On the Volume of Geo-referenced Tweets and Their Relationship to Events Relevant for Migration Tracking

Georg Neubauer, Hermann Huber, Armin Vogl, Bettina Jager, Alexander Preinerstorfer, Stefan Schirnhofer, Gerald Schimak, Denis Havlik

On the Volume of Geo-referenced Tweets and their Relationship to Events Relevant for Migration Tracking

Georg Neubauer[1], Hermann Huber[1], Armin Vogl[2], Bettina Jager[1], Alexander Preinerstorfer[1], Stefan Schirnhofer[1] and Gerald Schimak[1]

[1]AIT Austrian Institute of Technology GmbH, Vienna, Austria
{Georg.Neubauer,Hermann.Huber, Bettina.Jager, Alexander.Preinerstorfer, Stefan.Schirnhofer, Gerald.Schimak}@ait.ac.at
[2]Federal Ministry of Interior, Republic of Austria, Vienna, Austria
Armin.vogl@bmi.gv.at

Abstract. Migration is a major challenge for the European Union, resulting in early preparedness being an imperative for target states and their stakeholders such as border police forces. This preparedness is necessary for multiple reasons, including the provision of adequate search and rescue measures. To support preparedness, there is a need for early indicators for detection of developing migratory push-factors related to imminent migration flows. To address this need, we have investigated the daily number of geo-referenced Tweets in three regions of Ukraine and the whole of Japan from August 2014 until October 2014. This analysis was done by using the data handling tool Ubicity. Additionally, we have identified days when relevant natural, civil or political events took place in order to identify possible event triggered changes of the daily number of Tweets. In all the examined Ukrainian regions a considerable increase in the number of daily Tweets was observed for the election day of a new parliament. Furthermore, we identified a significant decrease in the number of daily Tweets for the Crimea for the whole examined period which could be related to the political changes that took place. The natural disasters identified in Japan do not show a clear relationship with the changes in the degree of use of the social media tool Twitter. The results are a good basis to use communication patterns as future key indicator for migration analysis.

Keywords: migration · tweets · geolocation · early indicator · push factor

1 Introduction

In recent years the importance of migration has increased considerably in EU member states, and is seen as one of the major challenges by the European Union. A number of events have led to issues of migration becoming more important in Europe, including the "Arab spring" crisis situations in North-Africa, natural disasters such as draughts or large scale flooding, and wars and warlike situations in the Middle-East, Sub-Sahara and South Asia. These events have caused an overwhelming migration

adfa, p. 1, 2011.

wave with a steadily growing number of refugees who are trying to reach European countries. European societies are struggling with the socioeconomic impact of these crises. Consequently, these societies are searching for political solutions which can cope with the humanitarian responsibilities, as well as the protection of their demographic and economic structures.

It is important for possible destination countries to be well-prepared for such large influxes of refugees so they can grant migrants a liveable and worthy reception. As an example, the southern border of Italy is facing a soaring migratory wave, as more than 100.000 illegal migrants had to be rescued in open-sea during 2014 and brought to the Italian coast. This mass immigration caused a socioeconomic crisis in this region. Human disasters, such as the drowning of hundreds of human beings, could be avoided if better information is available to alert search and rescue teams in timely manner. Currently, European authorities seem ill-prepared to cope with this overwhelming situation and therefore new information sources have to be found in order to better observe possible migration-causing incidents. Early and reliable indicators on migration movements are therefore imperative for multiple stakeholders such as border police forces or first responders in the field. So far, data on volume and flow of migration is mainly inconsistent, outdated or does not exist at all. This paper investigates the suitability of using the volume of geo-referenced Tweets versus time as potential early indicators of migration movements, whilst taking into account both long term impacting situations and specific high-profile events.

1.1 Related Work

In recent years, extensive research was done in identifying events by studying human communication behavior on social media. In this context the micro-blogging platform Twitter has been the focus of much research. The benefit of gathering information about events from social media services, such as Twitter, relates to the fact that people turn to these platforms in the face of exceptional circumstances [1]. This study stresses that irrespective of whether these events are emergencies, natural disasters or political protests, a significant factor for changes in communication patterns is a disruption to normal routines. Besides a certain content of Twitter messages, [2] pointed out, that volunteered information with geographic footprints is displaying people's current position, which can facilitate a wide range of possibilities to support situational awareness. The tracking of recent trends in migration patterns by using geo-located data gathered from Twitter has been discussed in [3] as a specific type of application.

Based on the monitoring of real-time migration flows, statements can be made on migration trends, i.e. increasing or decreasing mobility from a certain country to OECD countries. Political and civil events are frequently chosen as research use cases. An analysis of Twitter data in the reference period of five months exposed a relation between communications behaviour and the occurrence of exceptional events during the troubled time. As a main result of [4], highly publicised events in Egypt have been reflected in the amount of communication via Twitter in a certain spatial area. Furthermore, the paper states that the amount of communication in different Egyptian cities is comparable.

Another study focusing on the Middle East was made by [5] and showed the potential for the detection of trending topics on the basis of content-based geo-location of Arabic and English language Tweets. An extensive range of applications has been presented by [6], who have employed techniques for the SNOW Data Challenge 2014. By combining aggressive filtering of Tweets with hierarchical clustering of Tweets, events around the US presidential elections in 2012 and the recent events in Ukraine, etc. have been detected.

1.2 Developments in the Aftermath of the Ukrainian Conflict

In the course of the escalating conflict between Russia and Ukraine in the early 2014, a Ukrainian crisis emerged. Accompanied by demonstrations at the Independence Square in Kiev a strong commitment to Europe had been expressed by Ukrainian protesters, which has acquired visibility during the Euromaidan protest wave. Although the situation in the Crimea has seemed to calm down since the Crimean status referendum in March 2014, [7] pointed out, that the annexation of the Crimea has caused radical changes, especially for ethnic Ukrainians, Crimean Tatars and representatives of minority groups generally.

An increasing number of incidents have been recorded in the east and Crimea, which have been identified as an important factor for emigration from the Crimea. A research of the conflict events in the archive of Center for Strategic and International Studies (CSIS)[1] delivers therefore a huge number of hits. Table 1 illustrates the most relevant events for each category summarized on a monthly basis. The categories "civil events" and "political events" comprise the majority of results. In particular the combats in the eastern Ukraine have been highly publicized. The majority of events can be clearly attributed to one of the pre-defined categories. Solely, the category "natural events" remains vacant, which can be explained by the overwhelming media response to human induced events or the fact, that there were no natural induced disasters in the period from August 2014 to October 2014.

Table 1. Events in Ukraine in the second reference period (August 2014 – October 2014)

Month	No. of Events	Exemplary Event
August 2014	32	Over 300 Ukrainian soldiers flee across border to Russia European leaders threaten Russia with further sanctions Suspicion surrounds Russian convoy carrying aid to Ukraine
September 2014	28	Ceasefire strained as fighting near Donetsk kills nine Party of regions plans election boycott as fighting continues Drop in Russian gas supplies to Europe
October 2014	27	Separatists are violating ceasefire Ukraine elects new parliament Russia and Ukraine remain at impasse over gas

[1] http://csis.org/ukraine/index.htm#130

1.3 Environmental Events in Japan

Japan is one of the countries most affected by natural disasters. Due to its geographic location in the pacific in the geologic fracture zone of four tectonic plates, earthquakes happen frequently in the Japanese archipelago. Since Japan has over 100 active volcanoes eruptions are also occurring occasionally in Japan. Other relevant disasters caused by natural events in Japan are extreme temperatures, floods, storms (e.g. tropical cyclones) and mass movements [8]. Japan leads in the list of economic damage due to earthquakes in the world from 1900 up to the present day with the huge number of 360 billion US $ [9]. In the actual World Risk Report [10] Japan is in the top five countries with the highest urban risk. Urban risk is defined as the product of urban vulnerability and urban hazards. Several natural events occurred in Japan during the observation period from August 2014 up to October 2014, which can be seen in Table 2. These natural events were extracted using the search engine GLIDEnumber [11].

Table 2. Natural events in Japan in the period (August 2014 – October 2014) extracted from GLIDEnumber V2.0 [11]

Date	Type of Event	Event
08/08/2014	Tropical Cyclone	Typhoon Halong killed one person in Japan and injured 33, as authorities ordered 1.6 million people out of the path of the storm that battered the west of the country
20/08/2014	Landslide	At least 36 people were killed in Japan on 20 August 2014, when landslides triggered by torrential rain
27/09/2014	Volcanic Eruption	Mt. Ontake, a central Japan volcano popular with tourists particularly in the fall, erupted without warning just before noon 27 September 2014
06/10/2014	Tropical Cyclone	At least one person was dead and six were missing on 6 October as a strong typhoon whipped through the Tokyo metropolitan area after making landfall further south
12/10/2014	Tropical Cyclone	Typhoon Vongfong, the strongest storm to hit Japan this year, has made landfall on the country's main islands

2 Method

2.1 Data Acquisition

In contrast to many other platforms, Twitters data access policy is quite liberal. The programming interface provides access to 1% of the twitter traffic on a just-in-time basis. Users are free to downstream either a random sample or provide data filters in order to decrease data volume and ensure that Tweets match various criteria like the presence of certain hashtags. However, if the data volume after the application of filters still exceeds the 1% threshold, twitter randomly truncates the subset to meet the data limitation policy. In order to examine spatiotemporal frequencies, we applied a geo-location filter to only download Tweets that feature geographic positioning attributes in terms of latitude and longitude. In this way, Twitter provides roughly 7

million geo-referenced Tweets a day, originating from mobile devices equipped with GPS sensors.

During the three month evaluation period we downloaded 698 million Tweets (261GB). AIT has developed a data analysis tool called Ubicity in order to extract, aggregate and analyse social media data. It routes the data between all kinds of data stores and analysis components like search engines, graph databases or mobile phones. We used its geo-polygon query capability to extract the tweets emitted in certain areas of the Ukraine and Japan.

2.2 Data Analysis

We approximated the national borders of the Ukraine and Japan and retrieved all of the collected messages tweeted within the geo-polygons. In the case of Ukraine, we subdivided the territory into three subareas with respect to the distribution of the Russian language according to a national census in 2001 [12]. Polygon (a) in Fig. 1 depicts the area in which less than 10% of the population identified Russian as their native language. Polygons (b) and (c) depict areas with a rate higher than 50%. Area (b) encompasses the administrative districts Donetsk and Luhansk; (c) encompasses the Crimea peninsula. We selected the three Ukrainian areas because of their geo-political importance in the conflict between the Ukrainian government and the Pro-Russian separatists.

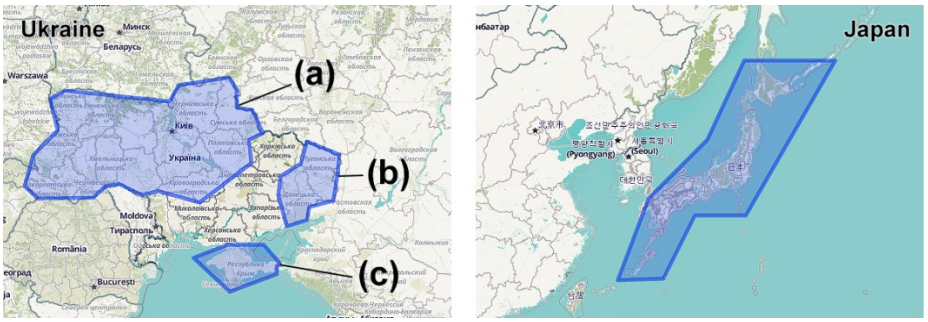

Fig. 1. The Geo-Polygons approximate the national borders of parts of the Ukraine and Japan. We analyzed all Tweets provided by the Twitter public streaming interface submitted within these areas – (a) West Ukraine, (b) East Ukraine and (c) Crimea

In the period of investigation, we leached from a total number of 698 million Tweets (261GB), 1.1 million Tweets in Ukrainian areas and 25.9 million in the Japanese polygon. The geo-polygon query consists of multiple latitude/longitude pairs to approximate the borders of the respective country. As our twitter stream was partly interrupted due to maintenance work, we identified those interruptions of at least one hour and excluded affected days from our investigation. Our analysis is based on the spatiotemporal dimension only. The main advantage of this approach is that it is language and text understanding independent. Advanced analysis might also take term

frequency and co-occurrence as well as hashtags and network relations into account. Depending on the use case we identified tight filters as a promising solution. Keeping the data rates low (below 1% of the overall traffic) avoids spurious frequency changes in the first place.

3 Results

Statistical tests highlight, that the communication behavior differs significantly between Japan, the Crimea and the regions in the Ukraine (Table 3). In comparison to the Japanese Tweets, the number of Tweets is lower by a factor of rounded 30 in the region of the West of Ukraine (see the mean values). However these numbers do not take the difference in the number of inhabitants between the regions from Ukraine and Japan into account.

Table 3. Statistical parameters for Tweets gathered in Japan, the Crimea and the East and the West of Ukraine

Statistical Parameter	Japan	Crimea	East Ukraine	West Ukraine
Mean	375316.8	1314.8	1416.3	13868.8
Median	375090.0	1105.0	1407.0	13774.0
Maximum	498847.0	2882.0	1855.0	17383.0
Minimum	212669.0	306.0	1094.0	10791.0
1st Quartile	328671.0	706.5	1309.0	13251.0
2nd Quartile	375090.0	1105.0	1407.0	13774.0
3rd Quartile	419217.0	1954.3	1498.0	14516.5

3.1 Results Obtained for the Crimea, the East and the West of Ukraine

As illustrated in Fig. 1 (left picture), the analysis of Tweets in the Crimea, the East and the West of Ukraine considers the three regions separately. Fig. 2 displays that the course of the curve shows a permanently downward trend from the beginning of the reference period to the end. Apart from the peaks in the second half of August, no comparable extent of Tweets could be reached in later stages.

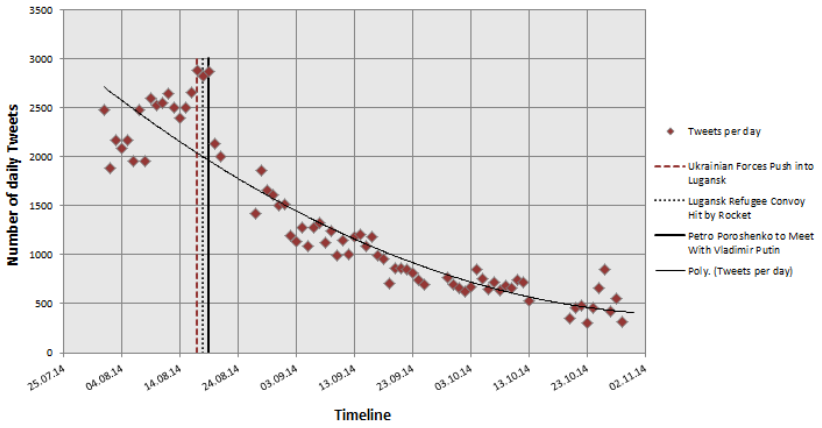

Fig. 2. Number of Tweets from August 2014 to October 2014 in Crimea

The flat curve shape of the number of Tweets (not shown here) indicates a moderate increase over the period of three months in the East of Ukraine. Determined by daily peaks, the events around the foray of Ukrainian forces into Lugansk and Donetsk are reflected in the communication identified in the eastern region of the Ukraine. Aside from the middle section of the curve in Fig. 3, a permanent decrease of the number of Tweets per day in the West Ukraine can be recognized. The line is slightly curved and the most disruptive events are not clearly reflected therein. Mirrored by monthly top rates a marked increase of communication can be noticed at the beginning of September 2014. In this time frame, NATO became more visible in Ukrainian conflict and in general, the largest number of political events had been observed regarding, foreign sanctions and economic degradation.

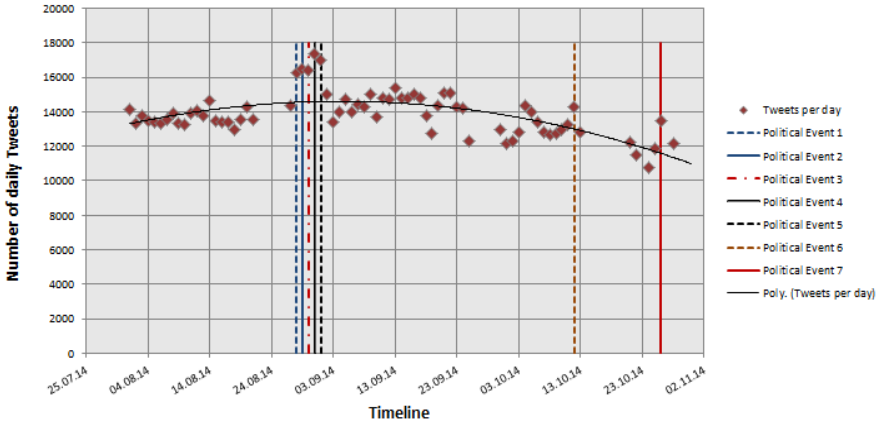

Fig. 3. Number of Tweets from August 2014 to October 2014 in West Ukraine

3.2 Results Obtained for Japan

The timeline of the daily number of geo-located Tweets in Japan describes a linear decrease over the whole period (Figure not shown). An interesting result is that the communication behavior around all the tropical cyclones is located over the trend line, but after the landslide no such trend is recognizable. This might be explained by the missing data from our dataset, which was caused by system maintenance. In this time period no political events with an increase of Twitter communication were found. On a culture-bounded basis, Japan naturally exhibits a higher number of Tweets when compared to the other observed areas – the Crimea, East and West Ukraine.

4 Discussion

4.1 Discussion of Communication Patterns Related to Natural Events

Analysis of the Tweets from Japan showed no clear patterns. Five major natural disasters and three political events were identified in the period of investigation. Looking at the two cyclones on August 8th and October 12th 2014, an increase of about 10% in number of Tweets was observed the day after the event. However, a decrease of about 16% was observed for the cyclone occurring on October 6th. The observed changes lay within typical variations of the daily number of Tweets. A landslide took place on August 20th causing 36 casualties. This event did not lead to an increase of messages on the day of the event nor in the days afterwards. This is in contrast to the results from [4], in which a snowstorm in Egypt was associated with the highest number of Tweets in the observed period. The reason for this difference might be the uniqueness of the Egyptian snowstorm, whereas cyclones are rather common events in Japan. The identified political events did not occur with an obvious increased number of Tweets.

4.2 Discussion of Communication Patterns related to Man - Made Events

We examined three regions in Ukraine in order to identify possible differences in communication patterns between these regions that were caused by different human induced factors. We identified for almost every day a political or civil event, however no natural disaster was documented. It is interesting to note that in all examined regions, the number of Tweets augmented on October 26th, the day the election of a new parliament took place.

The Crimea shows a remarkable decrease in the daily number of Tweets during the three investigated months. The median of daily Tweets in August is 2,279, whereas the median in the month of October went down to 647, corresponding to 28 % of the Tweets of August. Apart from the peaks in the second half of August, no comparable extent of Tweets was identified afterwards. There are a number of possible reasons for this decrease. For instance, new types of social media might have been introduced by the new potentates. Another possibility would be the restriction of communication services in general or a new taxing policy making use of social media not

affordable to a considerable part of the population. Finally, it is possible that a large proportion of former Twitter users just left the region.

On August 17th the highest number of geo-referenced daily Tweets was observed, this day a push of Ukrainian forces into Lugansk and Donetsk was reported. In the eastern region of Ukraine, the highest number of daily Tweets can be seen for September 20th, the day a large explosion took place in Donetsk. Compared to the median of all daily Tweets this corresponds to an increase of about 32% in the total number of Tweets.

In contrast to the investigations on Egypt performed by [4], in which the number of Tweets across the whole Egypt and the cities of Cairo and Alexandria correlated highly, correlations were low for the regions of Ukraine, e.g., 0.16 in case of the West of Ukraine and Crimea. Communication patterns using Twitter seem to differ considerably between the examined regions in Ukraine.

Looking at the methodology applied to extract Tweets several aspects need to be discussed in detail. The authors of [13] indicate that the 1% streaming limitation is not constant over time. Hence, long-term frequency analysis that is based on the public twitter programming interface requires advanced data normalization approaches to avoid misinterpretation. Our approach to focus on the amount of Tweets in a predefined area builds on the concept, which is based on the assumption that monitoring of the dimension of the crowd at a certain hotspot is more meaningful than the content of the communication. Without pursuing a thematic priority the approach of extracting geo-references from Tweets facilitate novel insights in human behavior and opens multiple application, as mentioned in [6], [14] [15] and [16]. Especially, the study of [16] have provided an outlook on how Social Media technology can be used as a support module for real time alerting systems for any type of event.

5 Conclusion

It is the purpose of our ongoing investigations to develop a method for the identification of early indicators of migration movements. The method has to optimize required man power of the stakeholders (e.g. border police forces) as well as their time investment, it has to serve to respond to developing or already ongoing migration movements in very early stages in order to mitigate the effects of such developments. The method consists of several steps:

1. Identification of typical communication patterns such as elevated number of daily Tweets associated with natural as well as man-made events having the potential to become push factors for migration based on analysis of historical data

2. Continuously observation of the quantity of social media communication for unstable regions without analyzing the content and comparison the developing patterns of communication with archival patterns being typical for migration triggering events. Identification of developing situations affecting the well-being of migration endangered populations leading to the imminent readiness for emigration by observing the individual base relevance of the region

3. In case of identifying communication patterns indicative for migration developments such as exceptional frequency of Tweets or non-typical increases or decreases of daily number of Tweets more detailed investigations such as text mining can be performed to obtain a more precise operational picture. Moreover, profound knowledge on the ongoing developments in specific critical regions is required to interpret changes in communication patterns correctly

So far, we examine the relation between communication patterns and historical data and restrict our work to item 1 of the above described approach. We intend to find specific types of events such as riots or natural disasters, that are associated with typical "communication curves". In some cases several developments in communication intensity may become potential indicators of migration movements. For instance, the very large decrease in transmitted Tweets in Crimea is likely to be a sign for one or several civil and political changes occurring in a period of several months.

References

1. K. Starbird, "Crowdwork, Crisis and Convergence: How the Connected Crowd Organizes Information during Mass Disruption Events," Colorado, 2012.
2. A. Stefanidis, A. Crooks and J. Radzikowski, "Harvesting ambient geospatial information from social media feeds," *GeoJournal,* vol. 78, no. 2, pp. 319-338, April 2011.
3. E. G. V. Zagheni and I. Weber, "Inferring international and internal migration patterns from Twitter data," in *Proceedings of the companion publication of the 23rd international conference on World wide web companion*, Geneva, Switzerland, 2014.
4. G. Neubauer, H. Huber, B. Jager und A. Vogl, „Detecting Events in Egypt Based on Geo-Referenced Tweets," in *22nd Proceedings of IDIMT-2014. Networking Societies – Cooperation and Conflict*, Poděbrady, Czech Republic, 2014.
5. S. Khanwalkar, M. Seldin, A. Srivastava, A. Kumar and S. Colbath, "Content-Based Geo-Location Detection for Placing Tweets Pertaining To Trending News on Map," in *The Fourth International Workshop on Mining Ubiquitous and Social Environments (MUSE'13)*, Prague, Czech Republic, 2013.
6. G. Ifrim, B. Shi and I. Brigadir, "Event Detection in Twitter using Aggressive Filtering and Hierarchical Tweet Clustering," in *Proceedings of the SNOW 2014 Data Challenge*, Seoul, Korea, 2014.
7. Office of the United Nations High Commissioner for Human Rights, "Report on the human rights situation in Ukraine," 2014.
8. preventionweb, [Online]. Available:

http://www.preventionweb.net/english/countries/statistics/?cid=87. [Accessed 5th November 2014].

9. Statista, [Online]. Available: http://de.statista.com/statistik/daten/studie/163492/umfrage/oekonomischer-schaden-durch-erdbeben-nach-laendern/. [Accessed 5th November 2014].
10. United Nations University-Institute for Environment and Human Security, "World Risk Report 2014," Bonn, Germany, 2014.
11. Glidenumber, [Online]. Available: http://glidenumber.net/glide/public/search/search.jsp. [Accessed 5th November 2014].
12. Washington Post, [Online]. Available: The main advantage is the independence in terms of language understanding. Advanced analysis might also take term frequency and co-occurrence as well as hashtags and network relations into account. We discovered the stability of the public twitter data de. [Accessed 5th November 2014].
13. F. Morstatter, J. Pfeffer, H. Liu and K. M. Carley, "Is the Sample Good Enough? Comparing Data from Twitter's Streaming API and Twitter's Firehose," in *Proceedings of ICWSM*, 2013.
14. H. Abdelhaq, C. Sengstock and M. Gertz, "EvenTweet: Online Localized Event Detection from Twitter," in *Proceedings of the VLDB Endowment*, Riva del Garda, Trento, Italy, 2013.
15. T. Kraft, D. Wang, J. Delawder, W. Dou, Y. Li and W. Ribarsky, "Less After-the-Fact: Investigative visual analysis of events from streaming twitter," in *IEEE Symposium on Large-Scale Data Analysis and Visualization (LDAV)*, 2013.
16. S. Schaust, M. Walther and M. Kaisser, "Avalanche: Prepare, Manage, and Understand Crisis Situations Using Social Media Analytics," in *Proceedings of the 10th International ISCRAM Conference*, Baden-Baden, 2013.