Discovering Events from Social Media for Emergency Planning

Momna Naseem
Department of Computer Science
Lahore University of Management Sciences (LUMS), Lahore, Pakistan,
15030038@lums.edu.pk

Basit Shafiq
Department of Computer Science
Lahore University of Management Sciences (LUMS), Lahore, Pakistan,
basit@lums.edu.pk

Soon A. Chun
Information Systems and Informatics Program
City University of New York, New York, NY, USA
soon.chun@csi.cuny.edu

Shafay Shamail
Department of Computer Science
Lahore University of Management Sciences (LUMS), Lahore, Pakistan,
sshamail@lums.edu.pk

Nabil R. Adam
Department of Management Science and Information Systems
Rutgers University, Newark, NJ, USA,
adam@adam.rutgers.edu

ABSTRACT

Social media is a popular platform for daily communication. It is the fastest medium to get real-time information about any event. Event identification and finding relations between them is important for information retrieval, which can be useful in many situations. For example, in case of disaster management this information can be helpful in better planning of response operations for future events. However, discovering the important events from a social media data is a challenging task due to the sheer volume of data. In this paper, we present an automated approach for discovering events and their relationships from Twitter feeds. Our proposed approach uses a two-level clustering approach. The first level clustering identifies major events among diverse tweets, and the second level clustering identifies sub-events of a given major event by considering their spatio-temporal and semantic relationships. We evaluate our approach on a dataset taken from twitter. Results show that the two level clustering could discover major events and associated sub-events with reasonable accuracy. We also discuss the implications of the automated approach of event discovery in emergency planning and emergency response evaluation.

dg.o 2019, June 18, 2019, Dubai, United Arab Emirates

ACM ISBN 978-1-4503-7204-6/19/06...$15.00
https://doi.org/10.1145/3325112.3325213

CCS CONCEPTS

• Human-centered computing • Collaborative and social computing • Collaborative and social computing theory, concepts and paradigms • Social media

KEYWORDS

Events clustering, sub-events extraction, spatial and temporal analysis, hierarchical clustering, emergency management plan

ACM Reference format:

Momna Naseem, Basit Shafiq, Soon A. Chun, Shafay Shamail and Nabil R. Adam. 2019. Discovering Events from Social Media for Emergency Planning. In *Proceedings of dg.o 2019: 20th Annual International Conference on Digital Government Research (dg.o 2019), June 18, 2019, Dubai, United Arab Emirates.* ACM, New York, NY, USA, 8 pages. https://doi.org/10.1145/3325112.3325213

1 Introduction

Social media is a fastest medium to get information. In the context of emergency management, quick and real-time information about disaster can be collected from people who are present at the spot through social media. We can use people as participatory sensors because they are the direct source of information, such as the extent of damage, need of the community, evolution of the event. This people generated data through text messages, calls, and social media enable responders to deal with current situation. Information from emergency management community can be combined with participatory sensing capability to get real time, quick and accurate situational awareness that can be helpful for better resource allocation and informed decision making leading to a better response to the emergency situation [2, 15, 16, 17, 18]. For example, in case of a disaster event like a major snow storm, information extracted from social media about related events such as dangerous icy or snowy road conditions, fallen trees, power outage, etc., can be useful for disaster management. The events extracted from social media can be used for early warnings in case of disaster

event anticipation, so that the emergency mangers could issue timely alerts to people. The events from social media can provide situation awareness throughout the disaster event enabling responders to allocate resources for a quick response or assess the resource requirement. The analysis of events in social media can also help in discovering the needs and problems faced by individuals and communities in different localities. For example, emergency responders can identify the specific neighborhoods and the streets/roads in such neighborhoods which were severely impacted and isolated due to snow storm and how such road/street conditions may affect individuals/communities in those neighborhoods (e.g., families may run out of food/medical supplies, children getting stuck in school, etc.). The post-event analysis of social media content could provide valuable information on the damage-related impacts of the devastation, or genuine feedback on emergency responses and management.

Emergency and disaster plan needs to be adaptive to provide a coordinated and co-operative response with capabilities of identifying urgent needs and allocating available resources. The emergency plans can be drafted in advance, but these should be adjusted depending on the localities, severity levels, and types of emergency [18]. Quick identification of important events and associated sub-events to detect needs and problems of the impacted community is essential for emergency response plan adjustment.

In this paper, we take up the research challenge to discover major events (so called event types or categories) that affect a given region and to identify their sub-events by considering their spatio-temporal and semantic relationships within a major event type by leveraging large social media dataset. Event type identification and finding sub-events help in understanding the types of events that are associated with a particular disaster as well as the relationships among co-occurring events. Particularly, we present a Machine Learning technique of hierarchical clustering to extract high quality event categories and to further identify the sub-events related to the major event categories/types from Twitter feeds.

In our proposed approach, we use two step clustering for major and sub-event extractions. First, we extract important keywords from tweets and identify important event categories through agglomerative clustering to find high quality events. Second, we identify sub-events associated with a major event category using their spatio-temporal and semantics relationships. In order to find temporal relationship, we consider the time stamps of tweets with reference to the time period identified for a given event. For spatial and semantic relationships, we consider location references as well as the named entities in the tweets text. We evaluate our approach on approximately 1.7M tweets related to several events, such as blizzard which swept across North America in early 2015, Nepal earthquake, Super Bowl 2015, Indian Premier League 2015, etc. The two level clustering approach is applied to the dataset. The first clustering step identified major events (event categories), e.g. blizzard, earthquake, football events, etc. Furthermore, the second step clustering results demonstrate that temporal, spatial and semantic relationships of tweets from two major event categories (Blizzard 2015 and Nepal's earthquake) could identify many important sub-events within each of these major event types.

The remaining paper is organized as follows. Section 2 discusses related work followed by the presentation of the two-level clustering approach in Section 3. Section 4 presents the data analysis of major event extraction and the associated sub-events, and discuss its results in Section 5. In Section 6 we presents the paper summary and future work.

2 Related Work

Atefeh and Khreich [13] present a detailed survey of event detection techniques in twitter streams. These techniques vary depending on event type, event detection method and features. Based on the target application and detection task, techniques may vary, such as retrospective event detection which uses archived data set or new event detection which uses live data feeds for detecting ongoing events in real time. The detection methods can be categorized as unsupervised, supervised, hybrid approaches. In [1, 3], Ritter et al. proposed an approach for extraction and categorization of significant events from twitter data. Their proposed approach is based on latent variable models. It first discovers event types that match the data. These event types are then used to classify aggregate events without any annotated examples. Benson et al. in [4] develop a supervised learning based model for record extraction from social media streams. This model uses a seed set of example records corresponding to different events. It uses these examples to analyze individual twitter messages and cluster them according to the event. Xing et al. in [5] present an approach for hash-tag based discovery of sub-event in Twitter data. The hash tags are considered as representative of sub-events in their approach. However, tweets referring to a particular event/sub-event may not include the appropriate hashtag and therefore such event/sub-event may be missed. In contrast, our approach detects events through tweet text and considers spatio-temporal and sematic relationship of tweets for events detection.

In [7], Abdelhaq et al. presented a framework named "Event Tweet" for detection of localized events from tweets in real-time. Localized events are those events which occur in small geographical region. For example: traffic jams, festivals and matches etc. Localized events are detected on the basis of two features: one is related keywords and the other is start time and geographical location of the localized event. For temporal resolution an event scoring scheme is proposed that provide specific score for each event to indicate its significance according to time. However, only geo-tagged tweets are used for spatial resolution. In our work, we consider major as well as sub-events and we consider location references provided in tweet text because majority of the tweets do not have geo-location encoding.

Sakaki et al. in [11] proposed a spatio temporal probabilistic model for event detection for an earthquake reporting system that detects earthquake and notify registered users about earthquake via email. Their main focus is on finding the center

and trajectory of the location of event. Popescu et al. in [12] detect events, their related entities and audience reactions on these events, from tweets. Two learning models are used; "Event Basic" which is supervised learning method that represents event snapshots; and "Event Aboutness" that augments the feature set of "Event Basic". Features contain different combinations of mean and standard deviations. They did not considered sub-events and spatio-temporal relationships of events.

Imran et al. [14] provide a comprehensive survey of methods for retrieval and processing of information related to mass emergency events from social media feeds. The objective is to help researchers and developers to develop tools that can be used by humanitarian organizations and response agencies to receive and disseminate incident-related information from social media feeds. Sakaki et al. in [10] present an approach for extracting driving information in real-time from social media feeds using text-based classification methods. The location information is extracted from tweets using natural language processing or if the tweets are geo-tagged, they use the embedded geo-coordinates. Google MAPS API is used to convert the geo-coordinates into location names. Song et al. [6] proposed a model for event-based spatio-temporal topic extraction from twitter data. In [8], a sketch-based model named "Topic sketch" is proposed for detection of bursty topics in real time. Zhao et al. in [9] studied the problem of topical key phrase extraction for summarization and analysis of twitter data by the extracting and organizing key phrases related to topics on twitter.

3 Methodology

The proposed methodology for discovering events and their sub-event is a 3-step process as depicted in Figure 1. We discuss these steps in detail in the following subsections.

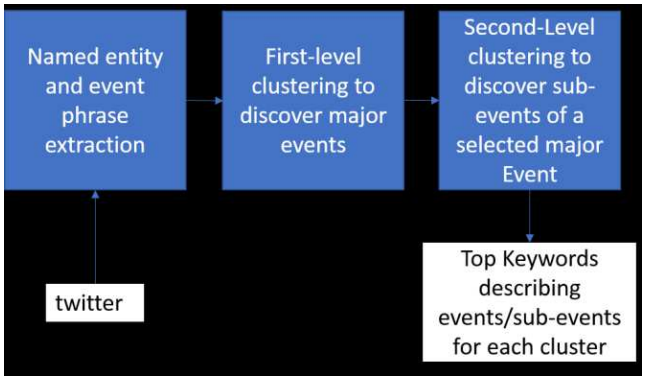

Figure 1: Proposed methodology for discovering events and their sub-events.

3.1 Discovering Major Events

We represent each tweet as a vector of keywords. These keywords are generated by performing named entity recognition and event phrase extraction. Named-entity recognition classifies the elements in twitter text into predefined categories such as the names of persons, organizations, locations. Event phrases may consists of verbs, nouns, or adjectives in the twitter text that refers to some event as illustrated below:

Verb: 13 persons were **evacuated** frm landslide zone of #Barpak 2 #Pokhara,y'day.

Noun: Scenes of **devastation** still visible in Chautara.

Adjective: In spite of life **threatening** damages due to #NepalEarthquake 278 inmates of Ramechhap jail stayed back, helped, and cleaned the debris.

For named entity recognition and event phrase extraction, we use the open domain event extraction approach by Ritter et al. [1]. Each tweet is represented as a vector of n points, where each point denotes the frequency of occurrence of specific keyword in particular tweet. Next we apply clustering on the resulting matrix consisting of the keyword vectors of all the tweets in the dataset. The objective is to group tweets that refer to a common event into a single cluster. We use agglomerative clustering for this purpose. Agglomerative clustering is a hierarchical clustering algorithm that takes distance matrix as input and divides the entire dataset according to different level of heights on the basis of their distances. We consider the height level in which the number of elements (i.e., tweets) in each cluster is greater than a minimum threshold number, e.g., 500. It is possible that a cluster may include multiple events because of the overlapping keywords characterizing those events. However, the distribution of events in a cluster is highly skewed – majority of the tweets in a cluster refer to the same event. In addition, tweets referring to the same event or related events may fall in different clusters. Each cluster is characterized by the top k (e.g., k=5) keywords.

Given the large volume of tweets on which clustering was applied, the resulting clusters partition the event space at a coarser granularity, e.g., Blizzard of 2015 in Northeastern states of the US, Nepal earthquake, Football World Cup, etc. We are interested in identifying the events at the finest granularity, which may correspond to sub-events of some major event (e.g., streets/roads that are closed in specific localities due to a snow storm, individuals request for assistance, etc.) as well as relationship between those events (e.g., power outage in a given locality due to snow storm and people requesting for assistance to prepare food or requesting backup generators to heat their houses, etc.). The tweets referring to these specific sub-events are lost in the major event clusters due to their low frequency count.

3.2 Discovering Sub-events of a Major Event

To discover interesting sub-events and their relationships in the context of a given major event (e.g., Blizzard of 2015), we select the cluster that represents the given major event. For the selected cluster, we compute the event timeline as well as location references from the underlying tweets. Next, we scan other clusters and select all those tweets that have a temporal, spatial, or semantic overlap with the selected cluster. For temporal overlap, we consider all the tweets that were posted within the event timeline of the given cluster and have at least 1-2 keywords matching with the top k keywords of major event

clusters. For spatial overlap, we consider all the tweets that refers to a location included in a set of locations associated with the major event cluster. We consider a location ontology (e.g. GeoNames) to resolve the geo-spatial references. For semantic overlap, we consider all the tweets that include some minimum number of keywords from the top k keywords of the selected major event cluster.

This filtered dataset that includes all the tweets from the selected major event cluster as well as tweets from other clusters that have a spatial, temporal, and semantic overlap with the given major event cluster is then used for extracting sub-events. For this purpose we run a second level clustering on this filtered dataset to group events at finer granularity. Again, we used agglomerative clustering for this step and choose a height level in which in which the number of elements (i.e., tweets) in each cluster is greater than a minimum threshold. This minimum threshold for sub-event identification is much smaller than the threshold considered for major events.

The clusters at the selected height level represent sub-events of a given major event. Based on the event timelines, location references, and characterizing keywords of these clusters, we can determine the spatial, temporal, and semantic relationships between the underlying events.

Following are the specific steps taken to discover sub-events and their relationships for a given major event cluster:

1. Identify the top k keywords related to the given event type.
2. Calculate time range of tweets in each cluster corresponding to the major event.
3. Extract all the location references from the given major event cluster(s).
4. Select all tweets from other clusters that have a temporal overlap (i.e., time-stamp of the tweet falls within the major-event timeline and have at least one keyword matching), spatial overlap (i.e., location word in the tweet text is related to the set of locations associated with the major event), or semantic overlap (i.e., at least m keywords matching) with the given major event cluster(s)
5. Perform second-level agglomerative clustering to generate clusters of sub-events. Characterize each cluster/sub-event by top k keywords.

4 Experiments and Results

For experimental evaluation of the proposed approach, we used Twitter dataset consisting of 1,726,559 tweets collected from January 2015 to July 2015. This dataset includes tweets related to several events, such as blizzard that swept across northeastern states of the US in early 2015, Nepal earthquake, Super Bowl 2015, Indian Premier League 2015.

4.1 Major Events Clusters

Table 1: Partial list of major events from clusters of tweets

Cluster Size	Top Key Words	Frequencies	Event
42420	blizzardof2015, traffic , rescue, snow, power, road, juno, from, home, time	16204, 13925, 5546, 4651, 3678, 3187, 2949, 2615, 2437, 2129	blizzardof2015 – traffic jams
1047	power, football, video, from, news, energy, ranger, death, rangers, sudan	1033, 892, 335, 332, 326, 256, 130, 52, 34, 33	Football
4798	boston, blizzardof2015, whoshoveledthefinishline, help, bostonstrong, police, snow, marathon, during, from	9641, 4400, 1863, 1804, 1734, 1689, 1595, 953, 917, 681	Boston – Shoveling of Snow from Boston Marathon Finish line
10087	traffic, free, from, drive, your, road, website, lane, bridge	11454, 665, 453, 416, 353, 324, 324, 299, 232,	Traffic
4030	school, life, sleep, your, power, high, time, star, home, love	3670, 2260, 2018, 300, 274, 217, 208, 96, 88, 82	School Life
40261	rescue, traffic, university, Nepal, school, power, relief, road	24820, 2644, 2303, 2288, 1921, 1462, 1453, 1275	Rescue
29388	time, school, times, power, Nepal, road, people, university, your, high,	29437, 13887, 10786, 5290, 3358, 3271, 1975, 1614, 1163, 1127	Delayed school opening
3264	bowl, super, superbowl, road, traffic, rescue, power, from, superbowlxlix, time	3353, 3097, 1547, 1354, 961, 355, 312, 252, 224, 201,	Super Bowl
36353	airport, incheon, from, michael, review, lane, changsha, super, time, your	36813, 4296, 3157, 2079, 1921, 1522, 1434, 1284, 1185, 905,	Airports

35634	power, ranger, your, news, star, supply, people, deals, high, road	35183, 2431, 1666, 980, 809, 796, 790, 765, 703, 664	Power Ranger actor kill his roommate
53027	Nepal, help, news, quake, rescue, donate, relief, from, people, foreign	51747, 10375, 3989, 3947, 3876, 3659, 3634, 2994, 1972, 1868	Nepal Earth Quake
14039	katy, perry, power, ranger, rangers, from, world, road, time, rainbow	14019, 13447, 13098, 11548, 3694, 2726, 902, 871, 627, 451	katyperry performance in super bowl
3246	girls, rescue, haram, boko, nigeria, school, nigerian, from, news, death	4275, 3870, 3267, 3262, 2105, 2085, 1621, 508, 306, 239	Girls who escaped from boko haram rescued
11569	week, school, national, counseling, happy, high, from, star, time, nscw15	14726, 11994, 1199, 1199, 768, 345, 249, 236, 233, 206	Student week at Schools

We managed to extract more than 37,000 keywords from this dataset. We found that the distribution of these keywords was skewed. We filtered out all those keywords that appeared in less than 100 tweets. After filtering we were left with top 600 keywords. After applying the first-level clustering, we got 508 clusters at height level 5. Each cluster at this height level included over 500 tweets. Table 1 shows some of these clusters with their cluster size (i.e. number of tweets), top keywords, frequencies of top keywords, and the corresponding major events referenced in the tweets.

4.2 Sub-events Clusters

We selected two major events from Table 1 clusters to identify sub-events and their relationships: "Nepal earthquake" and "Blizzard of 2015".

4.2.1 Sub-events related to Nepal Earthquake. We consider all those major event clusters (42 clusters) that include both Nepal and Earthquake as in their top 5 keywords. We took a union of the top 5 keywords of all these clusters, which resulted in 109 distinct keywords. We selected top 16 keywords out of the 109 keywords. These top 16 keywords are:

"*Nepal, Rescue, Quake, Earthquake, Nepalearthquake, death, donate, relief, school, Nepalquake, msghelpearthquakevictims, airport, Kathmandu, hospital, medical, prayforNepal*"

We scan all the remaining clusters (466 clusters) to find all those tweets that have a spatial, temporal, and semantic overlap with Nepal Earthquake clusters by following the steps listed in Section 3.2. Figure 2 shows the timeline of Nepal Earthquake major event, which is computed by considering the timestamp of all the tweets in the corresponding clusters. This timeline figure shows that the peak time interval related to Nepal Earthquake tweets is from April 25, 2015 to May 5, 2015. We consider this interval for temporal overlap.

We also extracted the location keywords from all the tweets belonging the Nepal Earthquake clusters. We found 91 location entities. We used the GeoNames Web service to filter out those locations that were not related to Nepal. Out of these 91 locations, we only found the following 6 location keywords that are in Nepal.

"*Nepal, Kathmandu, Katmandu, Gorkha, Chautara, Everest, Kavre*"

For the sub-event analysis, we consider a total of 229,377 tweets, with 215,125 tweets coming from 42 clusters that include both Nepal and Earthquake as in their top 5 keywords. The remaining 19,801 tweets were selected from other clusters that meet one of the following conditions:

- posted between April 25, 2015 and May 5, 2015 and include at least 1 keyword from the set of top 16 keywords listed above.
- included a reference to 1 of the 6 location keywords listed above.
- included at least 2 keywords from the set of top 16 keywords listed above.

We found 95 clusters corresponding to different sub-events related to Nepal Earthquake. Some examples of these sub-events are shown in Table 2. Out of these 95 sub-events, 16 sub-events were found from the tweets that were not part of the major event clusters as shown in Table 3. These sub-events could not have been discovered if we only used tweets from the major event clusters for second-level clustering. Therefore, it is important to consider tweets in other first-level clusters that are not associated with the major event.

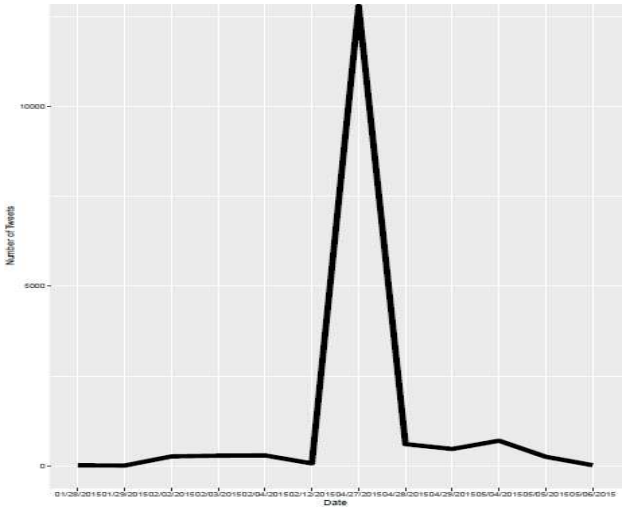

Figure 2: Timeline of Nepal Earthquake major event extracted from the related major-event clusters

Table 2: Examples of sub-events related to Nepal Earthquake

Sub-events			
RedCross help Nepal, Relief work in Nepal	WorldVision relief supplies	MusicforReliefproviding immediate medical care	Nepal quake victims flee to tents
Indian relief assistance to Nepal	People in US, Fedex are helping	Donate To UNICEF Nepal Appeal	Mount climbs 'almost impossible'
Help by firefighters	Utah's Tibetan community prays	Deaths in Kathmandu	South Jersey Woman Shares Story Of Surviving
International rescue efforts	legendary Gurkhas join quake rescue	Americans at Nepal, Relief efforts	Mobile field hospital
MeetRescueDogs	French 'Spiderman' scales Paris tower for Nepal	Firefighters from West Midlands	CatholicRelief to aid

Table 3: Sub-events related to "Nepal Earth Quake" from First-level clusters that do not include Nepal and earthquake as top 5 keywords

Cluster Size	Top Keywords	Sub-events
301	music, Nepal, relief, media, medical, quake, state, young, help, park, everest, mount	"MusicforRelief" providing immediate medical care
190	rescue, fire, Nepal, search, L.A. County, airport, news, breaking	Help by firefighters
256	Pakistan, Nepal, hospital, fort, army, relief, quake, earthquake, medical	Pakistan sent relief material for victims
24	Chicago, quake, earthquake, Nepal, home ,travel	Help from Chicago
49	fire, Nepal, quake ,earthquake, west, free, l.a., help, service, news, people	Firefighters from West Midlands
20	Irish, time, Nepal, relief, world, earthquake, Clifford	Irish relief workers
151	tech, Nepal, technology, medical, team, quake, fort, relief, rescue	Unique Medical Technology in Nepal
26	radio, earthquake, quake, Nepal, news, south, traffic, emergency, Nepalearthquake	Radio Assistance
160	Nepal, Nepalese, earthquake, news, community, south, bank	Nepal quake victims flee to tents
224	Bihar, quake, earthquake, hospital, India, Indian, mark, Nepal, Hindu	Hospital in Bihar
31	medical, rescue, during, charlotte, time	Nurse Dies in Fall From Helicopter
60	India, power, Nepal, kathmandu, city, electric, valley, IndiawithNepal, university	thanks Power Grid of India
13	Pakistan, Nepal, Pakistani, rescue, relief	Pakistanis rescued from Nepal
9	French, Nepal, quake, Paris, India, business	French 'Spiderman'scalesParis tower for Nepal
23	mount, Nepal, quake, earthquake, everest, singapore, Nepalquake,north	Mount climbs 'almost impossible'
14	Denver, Nepal, arts, relief, star ,Colorado	Denver-based nonprofit starts disaster relief fund

4.2.2 Sub-events related to Blizzard of 2015. We consider 6 clusters corresponding to the Blizzard of 2015 major event. The following top 10 keywords were used to select tweets from the remaining 502 clusters.

"*Blizzardof2015, Snow, York, Boston, Juno, Traffic, Rescue, Power, School, Home*"

The peak time interval related to Blizzard of 2015 tweet posting was [January 28, 2015, February 2, 2015]. The following location keywords were used for determining spatial relationship:

"*New England, Manchester, Worcester, Boston, London, Chicago, England*"

For the sub-event analysis, we consider a total of 225,408 tweets, with 232,934 tweets coming for 6 clusters that include blizzard, snow, Boston, or Juno in their top 5 keywords. The remaining 7,526 tweets were selected from other clusters that meet spatial, temporal, and semantic overlap criteria discussed above. We found 63 clusters corresponding to different sub-events related

to Blizzard of 2015. Some examples of these sub-events are shown in Table 4.

Table 4: Example of sub-events related to Blizzard of 2015

Sub-events			
blizzard at boston	BOSTON_EMS stuck during blizzard	whoshoveledthefinishline	'Power Ranger' actor arrested
snow in New York	Crazy couple crossing Africa in a camper van	rescue animals	Rescue the needy
Bill de Blasio NYC Mayor talk	traffic near the University of Phoenix	Whiteout at Central Park	Obama said to rescue women
help from USNationalGuard	West Bloomfield School District will be closed	seacoast dig out	Huron High School is closed

5 Discussion

In the context of emergency management, the proposed approach related to discovering major events and their related sub-events from social media data is useful in many ways, specifically for better response planning and training purposes. Post-incident analysis from social media data can help in identifying important events that may arise during an emergency which have not been considered in the default emergency response plans. These default response plans are generally designed for common events related to the given disaster incident. By considering the cause and effect relationship between events that happened earlier in some past incidents, emergency responders can infer how an ongoing disaster may evolve based on its similarity with the past incidents and be ready to respond to events missed in the default plans.

Moreover, the proposed work can help the emergency responders to have a better anticipation of the needs of different communities or assistance requests from citizens from varying geographical locations and for different incidents types. As discussed in [15, 16], citizens are increasingly relying on social media, mobile devices, and other online information portals to get information about ongoing disasters, seek assistance and safety information, and report their safety and well-being during or after emergencies.

Another related aspect where the proposed work can be utilized is in getting public feedback about the effectiveness and timeliness of response activities and how these could be improved. Citizens' suggestions, criticism, or complaints through social media can provide useful insights to how the incident was handled and how the response activities can cater to citizens' concerns and needs [2].

An important issue that needs to be addressed in the context of event extraction is to assess the veracity of information retrieved from social media. Recently, there has been some work done on detecting fake news or rumors spread through social media using both supervised and unsupervised learning methods as well as using crowd signals [19, 20]. These work can be applied to our approach to filter out the "unreliable" datasets to improve the quality of results.

6 Conclusion and Future Work

In this paper, we propose an approach for high quality event extraction and sub-event identification from Twitter feeds. Our proposed approach uses the hierarchical clustering machine learning technique to identify major event types from large tweet datasets and employ spatio-temporal and semantic relationship information to identify related sub-events. We used Parts of Speech (POS) tagger and Named Entity Recognition (NER) system to get named entities through tweets text. We proposed two level hierarchical agglomerative clustering machine learning techniques for identifying major event clusters. To extract related sub-events of a major event of interest, we select tweets from clusters directly related to that given event. To improve sub-events quality we add tweets from remaining clusters on basis of three filters which are: keywords filter, time filter, and location filter.

We tested our approach on a large data set consisting of more than 1.7 million tweets to identify major events first and then to identify sub-events from two natural disaster events: "Nepal earth quake 2015" and "blizzard of 2015". Results show that our approach found major events included in the tweet dataset through the clustering. It also identified sub-events associated with two major disaster events. We found many precise sub-events related to these two major events with reasonable accuracy.

The effective method in automated event identification as presented here will be of a great value for the disaster planning by understanding many sub-events related to major disaster events. We plan to improve our approach by incorporating the methods to address the data quality issues such as fake tweets. In the future, we plan to develop approaches for real-time extraction of these events/sub-events from live Twitter feeds. In addition, we plan to conduct a case study on how the disaster planning process can incorporate the presented event discoveries for dynamic adjustments.

7 ACKNOWLEDGMENTS

The work of Shafiq is supported by the Pakistan Higher Education Commission (HEC) NRPU Grant P#2305. In addition,

the project is partially funded by NSF CNS 1624503 and CNS 1747728, and partially by the National Research Foundation of Korea Grant from the Korean Government (NRF-2017S1A3A2066084).

REFERENCES

[1] Ritter, Alan, Oren Etzioni, and Sam Clark. "Open domain event extraction from twitter." Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining. ACM, 2012.
[2] Yuan, Qianli, and Mila Gasco. "Citizens' Use of Microblogging During Emergency: A Case Study on Water Contamination in Shanghai." In Proceedings of the 18th Annual International Conference on Digital Government Research, pp. 110-119. ACM, 2017
[3] Ritter, Alan, Sam Clark, and Oren Etzioni. "Named entity recognition in tweets: an experimental study." Proceedings of the Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics, 2011.
[4] Edward Benson, Aria Haghighi, and Regina Barzilay. "Event Discovery in Social Media Feeds" HLT '11 Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies, 2011.
[5] Xing C, Wang Y, Liu J, Huang Y, Ma WY. "Hashtag-based sub-event discovery using mutually generative LDA in Twitter". Thirtieth AAAI Conference on Artificial Intelligence 2016 Mar 5.
[6] Song S, Li Q, Bao H. "Detecting dynamic association among Twitter topics". Proceedings of the 21st International Conference on World Wide Web. ACM, 16 Apr 2012.
[7] Abdelhaq H, Sengstock C, Gertz M. "Eventweet: Online localized event detection from twitter." Proceedings of the VLDB Endowment. 2013 Aug 28;6(12):1326-9.(precise events)
[8] Xie W, Zhu F, Jiang J, Lim EP, Wang K. "Topic sketch: Real-time bursty topic detection from twitter." IEEE Transactions on Knowledge and Data Engineering. 2016 Aug 1;28(8):2216-29.
[9] Zhao WX, Jiang J, He J, Song Y, Achananuparp P, Lim EP, Li X. "Topical key phrase extraction from twitter." Proceedings of the 49th annual meeting of the association for computational linguistics: Human language technologies-volume 1 2011 Jun 19 (pp. 379-388). Association for Computational Linguistics.
[10] Sakaki T, Matsuo Y, Yanagihara T, Chandrasiri NP, Nawa K. "Real-time event extraction for driving information from social sensors." Cyber Technology in Automation, Control, and Intelligent Systems (CYBER), 2012 IEEE International Conference on 2012 May 27 (pp. 221-226). IEEE.
[11] Sakaki T, Okazaki M, Matsuo Y. "Earthquake shakes Twitter users: real-time event detection by social sensors." Proceedings of the 19th international conference on World Wide Web 2010 Apr 26 (pp. 851-860). ACM.
[12] Popescu AM, Pennacchiotti M, Paranjpe D. "Extracting events and event descriptions from twitter." Proceedings of the 20th international conference companion on World Wide Web 2011 Mar 28 (pp. 105-106). ACM
[13] Atefeh F, Khreich W. "A survey of techniques for event detection in twitter." Computational Intelligence. 2015 Feb 1;31(1):132-64.
[14] Imran M, Castillo C, Diaz F, Vieweg S." Processing social media messages in mass emergency: A survey." ACM Computing Surveys (CSUR). 2015 Jul 21;47(4):67.
[15] Nabil Adam, Jayan Eledath, Sharad Mehrotra, Nalini. "Social Media Alert and Response to Threats to Citizens (SMART-C)". 8th International Conference on Collaborative Computing: Networking, Applications and Work sharing ,2012
[16] Nabil R. Adam, Basit Shafiq, Robin Staffin, "Spatial Computing and Social Media in the Context of Disaster Management". IEEE Intelligent Systems, Vol 27, No. 6, November 2012, pp. 90-96.
[17] Lorenzi, D., S. Chun, J. Vaidya, B. Shafiq, V. Atluri and N. Adam. PEER: A Framework for Public Engagement in Emergency Response, International Journal of E-Planning Research, Vol4(3), 2015: 29-46.
[18] Shafiq, B., S. Chun, V. Atluri, J. Vaidya, and G. Nabi. Resource Sharing using UICDS Framework for Incident Management, Transforming Government: People, Process and Policy (TGPPP), Vol 6(1), 2012: pp41-61.
[19] Yang, S., Shu, K., Wang, S., Gu, R., Wu, F. and Liu, H., 2019. "Unsupervised fake news detection on social media: A generative approach." Proceedings of 33rd AAAI Conference on Artificial Intelligence.
[20] Tschiatschek, S., Singla, A., Gomez Rodriguez, M., Merchant, A. and Krause, A. "Fake news detection in social networks via crowd signals." In Companion Proceedings of The Web Conference 2018 (WWW '18) (pp. 517-524).