Social Network Analysis and Mining (2019) 9:61
https://doi.org/10.1007/s13278-019-0606-6

ORIGINAL ARTICLE

Activeness of Syrian refugee crisis: an analysis of tweets

Karsten Ladner[1] · Ruchishya Ramineni[2] · K. M. George[2]

Received: 12 November 2018 / Revised: 2 October 2019 / Accepted: 4 October 2019 / Published online: 17 October 2019

Abstract
In this paper, we propose and apply a method to analyze the activeness of an event based on related tweets. The method characterizes and measures activeness of an event by a set of indicators. The indicators proposed in this paper are original tweet count, retweet count, follower count, positive sentiment, negative sentiment, daily change in users count, and sparseness of user community. We present procedures to compute the last two indicators. All indicators collectively are used to determine the activeness of an event. This approach is used to analyze the Syrian-refugee-crisis-related tweets. Its generality is demonstrated by applying it to analyze "immigration"-related tweets.

Keywords Refugee crisis · Tweets · Sentiment · User community · Indicators

1 Introduction

Social media users are steadily increasing, and they routinely feed information into various sites such as Twitter and Facebook. As can be seen from Fig. 1 (Ref. Statista), the Internet is replacing the traditional media, and social media has become a platform of choice for production, consumption, and diffusion of information (either good or fake). Lotan et al. observe "The shift from an era of broadcast mass media to one of networked digital media has altered both information flows and the nature of news work. Mainstream media (MSM) outlets have adopted Twitter as a means of engaging with and enlarging audiences, strengthening their reach and influence while also changing how they rely on and republish sources. During unplanned or critical world events such as the Tunisian and Egyptian uprisings, MSM turn to Twitter, both to learn from on-the-ground sources, and to rapidly distribute updates" (Lotan et al. 2011). Also, participants in conflicts have used social media for organizing, to advance their ideologies, and for building public support.

Due to the massive increase in data (big data) produced by social media, mining information from the data by business and academic researchers has also increased. There are numerous papers available in the literature, (examples Nerghes and Lee 2018; O'Connor et al. 2010; Ribeiro et al. 2016; Yang and Leskovec 2010) that analyze Tweets and other text data to extract information.

In this paper, we propose a set of indicators as measures of activeness of an event in Twitter and apply them to analyze tweets related to the Syrian refugee crisis. We choose this crisis for evaluating the indicators because of the significance of the event and information extracted about it will be of interest to society. The Syrian Civil War began in 2011 creating a huge humanitarian crisis. Since the crisis began, from the war-torn country of Syria, millions of people have been displaced and well over four million refugees have left (https://partners.twitter.com/content/dam/partners-twitter/success-stories/pdf/SyrianRefugeeCrisisReport-Partnerships.pdf). "UN Calls Syria 'Worst Humanitarian Disaster' since Cold War," (Christian Science Monitor, June 20, 2013) (http://www.csmonitor.com/USA/Foreign-Policy/2013/0620/World-Refugee-Day-UN-calls-Syria-worst-humanitarian-disaster-since-cold-war). As of 2015 "Syrian civil war has to date claimed over 200,000 casualties, including over 8000 documented killings of children under 18 years of age. In a country of approximately 22 million

✉ K. M. George
kmg@cs.okstate.edu

Karsten Ladner
Karsten.Ladner@okbu.edu

Ruchishya Ramineni
rramine@okstate.edu

[1] Department of Computer Science, Oklahoma Baptist University, Shawnee, OK 74804, USA

[2] Computer Science Department, Oklahoma State University, Stillwater, USA

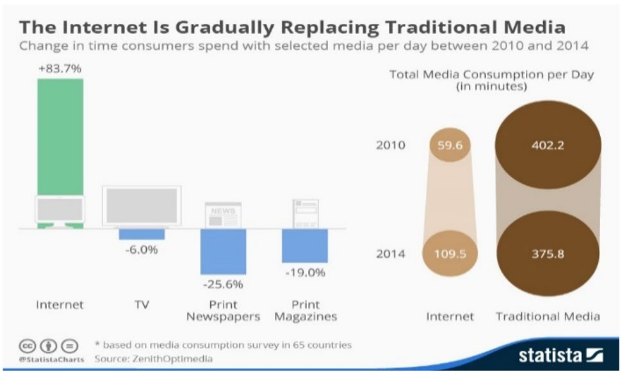

Fig. 1 Internet versus traditional media (Statista)

people, the bloody and prolonged conflict has resulted in 7.6 million internally displaced persons and an additional 3.2 million refugees, as well as approximately 12.2 million people (more than 1 in 2 Syrians) in need of humanitarian aid to survive" (Berti 2015). "Social media can act as the engine room for public engagement with refugees, allowing people to move beyond 'I should do something' to 'I will take action'." (http://theconversation.com/us/topics/syrian-refugees-12615). "Today, Syrians represent the largest refugee group in the world. Since the beginning of the Syrian civil war in 2012, more than 5.2 million Syrians have fled the country as refugees, and about half of these are children." (Smith and Aber 2018). While studies on Syrian refugee crises have addressed a wide range of issues (political, economic, humanitarian, immigration, etc.), the purpose of our research is to understand the activeness of the crisis based on tweets. Even though the crisis has been around for a long time, previous studies have not studied the activeness of the crisis as an event. Intuitively, activeness is the world's interest in or concern for the event. By studying the activeness, one may predict society's concern for the event. As mentioned earlier, social media data provide a wealth of information. For practical reasons, we have collected and analyzed English language tweets only.

In this paper, we characterize activeness by a set of seven indicators (features/variables) and measure those indicators. The changes in strength/magnitude of the indicators are used collectively as measure of activeness. We also examine the pairwise dependencies of the indicators by computing correlation. The indicators that we have considered are tweet counts, retweet counts, follower counts, tweet sentiment positive/negative, daily change in users and sparseness of user communities. Sentiment analysis has been used for analysis of the Syrian crisis (e.g., Öztürk and Ayvaz 2018). Those works compute sentiment of tweets to analyze Twitter users' opinions and compare user attitudes who speak English and other languages. But, there is no previous work (that we are aware of) that studies the crisis from an activeness perspective or combines different indicators that we have listed to form a collective measure of the activeness of the event. The last two of the seven indicators we consider are new proposals. This paper's contribution is to present a multipronged method to analyze an event of significance. New indicators, namely daily user change and community sparseness, are introduced as activeness indicators. We use a simple approach based on a water balance model (Güntner et al. 2004) of surface reservoir to model the change in users as an indicator. The daily users are modeled as a graph, and the number of connected components of the graph is used as the measure of sparseness of communities. Even though our focus is on the Syrian refugee crisis, the activeness analysis method is presented in general terms and applicable to all Twitter events. The positive and negative sentiments are counted as two indicators to capture the full impact on activeness. Internal consistency and principal component analysis are combined to evaluate results and possible prediction. The proposed activeness analysis method is applied to Syrian crisis-related tweets. This is a major crisis and human tragedy of our times, and so, is a significant problem to understand. Several results are computed.

2 Related work

The papers available in the literature approach the analysis of Syrian refugee crisis from different points of view. Berti (2015) examines the impact of the crisis on several aspects—health, education, employment and so on. Blitz (2017) reviews government policies on refugees and asylum seekers. O'Callaghan (2014) studied the role of social media in the Syrian conflict. Several papers used sentiment analysis of tweets to study public sentiment toward the crisis and the refugees who are generated by the crisis. Öztürk and Ayvaz (2018) investigated public sentiments and opinions toward the Syrian refugee crisis by analyzing tweets in English and Turkish languages. Their contribution is a comparison of Turkish and English tweet sentiments. They found the Turkish tweets were more positive than the English tweets. Pope and Griffith (2016) studied the refugee crisis in Europe including Syrian refugees. Their work also is based on sentiment analysis of tweets. Linguistic Inquiry and Word Count (LIWC) were used for sentiment computation. They compared English and German tweets. They found the negative sentiment increased significantly after the terror attacks in Paris and Germany (November 13, 2015 and December 31, 2015). "The sentiment categories of negative emotion, positive emotion, anger and anxiety were analyzed across two populations (English and German speaking) and across 68 days. Two significant events occurred during these 68 days (the Paris Terrorist attacks and the Cologne attacks), and these events were analyzed by considering the four sentiment categories in addition to the frequency of words used in tweets around those days" (Pope and Griffith 2016).

Guille et al. (2013) and Li et al. (2017) provided surveys of work related to information diffusion. They both classified related literature as explanatory models and predictive models. Information diffusion is defined as a piece of information flowing from individual to individual, or other entities such as communities (Li et al. 2017). These entities are represented as nodes in a network and information flow is modeled as contagion from node to node. Batrinca and Treleaven (2015) presented a comprehensive review of software tools for social networking media, wikis, really simple syndication feeds, blogs, newsgroups, chats and news feeds. Stieglitz et al. (2014) stated "Indeed, recent studies and surveys have revealed an emerging need to continuously collect, monitor, analyze, summarize, and visualize relevant information from social interactions and user generated content in various domains such as business, public administration, politics, or consumer decision-making." There are several papers that address specific events or topics. The events considered in Lotan et al. (2011) are the Tunisian and Egyptian revolutions.

One of the tools we used in our method of analysis is sentiment analysis of tweets. Sentiment analysis, also known by other names such as opinion mining and subjectivity analysis, is concerned with extracting subjective information from mostly natural language text. Sentiment analysis can be used to identify the opinion of groups of people. There are numerous papers on sentiment analysis. Ahmed et al. (2015) presented a survey of sentiment analysis, available tools, and applications. Ribeiro et al. (2016) provided a comparison of twenty-four sentiment analysis methods. Their objective is to expose the potential limitations, advantages and disadvantages of the methods. Bakliwa et al. (2013) analyzed tweets on the 2011 Irish general election. Tweets were manually annotated as positive, negative, neutral, or sarcastic for political parties or leaders. Then, they used machine learning approaches for 3-class classification (positive, negative, neutral). Coletto et al. (2016) used discussions of the refugee crisis (refugee crisis perception analysis) in Twitter as a case study to demonstrate a framework to analyze perceptions of social phenomena. They defined three dimensions—special, temporal and sentiment—as the basis for their framework. Yin et al. (2014) introduced social activeness as a means of describing a user's contribution to a community. They defined communities as a group of users and items that have similar interests and properties.

Community detection and characterization of community structure in networks are problems addressed by many researchers. Newman and Girvan presented a class of algorithms (Newman and Girvan 2004). One approach treats community detection as a modularity optimization problem (Newman 2006). Modularity is defined as a multiple of the number of edges falling within a group minus the statistically expected number of edges in a network in which edges are placed randomly. In another paper (Song and Kim 2013), Song and Kim presented a real-time Twitter trend mining system. One part of their work builds user network based on "mention" in tweets. They examine how strongly the mention based groups are structured. They used modularity as the measure of strength. They detected communities and their modularity. Korean presidential election data set was used for experimental analysis. It should be observed that the sparse connection referred to in their paper is a measure of the strength of inter-community connections. So, it is different from the "Sparseness of Community" measure we are presenting in this paper. Our definition of sparseness refers to quantification of disconnected communities formed by all the users. Our intuition based on Theodori (2003) is that activeness of an event decreases as number of disconnected communities related to the event increases.

Cárdenas et al. (2018) view crises as phenomena generated by complex social systems. They "explore a broader view of the crisis phenomena, particularly those affecting social systems, understanding them as natural, collective, unavoidable, and necessary processes for the evolution of a system in continuous adaptation and with increased complexity." This paper has a different approach that focuses on the measurement of social interest in crises.

3 Methodology

In this section, we outline the method proposed in this paper to analyze events. An event is a phenomenon that occurs during a time period. Examples of events are refugee crises, border crises, protests, etc. Activeness of an event signifies the importance of the event to human beings. The general methodology that we propose is to compute activeness indicators of the event at time intervals to construct time series of indicators and use the resulting time series for analysis. We define the terms below to describe our approach:

Event An *event* is defined by a set of words or "n-grams" using these as filters, tweets are collected.

Activeness of an event represents the state of the event. While activeness is an abstract term, a set of indicators is used to present a concrete view. We use these indicators as measures of activeness. A combination of internal consistency (Bollen 1984) and PCA is used for validation and prediction of trend.

In this paper, we compute the seven indicators listed above. As stated previously, the analysis will measure indicators of tweet counts, retweet counts, follower counts, tweet sentiment (positive and negative), change in users and sparseness of user community. Table 1 lists the indicators and measurements. Due to the large volume of data collected, for computational efficiency reasons, all indicators except change in users are counted at 12-h time intervals, while change in users is counted daily. Of these seven variables, the first three are straightforward to compute. To compute tweet sentiments, we made use of the lexicon-based, sentiment engine AFINN (https://finnaarupnielsen.wordpress.com/2011/03/16/afinn-a-new-word-list-for-sentiment-analysis/). Using a lexicon that maps a word to a vector that is positive or negative and has a magnitude that indicates the strength of that direction, AFINN parses a string of text and assigns a score to it. With the set of English tweets, we analyzed the sentiment by running the AFINN engine on the full text of the tweet, if present, or the shorter text field of the tweet, otherwise. Once it had been determined whether a tweet was negative (had a negative score), positive (had a positive score), or neutral (was neither), the corresponding count was incremented.

3.1 Daily change in users (*U*(*t*))

To model the change in users, we adopt ideas from water reservoir models (Güntner et al. 2004). We adopted a simple model $V_t = V_{t-1} + Q_c + Q_{\text{in}} - Q_{\text{out}} - U_{\text{RL}} + (P - E)A_{\text{RL}}$ for a large reservoir RL, "where V_t is the reservoir storage volume at day t, Q_c is the daily inflow from the sub-basin area adjacent to the reservoir after the passage of the cascade of small and medium-sized reservoirs in this sub-basin, Q_{in} is the inflow from all other upstream sub-basins via the river network, delayed by a simple streamflow routing scheme, Q_{out} is the outflow from the large reservoir, U_{RL} is water withdrawal, P and E are precipitation to and evaporation from the reservoir water surface, A_{RL}." Even though analogies (e.g., favorable and unfavorable followers in a period) can be found for the variables P, E, and A_{RL}, for simplicity's sake we count them as zeros. We adopt the model by making the following associations:

Table 1 List of features measured

Indicator name	Measure
Original tweets (OT)	Tweet count
Retweets (RT)	Retweet count
Followers (*F*)	Follower count of original tweeters
Sentiment	
(a) Positive (PS)	AFFIN sentiment value
(b) Negative (NS)	
User change (*U*)	$U(t) = \lvert S(t)\rvert + \lvert U_t\rvert - \lvert U_{t-1}\rvert$
Sparseness of community (SC)	Count *C* of eigenvalue 0 of the Laplacian

Reservoir model term	Change model substitutions
V_t	$U(t)$—number of users on day t
Q_c	Number of original tweeters on day t
Q_{in}	Number of retweeters on day t
Q_{out}	Number of original tweeters left on day t
U_{RL}	Number of retweeters left on day t
P	0
E	0
A_{RL}	0

Consistent with the model, we formulate a simple equation (1) for our computation. Let U_i denote the set of users active by either tweeting or retweeting during the time unit i. Let $U(t)$ denote the count of users during the time unit t. Then, the change model is defined as:

$$U(t) = \lvert S(t)\rvert + \lvert U_t\rvert - \lvert U_{t-1}\rvert, \tag{1}$$

In the reservoir model, $S(t)$ represents water V_{t-1} previously there. In our case, $S(t)$ can be associated with users who are active at the beginning of the time period t; U_{t-1} is the set of users active entering during the time period $t-1$; and U_t the set of users who leave during the time period t. For practical reasons, we assume that $S(t) = 0$, and U_{t-1} is active users during time $t-1$ and U_t is active users during time t. The model captures the daily change by considering the inflow and the outflow. Intuitively speaking, if this value stays constant, then there is no momentum for the event in either direction.

3.2 Sparseness of community (*C*(*t*))

A community is represented as a graph with individual users as vertices and relations as edges. Intuitively speaking, we may view the existence of unconnected subcommunities as weakness of the total community. Many subcommunities make the community sparse or fragmented. The community of an event is the set of users engaged in the event. So the activeness of an event can be captured by the strength/sparseness of the community.

In order to define community sparseness, we adopt ideas from spectral graph theory. We take advantage of the well-known relationship between the graph Laplacian and its eigenvalues. For the sake of completeness, we include related definitions and results below:

We define a simple graph $G = (V, E)$ where $V = \{v_1, \ldots, v_n\}$ is the set of vertices and $E = \{(v_i, v_j) \mid v_i$ and v_j are distinct vertices$\}$ is the set of edges. The adjacency matrix A of a graph G with n vertices is defined as a n-by-n square matrix A where the entry $(A)_{i,j}$ is 1 if there is an edge $e = (v_i, v_j)$ otherwise 0. Also, the diagonal elements of A are zeros. Let $D = \text{diag}(d_1, \ldots, d_n)$ be the diagonal matrix such that d_i = degree of the vertex v_i. The degree d of a vertex v is the number of vertices in G that are adjacent to v (i.e., $d_v = \sum_u e_{u,v}$). The sum of the degrees of all vertices is defined as the volume of the graph G ($vol\ G$). Then the Laplacian of G is defined as $L = D - A$. The Laplacian is a symmetric matrix of dimension n-by-n. Therefore, its eigenvalues are all real nonnegative. The smallest eigenvalue is zero. Its multiplicity k is equal to the number connected components of the graph. In this paper, we assume that G is a graph whose vertices represent Twitter users. So, G denotes the community of users who either tweet or retweet. There is an edge defined between two vertices if the user denoted by one vertex retweets the tweet of the user denoted by the other vertex. Let k be the multiplicity of the eigenvalue zero of the Laplacian of G. Then, we define k as the sparseness of the community defined by the graph G.

In this paper, we are concerned with the Syrian refugee crisis and tweets are used as the basis for analysis. So, the community in this case is the set of all users who tweet or retweet about the crisis. We build undirected graph G with vertices V representing Twitter users and (undirected) edges representing undirected retweet relation.

Let C denote the sparseness of G. Then $C = k$ where k is multiplicity of the eigenvalue zero of the Laplacian of G. Observe that $0 \leq C \leq n - 1$ where n is the number of vertices of G. If $C = 0$, then G is fully connected and has only one component. If $C = n - 1$, then G consists of only isolated vertices. As the sparseness measure can be the same for a very small graph and a very large graph, we also define a normalized sparseness measure as $C_{\text{norm}} = C/vol\ G$. We will make use of the following two results:

$vol\ G$ = trace L = sum of the eigenvalues of L

A high level algorithm for the computation of community sparseness (normalized and unnormalized) is given in Algorithm 1.

ALGORITHM I: SPARSENES COMPUTATION
Input: set of tweets.
Output: C, and C_{norm}
Step 1: Construct the adjacency matrix A from the tweets using the retweets as the relation
Step 2: Construct the diagonal matrix D where
$D_{ii} = \sum_{j=1}^{v} A_{ij}$
Step 3: Construct L = D - A
Step 4: Compute eigenvalues of L
Step 5: set C = multiplicity of eigenvalue 0.
Step 6: compute $vol\ G = \sum_1^n \lambda_i, \lambda_i\ eigenvalus$
Step 7: compute C_{norm} = C / vol G
Step 8: output C and C_{norm}.

The computation time of the algorithm will be dominated by step 4, which is the computation of the eigenvalues. The size of the matrix A is another concern. Due to implementation difficulties caused by matrix size, an ad hoc approach was taken for implementation of Algorithm 1. We split the algorithm into two parts and implemented it using Python and MatLab. Steps 1–3 of the algorithm were implemented using Python. Retweet is the relation used to build the adjacency matrix A. This was executed on the name node of a Hadoop cluster with approximately 98 TB of available storage. For eigenvalue computations, we used MatLab available in a different system.

3.3 Justification for indicators choice

The seven indicators and their measurements are listed in Table 1. Each indicator measures a different feature of the tweets associated with an event. The first three indicators are counts that measure three different groups in relation to tweets and increased count will indicate increased activeness of the event. The fourth and fifth indicators are sentiment measures of tweets. Sentiment is used to measure user opinions. So, the tweet sentiments (positive/negative/neutral) are reflections of how users feel about the event. Thus, a high level of sentiment, either positive or negative, indicate increased interest of users in the event, and hence, the event can be viewed as more active. If new users continue to participate that gives new life to the activeness of the event which is the justification for the sixth indicator to be considered. In other words, it is like diffusion of news. The more people participate, the more popular the news item becomes. The last feature measures the connectedness of users and interaction among them. We also find support for the justification in a previously published article titled "Activeness refers to the degree of interaction at the community level" (Theodori 2003). "Community-level interactions include activities such as participating in a community improvement project or working with other members of the community to try and solve local problems." So connectedness of the network is related to activeness. Sparseness indicates fragmentation and less fragmentation implies more activeness.

In the next section, we present the empirical results based on the data we have collected for a period of 2 months from July 18 to September 18, 2018, related to the event "Syrian Refugee Crisis." We also apply the methodology to a set of tweets collected with the key phrase "illegal immigration." Analyses based on the computed results are provided in Sect. 5.

4 Empirical results

The results presented in this section are outcomes of computations based on the data that we collected during July 18 through September 18, 2018, on the topic "Syrian Refugee Crisis." Measures of the seven indices are presented as time series covering the period of data collection. Our data collection started in late May 2018. However, there were some

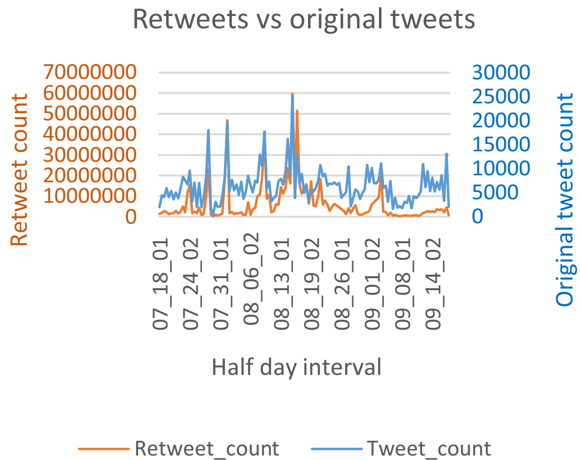

Fig. 2 Retweets compared with original tweets

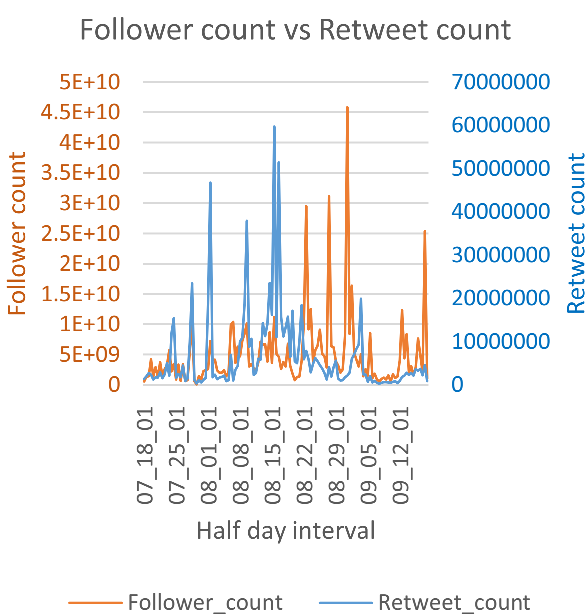

Fig. 3 Comparing retweets to follower count

discontinuities in data collection during June and so that data is not included for the analysis presented in this section. However, some insightful information derived from data collected during June and early part of July is given in the "Appendix." The key words used for data collection and filtering were Assad, chemical, chemical attack, ISIS, ISIL, Syria, Syrian, and refugee. Then we grouped the tweets by key words (second filtering).

Results presented in this section are based on two groups of tweets. One group has the term refugee(s) present in every tweet and the other group has the term Syria present in every tweet. For the sake of convenience, we call them the Refugee group and the Syria group. Figures 2, 3, 4, 5, 6, 7, 8, 9,

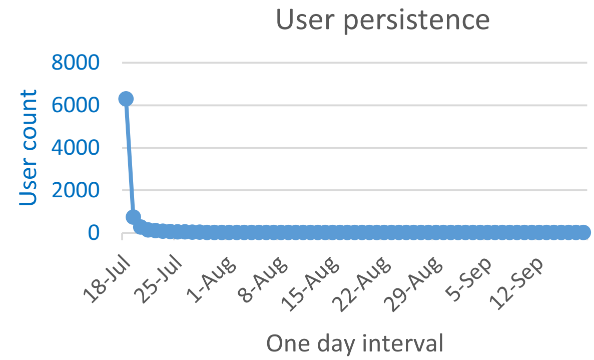

Fig. 4 Same user presence over time

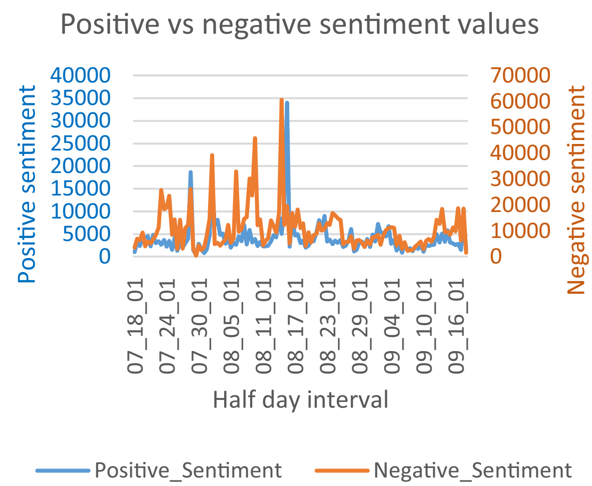

Fig. 5 Positive and negative tweets

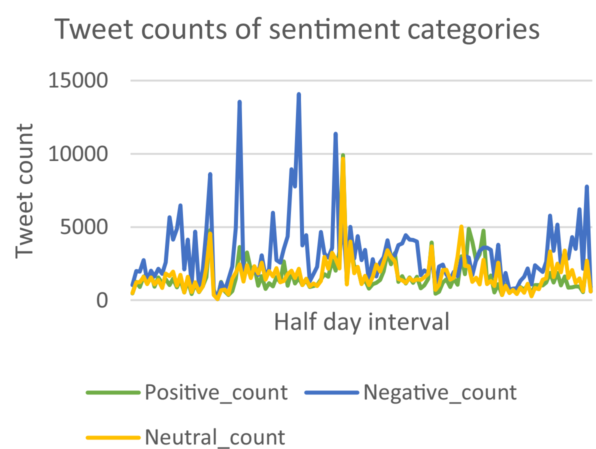

Fig. 6 Comparison of tweet counts by sentiment

10, and 11 are based on the analysis of tweets grouped as Refugee and Figs. 12, 13, 14, 15, 16, and 17 are based on the group termed Syria. We limited analysis of data only to these

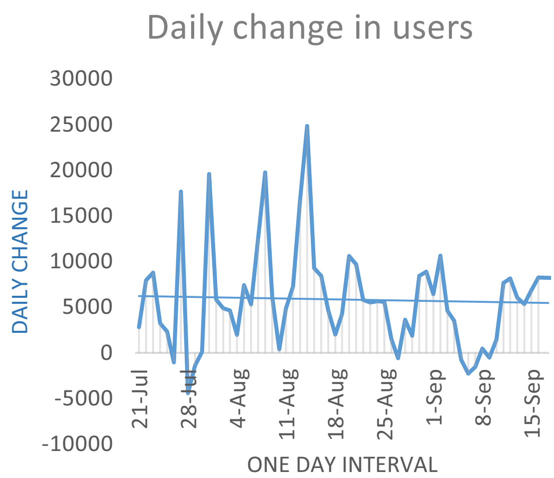

Fig. 7 Daily change of users, inflow–outflow

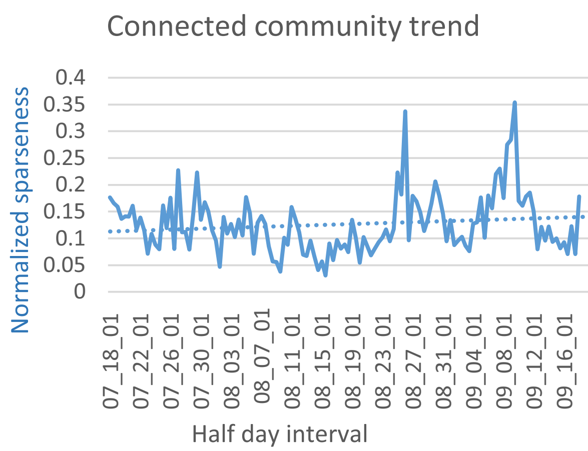

Fig. 8 Connected communities (fragmentation)

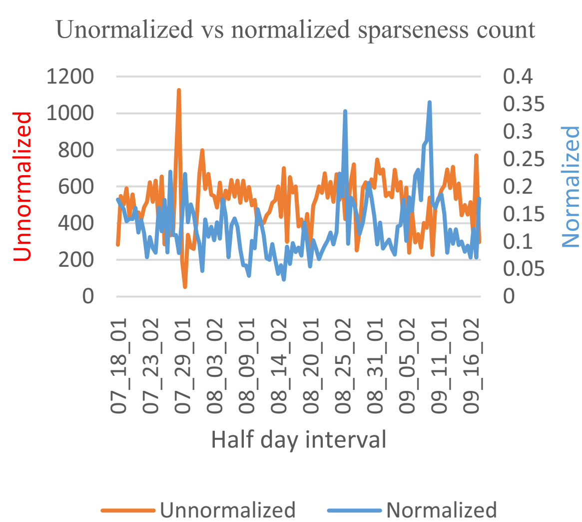

Fig. 9 Unnormalized versus normalized sparseness measure

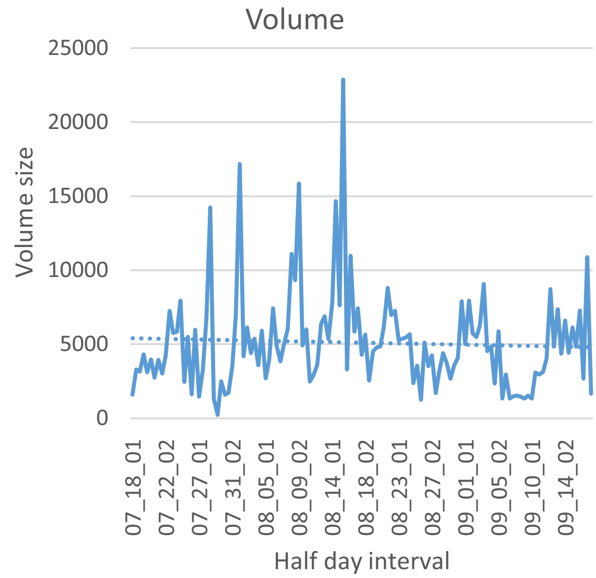

Fig. 10 Volume of the user graph

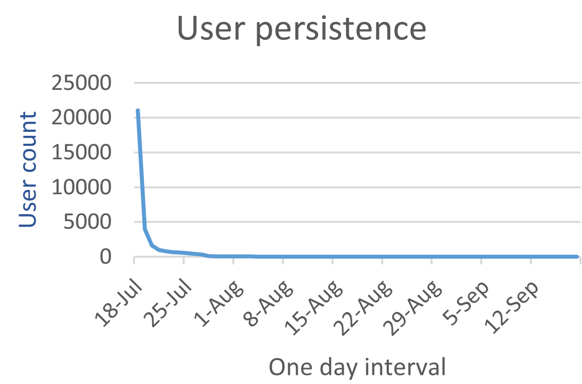

Fig. 11 Same user presence over time

two groups for the following reasons: (a) Random inspection of the grouped tweets suggest that the aforementioned two groups are the ones closely related to the topic of refugee crisis. (b) Another motivation for limiting the analysis based on only these two groups is to reduce data to a size manageable in our systems.

4.1 Refugees group

In Figs. 2 and 3, we provide the counts of tweets, retweets and followers as time series. They also provide visual comparison. As the volume of data happened to be high, these counts are presented per 12 h periods as time units. Figure 4 shows the length of time a user is active either by tweeting or retweeting measured in days. As can be seen from Fig. 4, a user is active for only a short period continuously. The

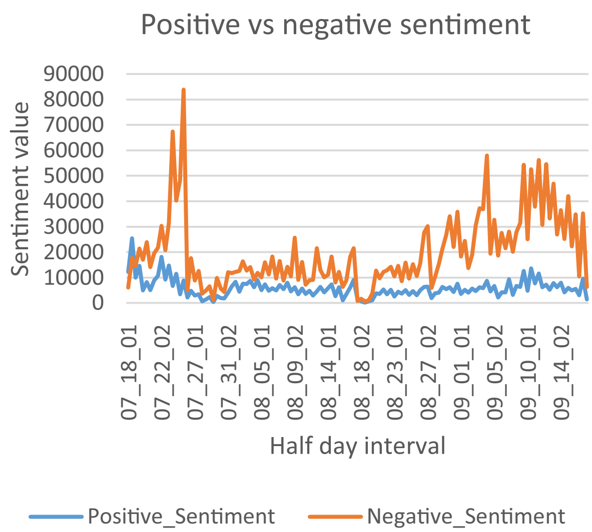

Fig. 12 Positive versus negative sentiment

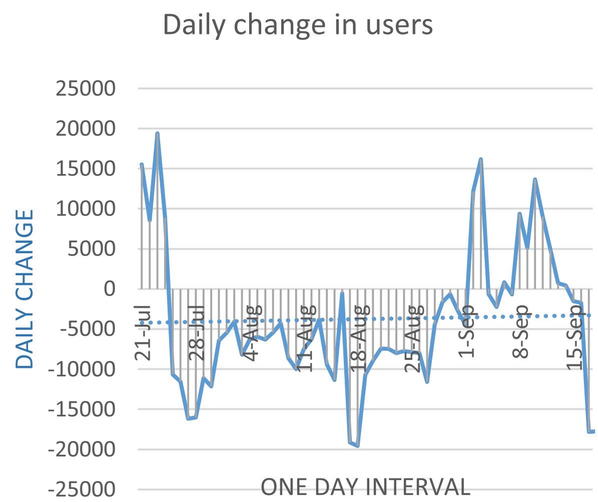

Fig. 14 Daily change in count of user, inflow–outflow

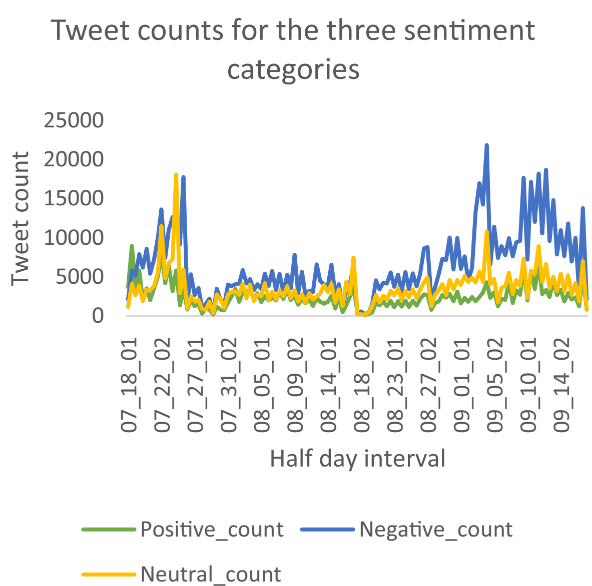

Fig. 13 Comparison of tweet counts by sentiment

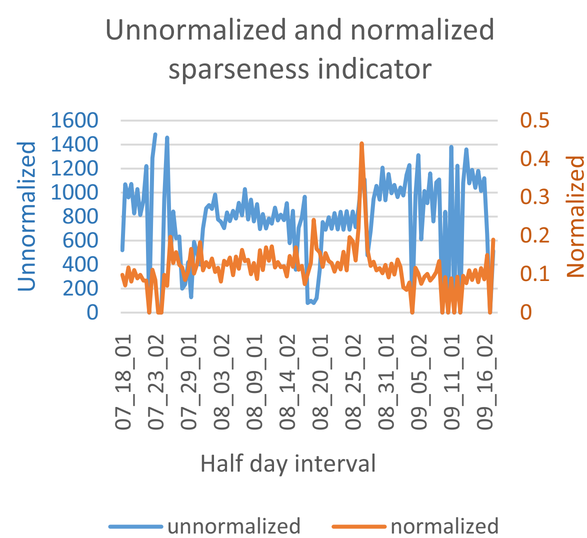

Fig. 15 Unormalized versus normalized sparseness measure

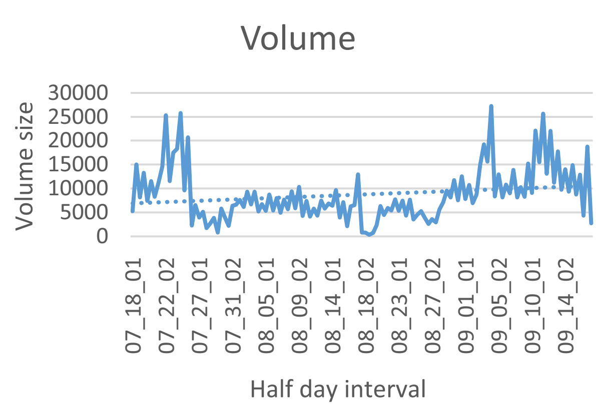

Fig. 16 Volume of the user graph

x-axis is the days. However, some organizational users or robots have a continuous presence.

Positive and negative sentiment values of tweets are shown in Fig. 5. The actual tweet counts producing the sentiments are given in Fig. 6. It shows the time series for tweet counts of positive, negative and neutral sentiment values. The tweet counts are positively correlated.

Figure 7 shows the daily change and the trend. Daily user count change is computed for a 24-h period instead of 12-h period. While daily change varies, the trend is negative. We computed the covariance of inflow and outflow of users, which is positive.

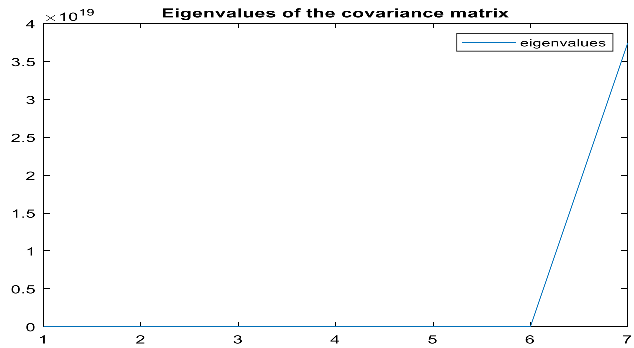

Fig. 17 Eigenvalues of the covariance matrix

Figures 8, 9 and 10 are related to the sparseness of the community indicator computed using Algorithm 1. Figure 8 shows the normalized community sparseness measure (i.e., connected sub-community normalized count C_{norm}) over the 2-month period. The trend is shown by the dotted line. As can be seen from the figure, the trend is almost neutral indicating no substantial change in the community behavior. However, the slight upward bias indicates the fragmentation is increasing. The values of C fluctuate during the period. Figure 9 shows the normalized and unnormalized sparseness measures C and $C_{norm.}$ for comparison. It turns out the correlation of C and C_{norm}. is negative. The spikes in C and C_{norm}. are on different days. This is due to the effect of normalization, as C and C_{norm}.are inversely proportional, and Table 2 shows the pairwise correlation of all indicators.

While all indicators are positively correlated, the normalized sparseness indicator C_{norm} is negatively correlated with all other indicators. This is the correct behavior since high sparseness is associated with weak activeness. (In order that increases in all measures indicate the same activeness behavior, we need to consider $1/C_{norm}$ Rather than C_{norm} in computing internal consistency.) Figure 10 shows the daily volume of the user graph, *vol G*, and the trend. While there is daily fluctuation in the volume of the graph, the trend does not change. The volume is positively correlated with the unnormalized sparseness indicator.

Figures 11, 12, 13, 14, 15 and 16 show results of the same computations based on the tweets grouped by the key word Syria. The pairwise correlation table is shown in Table 3. As can be seen, all indicators are pairwise positively correlated as in the previous case. Similar to the previous case, the unnormalized sparseness of the community indicator C_{norm} is negatively correlated with all other indicators.

4.2 Key word group Syria

Patterns of the graphs comparing tweets vs retweets and retweets vs followers are similar to Figs. 2 and 3. Therefore, the figures are not presented here. The correlation between original tweet count and retweet count is 0.53, and the covariance is 48,136,484,856. The correlation between retweet count and follower count is 0.78 and the covariance is 5.38048E+17. We have also measured the length of time a user might stay engaged. As can be seen from Fig. 11,

Table 2 Pairwise correlation and covariance of indicators—refugees tweet set

	OT	RT	*F*	PS	NS	*U*	NSC	SC
OT		0.79	0.49	0.80	0.77	0.23	0.84	−0.40
RT	+		0.15	0.70	0.61	0.20	0.80	−0.18
F	+	+		0.32	024	0.05	0.27	−0.23
PS	+	+	+		0.36	0.12	0.64	−0.27
NS	+	+	+	+		0.32	0.75	−0.28
U	+	+	+	+	+		0.36	−0.15
NSC	+	+	+	+	+	+		−0.21

NSC normalized SC

Table 3 Pairwise correlation and covariance of indicators—Syria tweet set

	OT	RT	*F*	PS	NS	*U*	NSC	SC
OT		0.53	0.41	0.70	0.88	0.65	0.68	−0.55
RT	+		0.78	0.30	0.50	0.44	0.54	−0.18
F	+	+		0.09	0.34	0.39	0.46	−0.11
PS	+	+	+		0.48	0.50	0.49	−0.42
NS	+	+	+	+		0.59	0.58	−0.44
U	+	+	+	+	+		0.50	−0.46
NSC	+	+	+	+	+	+		−0.23

NSC normalized SC

similar to Fig. 4 in this case also, users stay active continuously for only a short period of time.

Figure 12 shows the behavior of positive and negative sentiment. The two sentiments are positively correlated with the correlation 0.48. Figure 13 shows the tweet counts that produce these sentiment values counts for positive, negative, and neutral sentiments are shown. One obvious observation is the negative sentiment and corresponding user count being significantly higher even though they are positively correlated.

The correlations among tweet counts producing positive and negative sentiments are higher than 0.66. Figure 14 shows the daily change in user count. It looks different from Fig. 7 and the trend line is up. Figure 15 shows the normalized and unnormalized sparsity measures C and C_{norm}. As expected, they are negatively correlated. Figure 16 shows the volumes of the user graphs during the analysis period. The trend line is also shown which slopes slightly upwards.

5 Analysis of results

In the previous section, we provided the results that are measures of the proposed activeness indicators. The computations are based on two groups of tweets related to the Syrian refugee crisis. The tweets were grouped by key words "refugees" and "Syria." The two groups are not mutually exclusive. The "refugees" group data size is 37.45G, and the "Syria" group data size is 69.08G.

There are some obvious observations one can make: (1) negative sentiments and tweet counts are noticeably higher than positive sentiments and counts. Furthermore, positive counts are even lower than neutral counts. This would imply that Twitter users are not favorable to the refugee crisis or Syria. We performed ANOVA and F-tests on the positive and negative sentiment series. Both reject the null hypotheses indicating difference of the populations that are favorable and unfavorable. (2) There are spikes in all graphs. They probably are driven by news stories and should be viewed as outliers.

Based on Figs. 4 and 11, one can assert that most users do not stay active for a long time contiguously. Only interest groups and few individuals are persistent in their daily presence of tweets (examples are UNCHR, and National Refugee Council). We have repeated the computations on the tweets after removing all tweets originated by these groups. The results do not indicate any significant impact. We show total tweet counts-related indicators for the refugees group of tweets in Table 4. The corresponding data for Syria group of tweets is given in Table 5. We computed correlations of all corresponding indicator values. Except for daily user change, all correlations

Table 4 Effect of removing certain users (Refugee group)

Indicator	Count
Positive tweet—all users	197,818
Positive tweet count—institutional users removed	196,859
Negative tweet—all users	394,113
Negative tweet count—institutional users removed	394,113
Retweet count—all users	8.09E+08
Retweet count—institutional users removed	8.06E+08
Followers count—all users	6.1972E+11
Followers count—institutional users removed	5.65E+11

Table 5 Effect of removing certain users (Syria group)

Indicator	Count
Positive tweet—all users	303,653
Positive tweet count—institutional users removed	302,768
Negative tweet—all users	815,589
Negative tweet count—institutional users removed	813,358
Retweet count—all users	9.52E+08
Retweet count—institutional users removed	9.36E+08
Followers count—all users	2.63E+12
Followers count—institutional users removed	2.55E+12

are approximately equal to 1. Correlation of daily user change values is greater than 0.6. As can be observed from Tables 4 and 5, counts with removing users and without removing users do not have significant difference. The differences are relatively small. Therefore, we can conclude that institutional users have no influence on the outcome of activeness.

The unnormalized and normalized sparseness indicator shows the community of all users is sparsely connected. This is possibly a consequence of the users being mostly negative and being active for a short period of time. The activeness of the event is mostly driven by the negatively biased users. Also, the correlation between the inflow and outflow of users is positive. In both Tables 2 and 3, we have used the reciprocal of the measures C and C_{norm}. This is to make sure all indicators are consistent in reflecting the direction of activeness.

By looking at the behavior of indicators individually, we can conclude that the activeness of the crisis will be trending low slowly. As there is no benchmark or method available for validating the indicators, we tried other analysis methods for determining the validity of the computed results. We propose to use two measures (a) internal consistency and (b) PCA analysis. Furthermore, we applied the methodology to a different dataset collected using key phrase "illegal immigration."

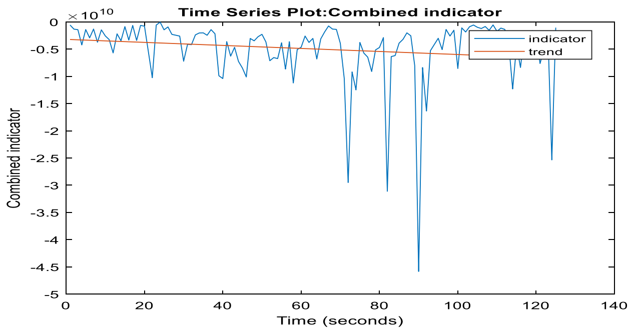

Fig. 18 One-dimensional representation of the indicators

5.1 Internal consistency

Internal consistency of indicators measuring a concept (in our case, activeness) refers to the property that all indicators are good indicators of the concept. Positive correlations between the indicators are a necessary condition for validity (Bollen 1984). Cronbach's alpha (Cronbach 1951) is a measure used to evaluate internal consistency. We adopt the formula: $\alpha = \frac{Nr}{1+(N-1)r}$ where N is the number of indicators and r is the average pairwise correlation among all indicators. A higher alpha indicates higher consistency. Tables 2 and 3 contain the pairwise correlations of the indicators related to refugee tweet set and Syria tweet set, respectively. The average of correlations is 0.47 and 0.52. The alpha values are 0.86 and 0.88, respectively. Alpha values are close to 1, showing that the indicators are internally consistent.

5.2 PCA analysis

We performed a PCA analysis. This is to combine the results of all indicators by dimension reduction. The results are shown for the "refugees" data. The covariance matrix of the indicators was used for the analysis. A plot of the seven eigenvalues $(\lambda_1, \ldots, \lambda_7)$ of the covariance matrix is shown in Fig. 17. Furthermore, $\frac{\lambda_7}{\lambda_1+\cdots+\lambda_7} \approx 1$ implies λ_7 can explain all variance in the indicators. By projection, we obtain the time series shown in Fig. 18. It shows a negative trend for the activeness which is mostly negative. Graphs of user networks of the two groups of tweets at the beginning of the analysis period and end of the analysis period are shown in Figs. 19, 20, 21 and 22. The changes in the networks are not significant. This agrees with the other results which suggest that the activeness of the event has not changed significantly during the 2-month analysis period. The total data sizes at the beginning of the analysis period and at the end were the same for the "refugees" group and down for the "Syria" group.

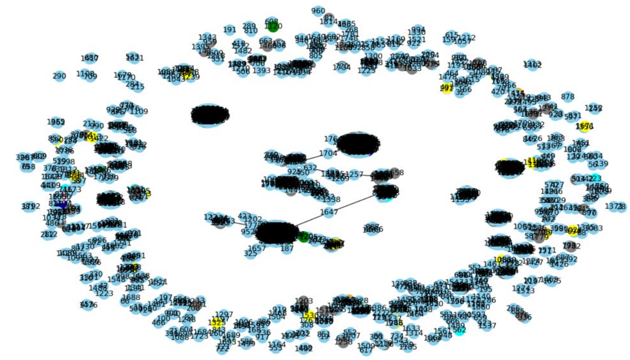

Fig. 19 User network Refugees group 7-18-2018

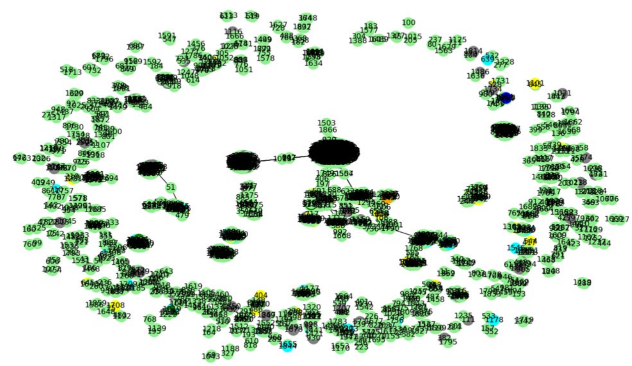

Fig. 20 User network Refugees group 9-18-2018

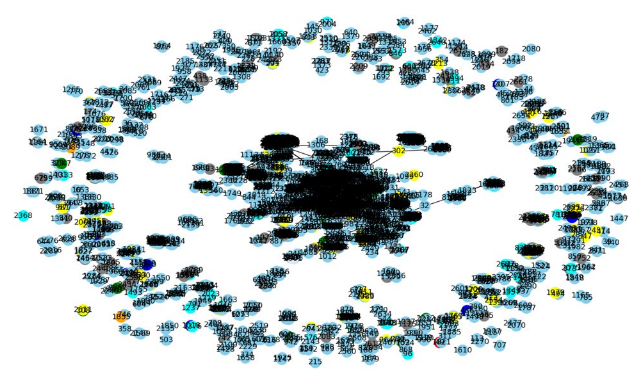

Fig. 21 User network "Syria" group 7-18-2018

In the "Appendix," we show geographic locations of English tweeters. Results are shown in Figs. 30, 31 and 32. The data collection period was from May 31 to July 16, 2018. We have only considered English tweets. This is the reason for Middle Eastern countries not showing up in Fig. 31. The figures show that most tweeters are from the USA and within the USA, and most active users were in the state California.

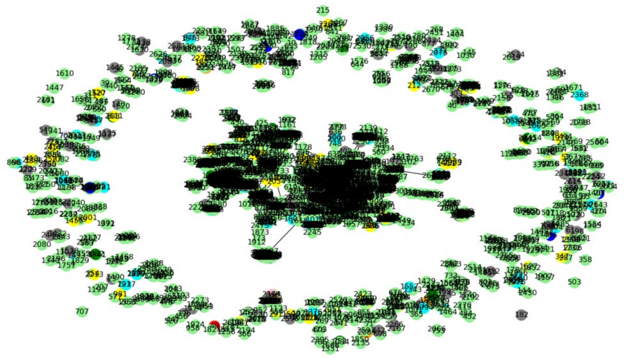

Fig. 22 User network Refugees group 9-18-2018

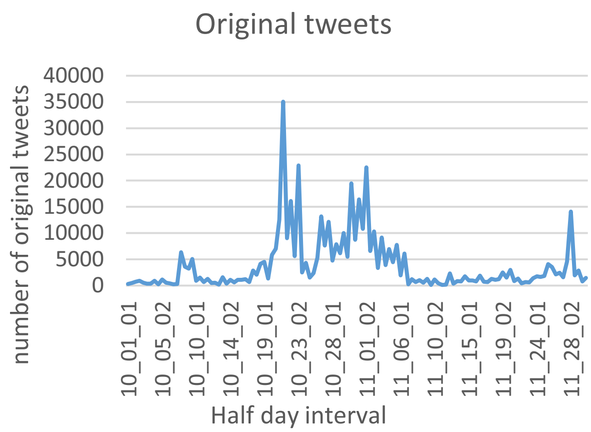

Fig. 23 Original tweet counts

5.3 Validation by applying to a different event

In order to demonstrate the usability and validity of the model to analyze other events, we applied it to a different dataset. This data set consists of tweets filtered by the phrase "illegal immigration" from a larger set of tweets collected using key words "immigration," "separation," "crime," "illegal," "boarder," and "parent." We filtered the data to reduce the size so that we can complete computations using our available resources. The reduced data size is 2.1 GB. We searched for persistent users in the filtered tweet set. Only one user is found to be persistent. It is a legitimate user and not a bot or news media. Figures 23, 24, 25, 26, 27, 28, and 29 show the results of computation of indicators as time series. The period of spikes in the indicators can be explained by news about "migration caravan" and US midterm election speeches.

Table 6 shows pairwise correlation of indicators. Average correlation of pairwise indicators is 0.66. Therefore, the Cronbach's alpha value is 0.93 suggesting very good internal consistency of the indicators.

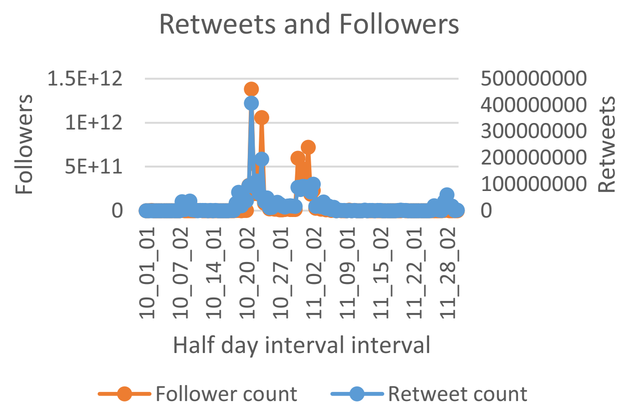

Fig. 24 Retweets and followers

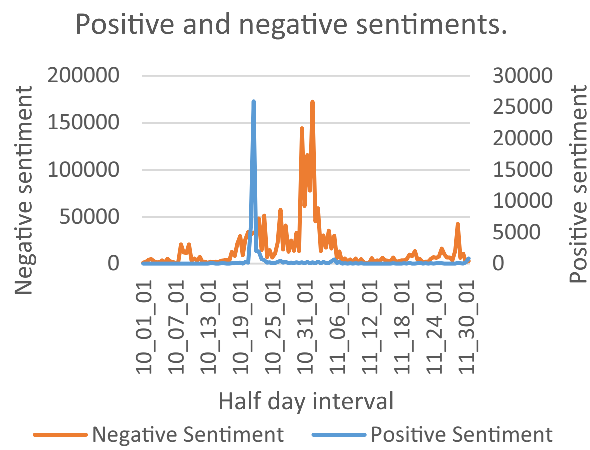

Fig. 25 Positive and negative sentiments

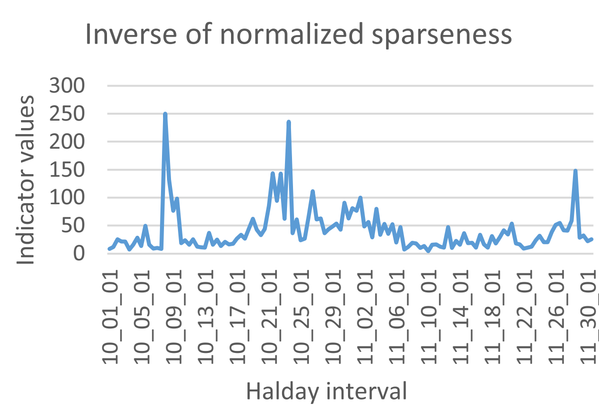

Fig. 26 Sparseness indicator

PCA is used to combine all indicators into one. The eigenvalues of the covariance matrix of indicators are shown in Fig. 28. As in the case of the "refugee data," there is one dominant eigenvalue and $\frac{\lambda_7}{\lambda_1+\cdots+\lambda_7} \approx 1$.

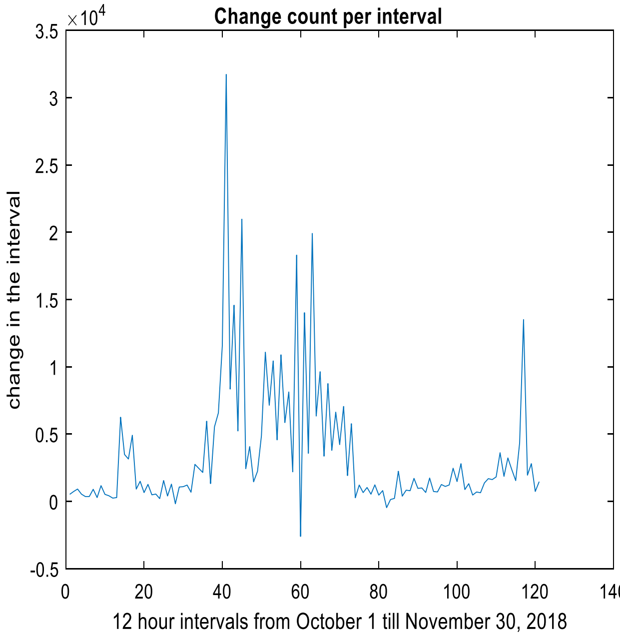

Fig. 27 User change inflow–outflow: 12-h periods

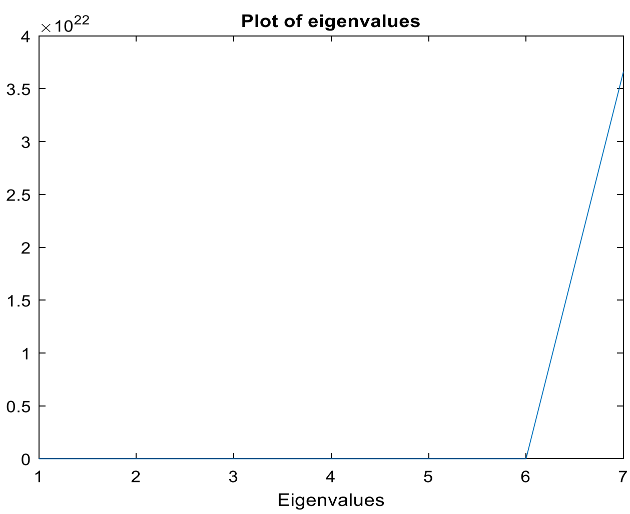

Fig. 28 Eigenvalues of covariance matrix

As there is only one dominant eigenvalue and all indicators satisfy high internal consistency, we can combine them into one indicator which can be used to forecast activeness of the event. Values of the combined indicator are shown in Fig. 29 as a time series. The spikes in the graph correspond to the spikes in other graphs. The trend is negative which indicates the activeness after the US midterm election.

As stated in Sect. 3, activeness is an abstract term with respect to an event. A concrete view is presented by a set of indicators. With that perspective, results about the indicators translate to results about activeness of the respective event.

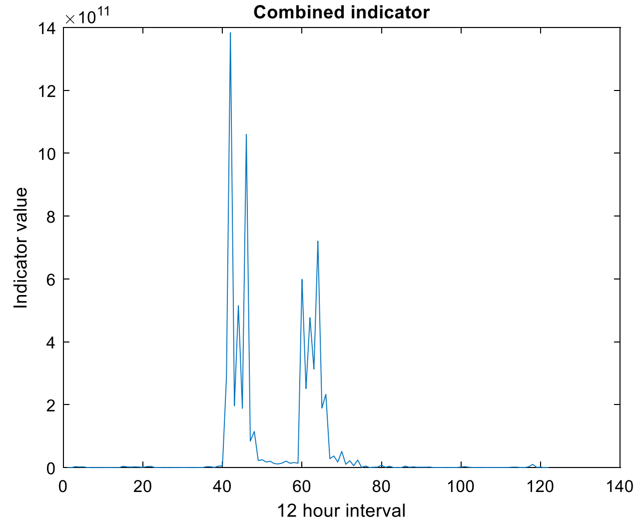

Fig. 29 Combined indicator (projection to first principal component)

"Syrian Refugee Crisis" is the event studied in this research. Twitter data were collected using several key words. The analyses presented in this paper are based on two data sets filtered using the key words "refugees" and "Syria." Severn indicators are proposed.

Pairwise correlation of the indicators are all positive (refer to Tables 2, 3). The indicators are internally consistent meaning they represent the same concept/topic. So, they collectively present a concrete view of activeness at any given time. As correlation is a measure of similarity, we can assume that one indicator will capture properties of the event missed by other indicators.

6 Conclusion

In this paper, we analyzed the Syrian refugee crisis using related tweets. This is an event which is viewed as a great human tragedy of our times. To do the analysis, we present a new approach by defining activeness of an event and choosing several indicators as measures. According to Jacobson and Lalu (1974), use of only one indicator to measure an underlying variable is limited due to varying reasons. One of the reasons is "the abstract quality of most theoretical variables, which does not permit a useful summarization when a single indicator is employed." So, a multiple indicator approach is desirable. The indicators together are expected to explain human engagement with respect to the event and people's opinion. In this paper, we select seven indicators from the tweets to explain the concept of activeness of the event. We have considered only English tweets and performed extensive experiments. Generally speaking, our analysis shows that people who are active users of Twitter are positioned against the

Table 6 Correlation between indicators

1.00	0.88	0.87	0.59	0.78	0.75	0.39
+	1.00	0.92	0.80	0.53	0.64	0.39
+	+	1.00	0.68	0.63	0.60	0.25
+	+	+	1.00	0.11	0.30	0.24
+	+	+	+	1.00	0.53	0.28
+	+	+	+	+	1.00	0.30
+	+	+	+	+	+	1.00

Syrian refugee crisis. The magnitude of negative sentiment and the corresponding user count far exceeds the positive sentiment and count. The user community as a whole is sparsely segmented into smaller disconnected communities. All indicators are internally consistent, and so we build one combined indicator. We have used the first principal component for this purpose. The trend of the combined indicator is negative, which is the current trend. Our indicator approach could be used to analyze other events as demonstrated by the analysis of immigration related tweets.

Acknowledgements Karsten Ladner was supported by NSF REU site grant.

Appendix

In this appendix, we provide the results of our analysis based on data collected during May 31 through July 16, 2018. Due to system problems, data were not available on some days. So we were unable to perform daily analysis reliably. The summary of results is provided. Figure 30 compares the tweet sentiments from different countries on three different random days. It can be seen that USA dominates on all 3 days. As we considered only English tweets, countries from the Middle East appear in yellow region.

Figure 31 shows the tweets originating from countries where English is spoken. As can be seen, most tweets originate from the USA. Figure 32 shows the states from where the tweets are originating.

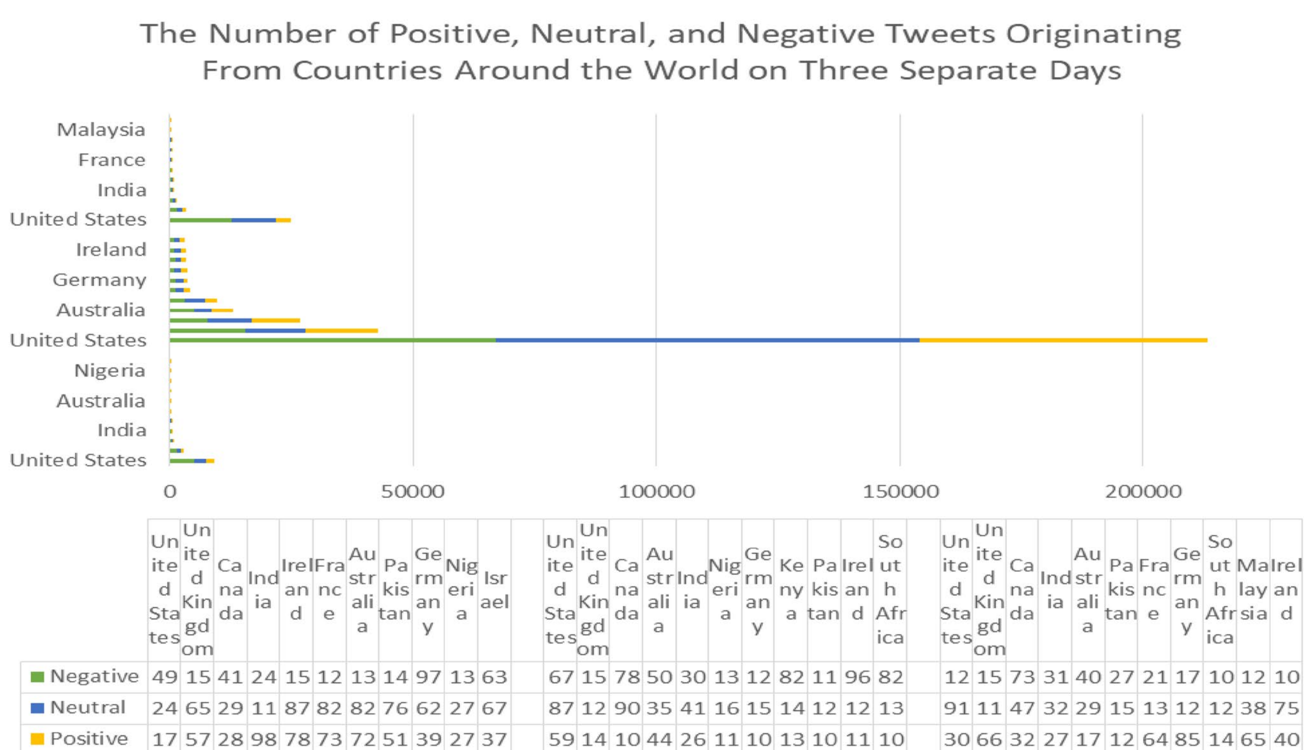

Fig. 30 Comparison of counties based on tweet sentiments for 3 days

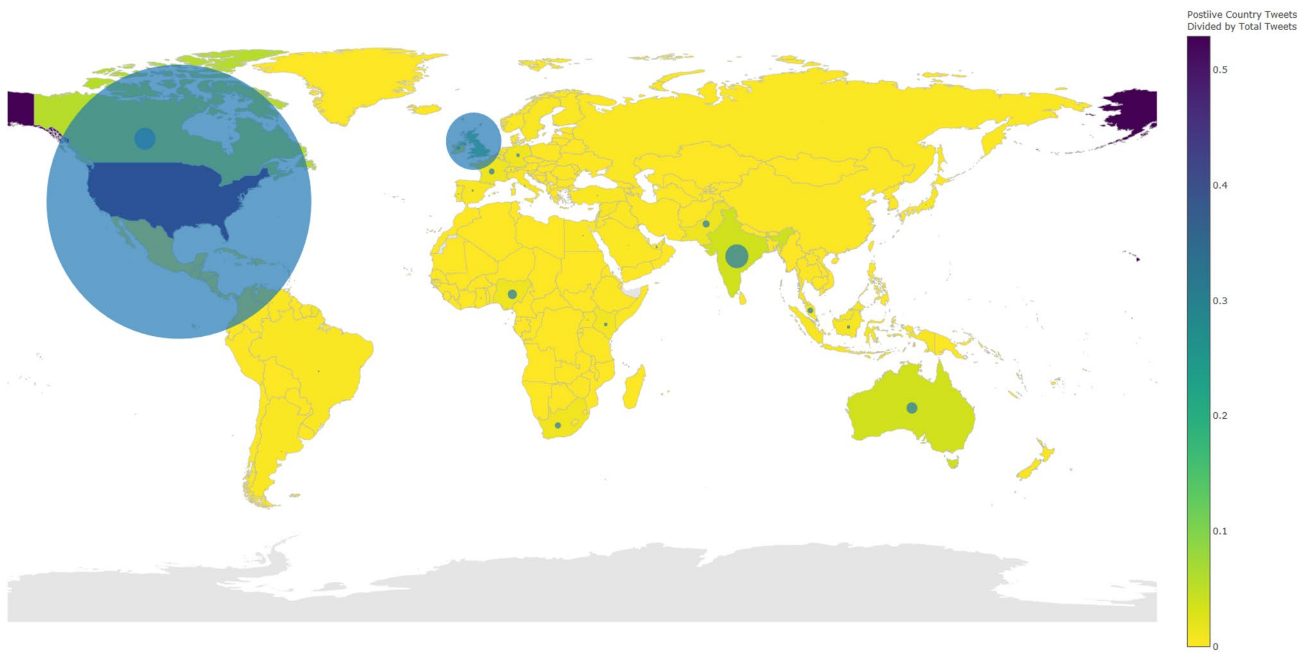

Fig. 31 Tweets originating from countries

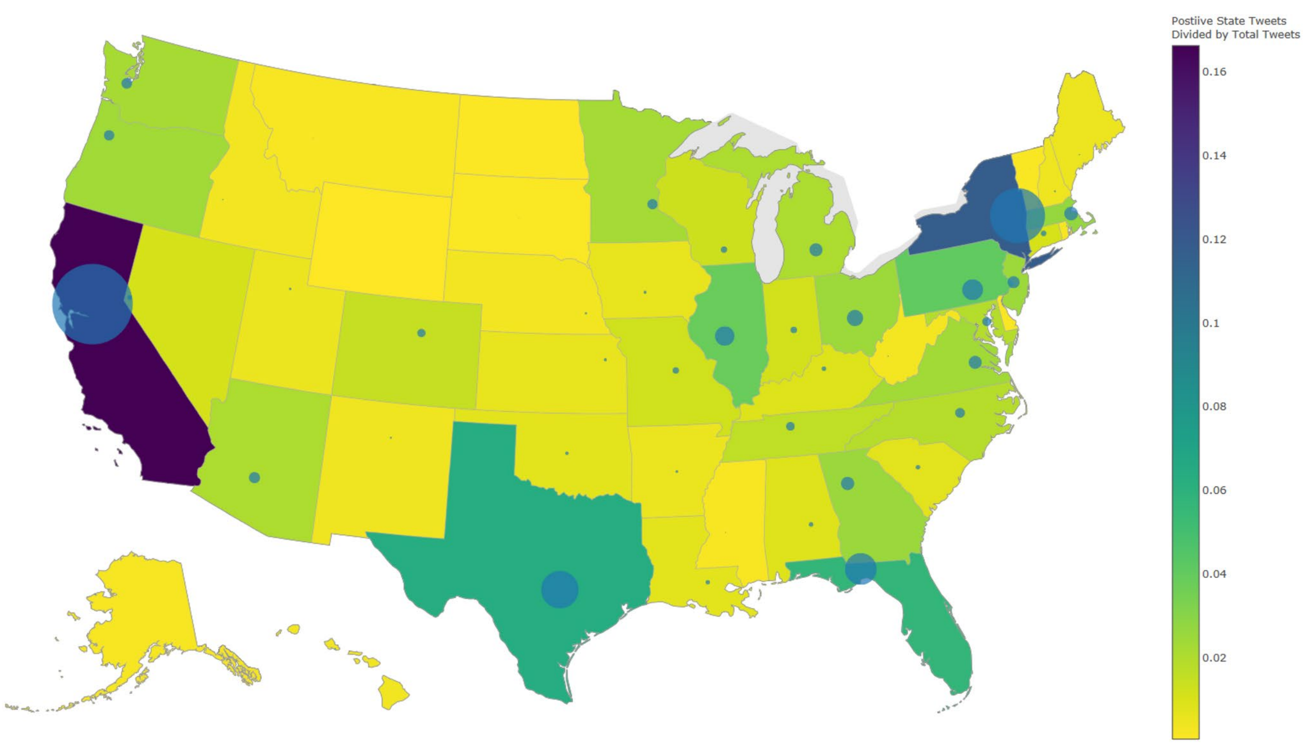

Fig. 32 Tweets originating from the states of the USA

References

Ahmed K, Tazi NE, Hossny AH (2015) Sentiment analysis over social networks: an overview. In: 2015 IEEE international conference on systems, man, and cybernetics, pp 2174–2179

Bakliwa A et al (2013) Sentiment analysis of political tweets: towards an accurate classifier. In: Proceedings of the workshop

on language in social media (LASM 2013), Atlanta, Georgia, June 13 2013, pp 49–58
Batrinca B, Treleaven PC (2015) Social media analytics: a survey of techniques, tools and platforms. AI Soc 30:89–116
Berti B (2015) The syrian refugee crisis: regional and human security implications. Strateg Assess 17(4):41–53
Blitz B (2017) Another story: what public opinion data tell us about refugee and humanitarian policy. J Migr Hum Secur 5(2):379–400
Bollen KA (1984) Multiple indicators: internal consistency or no necessary relationship? Quality Quant 18(1984):377–385
Cárdenas JP, Vidal G, Urbina C, Olivares G, Rodrigo P, Fuentes M (2018) Social crises: signatures of complexity in a fast-growing economy. Complexity. https://doi.org/10.1155/2018/9343451
Coletto M, Esuli A, Lucchese C, Muntean CI, Nardini FM, Perego R, Renso C (2016) Sentiment-enhanced multidimensional analysis of online social networks: Perception of the mediterranean refugees crisis. In: En 2016 IEEE/ACM international conference on advances in social networks analysis and mining (ASONAM), 2016, pp 1270–1277
Cronbach LJ (1951) Coefficient alpha and the internal structure of tests. Psychometrika 16:297–334
Guille A, Hacid H, Favre C, Zighed DA (2013) Information diffusion in online social networks: a survey. SIGMOD Rec 42(2):17–28
Güntner A, Krol MS, De Araújo JC, Bronstert A (2004) Simple water balance modelling of surface reservoir systems in a large datascarce semiarid region. Hydrol Sci J 49(5):901–918
Jacobson AL, Lalu NM (1974) An empirical and algebraic analysis of alternative techniques for measuring unobserved variables. In: Blalock HM (ed) Measurement in the social sciences. Aldine, Chicago, pp 215–241
Li M, Wang X, Gao K, Zhang S (2017) A survey on information diffusion in online social networks: models and methods. Information 8(4):118. https://doi.org/10.3390/info8040118
Lotan G, Graeff E, Ananny M, Gaffney D, Pearce I, Boyd D (2011) The revolutions were tweeted: information flows during the 2011 Tunisian and Egyptian revolutions. Int J Commun 5:1375–1405
Nerghes A, Lee JS (2018) The refugee/migrant crisis dichotomy on Twitter: a network and sentiment perspective. In: WebSci'18, May 27–30, 2018, Amsterdam, Netherlands, pp 271–280
Newman MEJ (2006) Modularity and community structure in networks. PNAS 103(23):8577–8582
Newman MEJ, Girvan M (2004) Finding and evaluating community structure in networks. Phys Rev E 69:026113
O'Callaghan et al (2014) Online social media in the Syria conflict: encompassing the extremes and the in-betweens. In: IEEE/ACM international conference on advances in social networks analysis and mining (ASONAM 2014)
O'Connor B, Balasubramanyan R, Routledge BR, Smith NA (2010) From tweets to polls: linking text sentiment to public opinion time series. In: Association for the advancement of artificial intelligence, pp 122–129
Öztürk N, Ayvaz S (2018) Sentiment analysis on Twitter: a text mining approach to the Syrian refugee crisis. Telemat Inform 23:136–147
Pope D, Griffith J (2016) An analysis of online Twitter sentiment surrounding the european refugee crisis. In: Proceedings of the 8th international joint conference on knowledge discovery, knowledge engineering and knowledge management (IC3 K 2016)—Volume 1: KDIR, pp 299–306
Ribeiro FN, Araújo MA, Gonçalves P, Gonçalves MA, Benevenuto F (2016) SentiBench—a benchmark comparison of state-of-the practice sentiment analysis methods. EPJ Data Sci 5:23
Smith SR, Aber JL (2018) Increasing understanding for syrian refugee children with empirical evidence. In: Vulnerable children and youth studies, Volume 13, 2018—Issue 1: Special Section: Increasing Understanding for Syrian Refugee Children with Empirical Evidence, pp 1–6
Song M, Kim MC (2013) RT^{2}M : Real-time Twitter trend mining system. In: 2013 International conference on social intelligence and technology, pp 64–71
Stieglitz S, Dang-Xuan L, Bruns A, Neuberger C (2014) Social media analytics an interdisciplinary approach and its implications for information systems. Bus Inf Syst Eng 6(2):89–96
Theodori GL (2003) The community activeness—consciousness matrix. J Ext 41(5). https://www.joe.org/joe/2003october/tt2.php
Yang J, Leskovec J (2010) Modeling information diffusion in implicit networks. In: IEEE international conference on data mining, pp 599–608
Yin B, Yang Y, Liu W (2014) Exploring social activeness and dynamic interest in community-based recommender system. In: International World Wide Web conference (IW3C2), April 7–11, 2014, pp 771–776