered that existing stemmers for the Arabic language are not yet fit for real world problems. In our efforts, we tested three stemmers: the Arabic stemmer created by [10], the Arabic stemmer provided with Apache Lucene[4], and the Tashaphyne stemmer[5]. All three of the aforementioned stemmers produced inconsistent output that could not be understood by native Arabic speakers helping our team, making it impossible for the authors to know if their results were correct. Therefore, to remain consistent, we eliminated stemming from our preprocessing treatment for all languages.

**Table 3: Sample of words from a subset of topics in Tunisia with justification for their selection.**

| Topic Keywords | Selected | Reason |
|---|---|---|
| `forget, tonight, ..., proud, site` | No | Disagreeing |
| `police, protest, ..., situation, shot` | Yes | Agreeing |

Next we discuss our approach to identifying information leaders, or users to follow in an event.

## 3. GEO-TOPICAL USER IDENTIFICATION

Social media sites now have millions of users and information travels easily and quickly through this medium. Due to noise and credibility concerns, it is not sufficient to simply pick users who produce more information. Tracking all users is also not a viable option to acquire information. To identify a subset of the users who are likely to publish useful information on a crisis we need to come up with a more effective strategy. Two factors play an important role in a crisis: 1) the *topic of discussion* which relates the user to the event, and 2) the *location of the users* which is important to establish the credibility of the content being published by the user. Every user who has tweeted on a topic can be associated with each of these dimensions with a specific score that represents his relevance along that particular dimension. Below, we discuss the procedure to compute these scores and also explain the significance of scoring well along a particular dimension. Our first step is to identify the topics of discussion in the tweets.

### 3.1 Topic of Discussion

Tweets can be considered as small documents of length at most 140 characters. The topic of discussion of the tweets can be manually labeled as one of several topics of discussion or factors that initiate these discussions. In the context of Arab Spring, these factors may include economic factors, torture and brutality, protest, etc. Alternatively, an automated approach of topic detection in documents is the Latent Dirichlet Allocation (LDA) [2].

In this work, we use LDA to evince topics in the various events in the Arab Spring. We utilize the Gibbs sampler LDA[6] to discover topics from the tweets. To tune the hyperparameters on the Dirichlet priors ($\alpha$, $\beta$) and the number

[4]http://lucene.apache.org/
[5]http://pypi.python.org/pypi/Tashaphyne/
[6]http://tinyurl.com/783o3nw

**Table 1: Parameters Used to Collect the Tweets**

| Country | Keywords/Hashtags | Geographic Boundary |
|---|---|---|
| Egypt | `#egypt,#muslimbrotherhood,#tahrir,#mubarak,#cairo,#jan25,#july8,#scaf,#noscaf` | (22.1,24.8),(31.2,34.0) |
| Tunisia | `#tunisia,#tunisian,#tunez` | (30.9, 9.1),(37.0,11.3) |
| Syria | `#syria,#assad,#aleppovolcano,#alawite,#homs` | (32.8,35.9),(37.3,42.3) |
| Libya | `#libya,#gaddafi,#benghazi,#brega,#misrata,#nalut,#nafusa,#rhaibat` | (23.4,10.0),(33.0,25.0) |
| Yemen | `#yemen,#sanaa,#lbb,#taiz,#aden,#saleh,#hodeidah,#abyan,#zanjibar,#arhab` | (12.9,42.9),(19.0,52.2) |

of topics $N$, we performed several iterations of LDA using the Tunisia dataset and did manual inspection to see which parameter values perform the best. To start we began from $\alpha = 0.1$ to 1.0 in intervals of 0.1, and $N = 10$ to 100 in intervals of 10 for a total of 100 iterations. We then manually went through these results and chose the parameters that made the most sense to us as topics. As criteria, we looked to the coherency of the words in a topic to make up what we viewed as a theme, regardless of the content. Once we obtained a value of $\alpha$ and $N$, we next proceeded to tune $\beta$. To do this, we iterated $\beta = 0.1$ to 1.0 in intervals of 0.1 with $N$ fixed at 40. After analyzing the results, we found that the best results resided between 0.1 and 0.2. Iterating between 0.1 and 0.2 at an interval of 0.01, we found that the best value for $\beta$ was 0.11. However, we found that some topics produced were not coherent. In the next section, we discuss how we trimmed the irrelevant topics, to ensure that all topics investigated present a coherent idea.

#### 3.1.1 Topic Pruning

Upon inspection of the topics produced by LDA, we soon realized that many topics were unfit to inspect further, i.e., many contained unrelated keywords, or sets of keywords that did not add up to a distinct topic. To remove the unrelated topics, the authors, along with native Arabic speakers, manually went through the topics and eliminated those that were not related to the event of that country. In Table 3, we show an example of an English topic for the events that were deemed appropriate for our studies and ones that were not. After careful pruning, we were left with the following number of topics for each country: Egypt - 11, Libya - 23, Syria - 17, Tunisia - 14, Yemen - 21.

Using the final set of topics from each country, we can identify user relevancy through a topic affinity score.

#### 3.1.2 Topic Affinity Score

Let $S$ be the set of words that define the topic. These words are the top 25 most probable words for the topic, as determined by LDA, i.e., $|S| = 25$. Let $\mathcal{T}$ be the collection of a user's tweets. Let $T \in \mathcal{T}$ be a user's tweet, i.e., a set of words. We can define a user's topic affinity score as in Equation 1.

$$topic\_score(S, \mathcal{T}) = \frac{\sum_{T \in \mathcal{T}} sgn(|S \cap T|)}{|\mathcal{T}|}, \qquad (1)$$

where, $sgn$ represents the sign function. Using this formulation we see that a user's topic affinity score is in the interval $[0, 1]$. Score value 0 indicates that they never tweeted in the topic and a score of 1 indicates that all of the tweets overlapped with the topic.

### 3.2 Location of the User

During a crisis, the location of the user is an important factor which can help us determine which user is likely to publish information relevant to the crisis. For example, in an earthquake, tweets coming from a location closer to the earthquake are likely to be more pertinent to the crisis than tweets from outside the location. In the case of the Arab Spring, tweets coming from within the country are more likely to contain relevant information than those from outside the respective countries. To identify a user's relevancy to the event based on his location, we propose the measure *geo-relevancy score*.

#### 3.2.1 Geo-Relevancy Score

A user's location can be determined using the location from his tweets. The location of a tweet can be determined in one of two ways:

1. **Geolocated Tweet** - A tweet that has been located through the GPS sensor on a mobile device, or through IP location capabilities of the browser. This information is metadata that the individual tweeting chooses to share when publishing the tweet.

2. **Profile-located Tweet** - A tweet whose location data is obtained by analysis of the user's profile. Users can provide geographic location information in their profile, and we analyze this by geolocating it through the OpenStreetMaps Service[7].

Using the location information from the user's tweets his geo-relevancy score is a value in the interval $[0, 1]$, calculated as follows:

1. If the user never produced a geolocated tweet, then his geo-relevancy score is the average number of his tweets that were profile-located to be within the crisis region. A user is represented as a tweet location vector $tweet\_loc \in \mathbb{R}^T$, where $T$ is the number of tweets published by the user. $tweet\_loc_i = 1$, indicates that the user's profile information at the time of the $i$th given tweet resolves to within the crisis region and a $tweet\_loc_i = 0$ indicates that the user was outside or that the location information was missing. Then, we can compute the geo-relevancy score as follows:

$$geo\_rel\_score(tweet\_loc) = \frac{||tweet\_loc||_0}{T}, \qquad (2)$$

   where $||.||_0$ denotes the zero-norm.

2. If a user is geolocated and their location is within the crisis region, then his geo-relevancy score is 1.

3. Conversely, if a user produces a geolocated tweet that is not within the crisis region, then their geo-relevancy

[7] http://nominatim.openstreetmap.org/

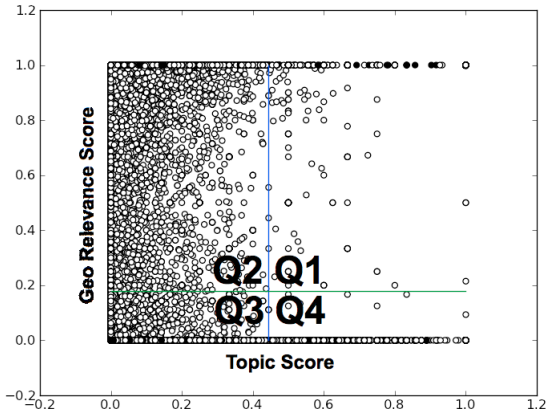



**Figure 1: User visualization with both geo-relevancy and topic affinity for a topic in Egypt.**

score is set to 0 as they have demonstrated that they are not within the location and do not have access to the temporally-sensitive information as someone experiencing the event firsthand.

We note that the user has a different topic affinity score for each topic in the revolution, but the same geo-relevancy score across the topics.

## 3.3 Visualizing Users in Two Dimensions

After obtaining the geo-relevancy score and topic score for each user in every topic, we create a scatter plot to see how users are related to each other. An example of one such plot is shown in Figure 1. In this plot, each dot is a user. The black dots are the users who received their score through geolocation (rules 2 and 3 of the previous section). The white dots are users who received their geo-relevancy score from resolving their profile information (rule 1 in the previous Section). The $x$-axis represents the user's topic affinity, and the $y$-axis represents the user's geo-relevancy score. The vertical and horizontal bars represent the averages for the distance and topic scores, respectively. In Figure 1 we can see that, based on the location of these average bars, the plot breaks down into four quadrants.

### 3.3.1 Understanding the Quadrants

By laying out the quadrants in the method prescribed above, we observe that each quadrant has certain unique characteristics. Using the same numbering system as the Cartesian coordinate system, we define the following quadrants:

**Quadrant I** (Q1): This quadrant contains users with both topic and geo-relevancy scores above the average. This quadrant contains users who are both on the ground and actively discussing the topic at hand. We call these users "*Eyewitness*" users.

**Quadrant II** (Q2): This quadrant contains users whose topic score is below average, but their location score is very high. These people are in the vicinity of the revolution, but not discussing the topic. We call these users "*Topic Ignorant.*"

**Quadrant III** (Q3): This quadrant contains users with topic and geo-relevancy scores below the average. We call these users "*Apathetic*", as they are neither within the region nor discussing the topic at hand.

**Quadrant IV** (Q4): This quadrant contains users with topic scores above the average, but geo-relevancy scores below. These users are outside of the country, but are still producing information relevant to the event. We call these users "*Sympathizers*".

Users in Q1 can be considered the most relevant to the crisis, as they have high scores across both the dimensions. Users in both Q1 and Q4 are considered "*topic-aware*" as they have a better-than-average discussion rate on the given topic. These are users who have spent a lot of time talking about topics relevant to the Arab Spring. Hence, we propose to study the tweet characteristics of the users in Q1 and Q4. This study would clarify the utility of following Q1 for the purpose of obtaining information about an event.

# 4. UNDERSTANDING USERS IN Q1 AND Q4: SPECIALISTS AND GENERALISTS

Having identified a measure to discover users who are involved in the topic (users who appear in either Q1 or Q4), we are left to uncover the relationship of these users with similar topic-aware users across other topics in a region. Do differences exist between topic-aware users who experience the event first-hand (Q1 users) and those who do not (Q4 users)? In this section we discuss the interrelatedness of the quadrants across topics. The intuition behind conducting this experiment comes from the fact that Q1 users in each topic of a country directly experience the crisis.

## 4.1 Similarities in Specialists and Generalists

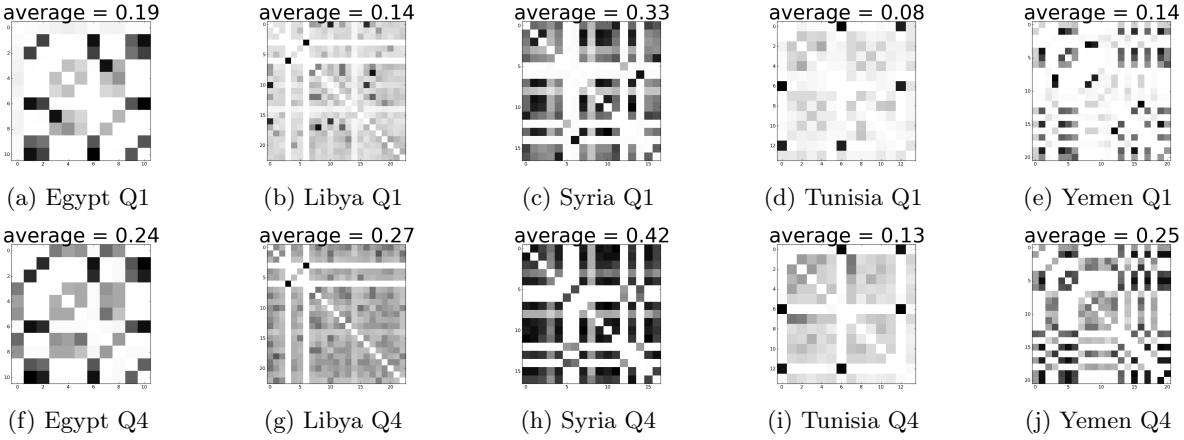
First, we investigate the overlap of information in Q1 and Q4 users across topics. Here the information will be represented by the top 35 most frequently used keywords in their tweets. To measure the overlap we employ the Jaccard similarity measure. Jaccard score has the benefit of ignoring the position of the words, giving the advantage of not paying heed to the frequency of a word, but instead just focuses on the information covered by the users in the quadrant. Two topics which yield a low Jaccard score will likely cover different sets of information. Conversely, two topics which have a high Jaccard score will cover similar information.

Before computing the Jaccard scores, we first eliminate all comparisons between topics which are not in the same language. To determine whether or not the languages are similar enough for comparison, we used the following measure:

$$lang\_sim(W_i, W_j) = 1 - \frac{|arabic(W_i) - arabic(W_j)|}{(|W_i| + |W_j|)/2}, \quad (3)$$

where $W_i$ is the set of keywords in topic $i$ and $arabic$ is a function that returns the number of Arabic words in the set. This measure ensures that we do not make unreasonable comparisons between topics which are in different languages. We control the comparisons through a threshold, $\epsilon$, which represents the language similarity we require between two topics being compared. This parameter is set $\epsilon = 0.80$, meaning that at least 80% of the topic words must agree in language for the comparison to occur.

The heat maps in Figure 2 show the inter-topic Jaccard scores along with the average Jaccard values for each country.




(a) Egypt Q1 (b) Libya Q1 (c) Syria Q1 (d) Tunisia Q1 (e) Yemen Q1

(f) Egypt Q4 (g) Libya Q4 (h) Syria Q4 (i) Tunisia Q4 (j) Yemen Q4

**Figure 2: Heat maps showing inter-topic Jaccard similarity scores for Q1 (crisis eyewitness) users and Q4 (crises sympathizers around the world) users in different countries. White represents a Jaccard similarity score of 0 and black represents a Jaccard score of 1. It is clear that Q4's (sympathizers) are much darker (similar in discussion) than Q1 (eyewitness)'s.**

Darker tiles indicate a Jaccard score closer to 1.0 and lighter tiles indicates a Jaccard score closer to 0.0. Results show that, for all countries, the Jaccard scores across the topics for users in Q1 are lower than the Jaccard scores across the topics for users in Q4.

This shows that the discussion of Q1 users across topics will be centered on specific issues that they perceive as relevant. Here, location is an important influencing factor. On the other hand users in Q4 are located outside the region of crisis and do not experience the crisis first hand. Therefore, their discussion is expected to be focused on a wide range of topics. Indeed, this pattern can be seen across the Q1s and Q4s for each country. This tells us that users who are actually in the affected region, are tweeting about different topics. In this sense, we can term Q1 users as "*specialists*". On the other hand, we see a large amount of overlap between the "sympathizers" in Q4. The users in Q4, though in different topics, discuss the same top words in their tweets and, by extension, are largely talking about the same things. For this reason, we term Q4 users as "*generalists*". This behavior is very different from the users in Q1 who are in the region, and are largely discussing very different things. In the next section, we will explore the overlap in the topics of generalist (Q4) users further and how they prioritize the information being discussed.

## 4.2 The Disaccord of Generalists

From the previous section we know that "generalists" have considerable overlap in their discussion. In this section, we delve into studying the users in Q4, showing that while they share many of the top words with users in other topics, they exhibit originality in the ranking of their keywords based on frequency. That is, the ideas they discuss are similar, however the importance they attribute to individual ideas differs from topic to topic. Previously, we employed only the Jaccard score to compare the similarity of topics, a measure that works well when trying to see the disharmony of the top keywords in the topics. Comparing the content produced by Q4 users across topics gives us more insight into the individuality of topics.

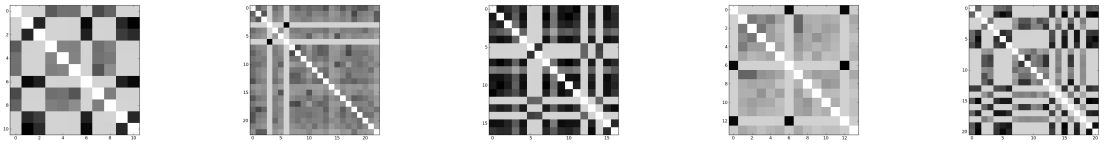
To compare the ranking of the keywords from Q4 users we will use the Kendall's $\tau$ rank correlation coefficient between two lists consisting of the top 35 words from Q4 users belonging to two different topics, in descending order of the number of occurrences. Kendall's $\tau$ measures rank correlation of two lists by counting the number of agreeing and disagreeing pairs in the two lists. Using the same value of $\epsilon$ as mentioned previously, we generate the $\tau$ scores for Q4 across topics for each country. These scores are presented as a heat map in Figure 3. The heat map represents the Kendall's $\tau$ score across topics, with a darker square indicating a higher score (that is, a score closer to 1). Lighter squares indicate that the top keywords contain much overlap, but the ordering of the words are different. Darker squares indicate that there is much overlap, and the ordering is similar. Figure 3, confirms our previous observation that the ranking of the words for the users in Q4 across topics is quite similar. Only in the case of Tunisia this phenomena is less pronounced. We suspect that this might be due to the size of the dataset for Tunisia, which is significantly smaller than those for other countries.

# 5. EVALUATION

In this section, we will show that users in Q1 generate higher quantity of information. Later, we will evaluate the quality of the information generated by these users.

## 5.1 Information Quantity from Q1 Users

To evaluate the quantity of information generated by Q1 users, one can measure the quantity of tweets published by Q1 users from each country. To show that these users produce more information and the quantity is statistically significant, we need to compare quantities produced with a set of representative users from within our dataset. Uniform sampling provides theoretical guarantees on generating accurate representative datasets. Hence, we uniformly sample an equal number of users from the dataset as contained in Q1 and consider it as a representative set. To avoid any sampling bias in the results of comparison, we generate 100 such sets of randomly selected users $U_{Rand}$ and take the average of the number of tweets generated by them to the number of tweets generated by Q1. A comparison of the



(a) Egypt (b) Libya (c) Syria (d) Tunisia (e) Yemen

**Figure 3: Pairwise rank correlation, computed using Kendall's $\tau$ coefficient, for generalist (Q4) users of each country**

**Table 4: Comparison of the quantity of tweets generated by Q1 users and a random set of users $U_{Rand}$**

| | Tunisia | | | Egypt | | | Syria | | | Yemen | | | Libya | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Q1 | $U_{Rand}$ | $p-value$ | Q1 | $U_{Rand}$ | $p-value$ | Q1 | $U_{Rand}$ | $p-value$ | Q1 | $U_{Rand}$ | $p-value$ | Q1 | $U_{Rand}$ | $p-value$ |
| Feb | 1,817 | 3,706 | | 74,956 | 138,379 | | 228 | 20 | | 1,247 | 211 | | 1,101 | 926 | |
| Mar | 805 | 1,521 | | 84,856 | 34,346 | | 199 | 113 | | 2,546 | 345 | | 1,971 | 668 | |
| Apr | 1,006 | 2,062 | | 137,562 | 48,472 | | 12,271 | 4,817 | | 4,480 | 599 | | 2,840 | 319 | |
| May | 144 | 234 | <0.0001 | 22,335 | 7,939 | <0.0001 | 47,496 | 2,347 | <0.0001 | 639 | 40 | <0.0001 | 2,419 | 168 | <0.0001 |
| Jun | 12 | 5 | | 29,569 | 11,610 | | 40,458 | 2,514 | | 1,568 | 161 | | 2,068 | 187 | |
| Jul | 296 | 364 | | 274,446 | 89,348 | | 113,069 | 5,550 | | 4,666 | 444 | | 4,488 | 111 | |
| Aug | 2,081 | 1,716 | | 232,288 | 67,608 | | 79,428 | 10,018 | | 3,920 | 224 | | 4,624 | 961 | |

tweets generated by Q1 and $U_{Rand}$ is presented in Table 4. Looking at the first two columns for each country, it is clear that the Q1 users generally tweet more than $U_{Rand}$. In cases such as Syria, we found that Q1 users comprised around 0.006% of the users and yet contributed more than 10% of all the tweets for the region. To show that the observed difference is also statistically significant we employ the $\chi^2$ test. Our null hypothesis is as follows:

> $H_0$: Q1 users and randomly selected users $U_{Rand}$ generate similar numbers of tweets during a crisis.

Given 7 months of data, we run the $\chi^2$ test with 6 degrees of freedom and a significance level of $\alpha = 0.05$. As observed from Table 4, we reject the null hypothesis for all the countries. The results of the $\chi^2$ test show that the difference between the rate at which tweets are generated by Q1 and $U_{Rand}$ is statistically significant.

## 5.2 Information Quality of Q1

The previous section establishes that Q1 users generate a significant amount of information. Here, we show that this information is of high quality.

### 5.2.1 Meaningful Patterns

In this section, we show that Q1 users generate information that captures the current trends in the region. Consequent to our methodology, these users are well placed to generate firsthand accounts, as they are in the crisis region and have access to information that most others outside the region do not. By meaningful information here, we mean information that does not correlate highly with the information an average *random* user concerned about the event would publish. Our assumption while performing this experiment is that information leaders should (1) post more often about the specific events *when these events exist* and at other times, (2) post information that is closer to the general discussion about the crisis. In this experiment, we compare the content from the tweets of Q1 users with the (i) general topic of discussion and to (ii) those of a randomly selected set of users. This experiment is performed over $M$ days spanned by our dataset. Here $M = 212$. The topic of discussion for any set of users $U$ is defined as a collection of the top 35 most popular keywords occurring in the tweets of $U$. Here Q1 users are the union of all Q1 users that exists across topics for a country, i.e., $Q1 = \cup_{j=1}^{n} Q1_j$, where $Q1_j$ is Q1 users for topic $j$ of a specific country and $n$ is the number of topics in that country.

**Table 5: Q1's ability to diverge from the general topic.**

| | Egypt | Libya | Syria | Tunisia | Yemen |
|---|---|---|---|---|---|
| $\|Q1\|$ | 19565 | 337 | 946 | 654 | 202 |
| Position | 2 | 1100 | 3307 | 1 | 3751 |

- Let $U_{Rand}$ represent a random set of users selected from the dataset. $T_{Rand}$ represents the topic of $U_{Rand}$, $T_{Q1}$ represents the topic of Q1 users and $T_{general}$ represents the general topic of discussion among all users.

- For day $i$, where $1 \leq i \leq M$, we compute the distance $d_i$ between $T_{Q1}$ and $T_{general}$ using Jaccard distance,

$$d_i = \frac{T^i_{Q1} \cap T^i_{general}}{T^i_{Q1} \cup T^i_{general}}. \quad (4)$$

  Then, we can represent daily distances using a vector, $D = (d_1, d_2, \ldots, d_M) \in \mathbb{R}^M$. We can generalize the distance to vector format using any vector norm. Here, we use the $L_1$-norm,

$$d(T_{Q1}, T_{general}) = ||D||_1 = \sum_{i=1}^{M} d_i. \quad (5)$$

  $d(T_{Rand}, T_{general})$ can be calculated similarly. To remove random bias, we generated 5,000 random user sets $\{U^i_{Rand}\}_{i=1}^{5,000}$ 's and their corresponding topics $\{T^i_{Rand}\}_{i=1}^{5,000}$ 's.

  We computed the distance between $T_{general}$ and all 5,000 randomly generated topics $\{T^i_{Rand}\}_{i=1}^{5,000}$, i.e., $d(T_{general}, T^i_{Rand})$. We also computed the general topic $T_{general}$ distance to $T_{Q1}$, i.e., $d(T_{Q1}, T_{general})$. After these computations, we are left with 5,001 distances to the general topic (5,000 distances from random topics + 1 from Q1 users). The list of 5,001 distances is then

**Table 6: Evaluation of Tweet Quantity by Q1 and Followers**

| | Tunisia | | | Egypt | | | Syria | | | Yemen | | | Libya | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Follow | Q1 | $p-value$ | Follow | Q1 | $p-value$ | Follow | Q1 | $p-value$ | Follow | Q1 | $p-value$ | Follow | Q1 | $p-value$ |
| Feb | 1,117 | 1,817 | | 204,023 | 74,956 | | 48 | 228 | | 237 | 1,247 | | 1,720 | 1,101 | |
| Mar | 570 | 805 | | 41,275 | 84,856 | | 135 | 199 | | 364 | 2,546 | | 1,670 | 1,971 | |
| Apr | 664 | 1,006 | | 46,187 | 137,562 | | 4,967 | 12,271 | | 497 | 4,480 | | 616 | 2,840 | |
| May | 105 | 144 | <0.0001 | 9,310 | 22,335 | <0.0001 | 5,793 | 47,496 | <0.0001 | 148 | 639 | <0.0001 | 270 | 2,419 | <0.0001 |
| Jun | 5 | 12 | | 14,928 | 29,569 | | 5,265 | 40,458 | | 135 | 1,568 | | 170 | 2,068 | |
| Jul | 152 | 296 | | 122,505 | 274,446 | | 12,036 | 113,069 | | 461 | 4,666 | | 336 | 4,488 | |
| Aug | 855 | 2,081 | | 67,525 | 232,288 | | 12,316 | 79,428 | | 240 | 3,920 | | 1,603 | 4,624 | |

sorted in ascending order. The first element in this list is the farthest away from the general topic of discussion, and the 5001st is the closest. The ranking of Q1 users is presented in Table 5. Q1 users deviated from the general topic of discussion in Egypt and Tunisia. In Syria and Yemen users in Q1 were closer to the general topic of discussion.

## 5.3 Unique Attributes of Q1 Users

It is important to distinguish the users found by the method from influential users found using other methods. These influential people are expected to generate crisis-relevant information. In this experiment, we show that Q1 users generate more information than influentials, later we will show that this information is also more focused compared to the influentials. To measure influence in a directed network such as Twitter one way is to consider the number of followers a user has accrued. We use this approach in this paper, although techniques, such as PageRank could be employed.

To conduct this experiment, we first identify the number of tweets generated by users from Q1 and the Influentials from each country in each month $m$, spanned by our dataset. Our results are presented in Table 6. We observe that amount of information generated by Q1 is much higher than Influentials. To see if the difference is statistically significant, we run the $\chi^2$ test on the results and we find that the difference between the quantity of tweets generated by the two types of users is statistically significant in all cases.

To further investigate the uniqueness of Q1 users, we compare the underlying word frequency probability distribution of their most-used words with that of the Influentials. To compare these distributions, we could use the Jensen-Shannon (JS) divergence [11],

$$JS(P||Q) = \frac{1}{2}[D(P||M) + D(Q||M)], \quad (6)$$

where $M = \frac{1}{2}(P + Q)$, and $D$ is the Kullback-Leibler (KL) divergence [4],

$$D(P||Q) = \sum_{i=1}^{|P|} P_i \cdot log(\frac{P_i}{Q_i}). \quad (7)$$

Here, $P$ and $Q$ are the normalized occurrences of the top 500 words used by each of the 10 groups ((Influentials + Q1s) $\times$ 5 countries). Using the JS divergence on can create a distance matrix between the 10 groups. From the distance matrix, we can generate an embedding of the groups based on embedding techniques. The embedding will demonstrate how different groups are situated with respect to one another in a 2-dimensional space. When seeking an embedding of the matrix, it is more desirable to have a distance metric since distances will be comparable (due to triangle inequality). It has been proven that the square root of the JS divergence is a metric [5]; therefore, we use that instead of the JS divergence.

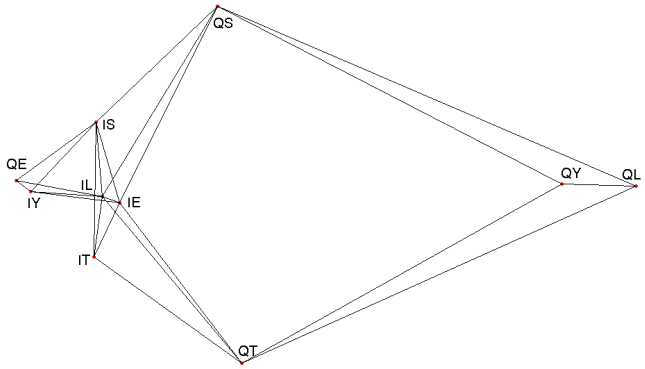



**Figure 4: Group embedding (using Isomap) of influentials and Q1 users. The probability distribution is the frequency of top-words and the distance is computed using the square root of JS divergence. In this figure, each point is labeled by a two-character code. The first character is either 'I' or 'Q', indicating a Influentials or Q1 group, respectively. The second character is the first letter of the group's representative country. For example, QE represents the Q1 group in Egypt, and IY represents the Influentials group in Yemen.**

For this work, we investigated classical embedding techniques such as the classical PCA or Multi-Dimensional Scaling (MDS), and decided to employ the more robust Isomap technique [18], capable of extracting non-linear relationships using geodesic distances between points. The resulting 2-dimensional embedding can be seen in Figure 4. This figure shows the different Q1s at a distance from each other surrounding a dense group of Influentials.

## 5.4 Visualizing the Topics

At this point, visualizing the topics of the above mentioned Q1 users and influential users would aid our understanding of the differences between them. A commonly-used technique to visualize such textual information is tag clouds. To generate the word clouds, we first extracted the top 35 words for all of the Q1 users for each topic based on frequency. Using the same method we extracted the top 35 keywords for the influential users in that topic (with the most followers), limiting this group to size $|Q1|$ for fair comparison. This process generates two lists of the 35 words for each topic (one for Q1s, and another for Followers). Next, we take the content of these lists, and use Wordle[8] to generate word clouds. Representative examples of these word clouds from a topic from Libya and another from Syria are presented in Figure 5.

[8]http://www.wordle.net/


(a1) Followers (a2) Q1s

(a) Sample Topic from Libya


(b1) Followers (b2) Q1s

(b) Sample Topic From Syria

**Figure 5: Two representative topics from Libya and Syria selected randomly from the pool of discovered topics from the 5 countries.**

From the tag clouds in Figure 5, one can notice that the influential users in Figure 5(a1) are discussing information from both in and out of the region. This is expected and is the basis of the motivation for this work, as the identified influential members during an event may often be located outside the region and thus, have limited access to information from the event location. On the other hand, in the case of Q1 Libya users in Figure 5(a2), selected using our approach, we observe that the discussion is clearly focused on topics like killing, rebels, and reports of violence, and major cities of Libya are all common points of discussion, which shows that the discussion among these users is focused on the problems in the region.

## 6. RELATED WORK

The related work to our research falls into three intertwined areas: topic models, event detection, and finally Twitter analysis under events, especially disasters.

Topic models have been studied extensively in short-messaging environments. [8] analyzed tweets related to crises using topic models. Their approach employs topical clustering and their new technique, dynamic corpus refinement. They tune the term weights in order to get more accurate topic distributions and they also refine their corpus based on the initial topic distribution in order to get datasets that are more related to the disaster under study. [16] presents a partially supervised learning model, called the Labeled LDA, that maps tweets into dimensions. They argue that these dimensions correspond to substance (topics about events, ideas, things, and people), social characteristics (social topics), style (broader trends), and status (personal tweets). Their models take into account both users and tweets. Besides their latent dimensions in Twitter that can help identify broad trends, in order to identify smaller ones, several classes of tweet- specific labels were applied to tweet subsets.In another effort [7], the authors attempt to adapt the standard topic model system to the microblogging environments. Their results show that models trained on aggregated messages result in higher performance in real-world scenarios. It is interesting to know how topics found by these models change from microblogs compared to the ones found in traditional media. This has been done in [19] where they compare Twitter based topic models and the ones found in traditional media. They found that Twitter acts in many ways similar to social media. However, there are differences. For instance, Twitter acts as an invaluable source for "entity-oriented" topics. These are topics that have low-coverage in other sources of media. Another finding was that though Twitter users had low interest in international news, they actively engaged in helping spread important news.

Topic and event detection has also been an active area of research. In [3], the authors introduce both a topic detection and a real-time topic discovery technique. The topics are described as a set of terms. Terms have a life cycle and a term or set of terms is considered emergent if its frequency increases in a specified time interval and was relatively rare in the past. They also weight content based on the PageRank of the

authors and introduce a topic graph where users can identify semantically related emergent topics and keywords. [14] introduces a method for identifying controversial events. They formalize controversial events and approach the problem using regression methods. Controversial events are ones that provoke public discussions in which audience express opposing opinions. Their feature set includes Twitter-based (linguistic, structural, buzziness, sentiment, and controversy) and External features (News Buzz and Web News Controversy). Various systems have also been developed to monitor tweets and events on Twitter [9]. TwitterMonitor [12] is a system that performs trend detection over a stream of tweets. The system detects emerging topics or trends, and provides meaningful analytics. Emergent topics are detected by identifying bursty keywords.

Twitter, and in general microblogging, has shown to be highly effective when it comes to disaster relief and rapid communication during a natural disaster. Recent studies related to the disasters in Yushu [15], Japan [17], and Chile [13], and Haiti [6, 1] show its usefulness in these situations.

# 7. CONCLUSION AND FUTURE WORK

Identifying information quickly and efficiently is crucial during crises. In this paper, we presented an innovative approach to efficiently access information in social media. Using Twitter as an example, we show that we can identify a subset of Twitter users who publish tweets about the event of interest and can help provide quick access to relevant information. In other words, we present an approach to detect "Information Leaders".

Our approach is based on two natural dimensions along which a user can be categorized, namely: topic of discussion and the user's location. Specifically, our contributions are:

- We present a novel approach to find users who provide quick access to relevant event information.
- Based on our approach we identify different categories of users who can provide different kinds of information. *Generalists* can be used to understand the global impact of a crisis. *Specialists* can be used to get access to information on various topics directly associated to a crisis from within the impact region.
- Our method gives all users equal opportunity to be information leaders. In the event of a crisis, most useful information usually comes from people who have personally experienced the impact or have access to such information. These users are not expected to have a large number of followers or play a central role in the Twitter network outside of the crisis.

Through comparison with a reasonable measure of identifying information leaders, we show that users selected using our approach produce information in more quantity and with better quality. Further work is needed to improve our approach, such as considering the network information of the users to identify influential individuals.

## Acknowledgments

This work was supported, in part, by the Office of Naval Research grant: N000141010091. We are also grateful to Salem Alelyani for insightful discussions.

# 8. REFERENCES

[1] G. Barbier, R. Zafarani, H. Gao, G. Fung, and H. Liu. Maximizing benefits from crowdsourced data. *Computational & Mathematical Organization Theory*, pages 1–23, 2012.

[2] D. Blei, A. Ng, and M. Jordan. Latent dirichlet allocation. *JMLR*, 3:993–1022, 2003.

[3] M. Cataldi, L. Di Caro, and C. Schifanella. Emerging topic detection on twitter based on temporal and social terms evaluation. In *Proceedings of the Tenth Int. Workshop on Multimedia Data Mining*, page 4, 2010.

[4] T. Cover, J. Thomas, et al. *Elements of information theory*. Wiley Online Library, 1991.

[5] D. Endres and J. Schindelin. A new metric for probability distributions. *Information Theory, IEEE Transactions on*, 49(7):1858–1860, 2003.

[6] H. Gao, G. Barbier, and R. Goolsby. Harnessing the crowdsourcing power of social media for disaster relief. *Intelligent Systems, IEEE*, 26(3):10–14, 2011.

[7] L. Hong and B. Davison. Empirical study of topic modeling in twitter. In *Proceedings of the First Workshop on Social Media Analytics*, pages 80–88. ACM, 2010.

[8] K. Kireyev, L. Palen, and K. Anderson. Applications of topics models to analysis of disaster-related twitter data. In *NIPS Workshop on Applications for Topic Models: Text and Beyond*, 2009.

[9] S. Kumar, G. Barbier, M. A. Abbasi, and H. Liu. TweetTracker: An Analysis Tool for Humanitarian and Disaster Relief. In *ICWSM*. The AAAI Press, 2011.

[10] L. S. Larkey and M. E. Connell. Arabic information retrieval at umass in trec-10. Technical Report ADA456273, University of Massachussetts, 2006.

[11] J. Lin. Divergence measures based on the shannon entropy. *Information Theory, IEEE Transactions on*, 37(1):145–151, 1991.

[12] M. Mathioudakis and N. Koudas. Twittermonitor: trend detection over the twitter stream. In *SIGMOD*, pages 1155–1158. ACM, 2010.

[13] M. Mendoza, B. Poblete, and C. Castillo. Twitter under crisis: Can we trust what we rt? In *Proceedings of the First Workshop on Social Media Analytics*, 2010.

[14] A. Popescu and M. Pennacchiotti. Detecting controversial events from twitter. In *CIKM*, pages 1873–1876, 2010.

[15] Y. Qu, C. Huang, P. Zhang, and J. Zhang. Microblogging after a major disaster in china: a case study of the 2010 yushu earthquake. In *CSCW*, pages 25–34, 2011.

[16] D. Ramage, S. Dumais, and D. Liebling. Characterizing microblogs with topic models. In *ICWSM*, 2010.

[17] T. Sakaki, M. Okazaki, and Y. Matsuo. Earthquake shakes twitter users: real-time event detection by social sensors. In *Proceedings of the 19th international conference on World wide web*, pages 851–860, 2010.

[18] J. Tenenbaum, V. De Silva, and J. Langford. A global geometric framework for nonlinear dimensionality reduction. *Science*, 290(5500):2319–2323.

[19] W. Zhao, J. Jiang, J. Weng, J. He, E. Lim, H. Yan, and X. Li. Comparing twitter and traditional media using topic models. *Advances in IR*, pages 338–349, 2011.