**Note:**

The text in black is reviewers’ comments, blue text is authors’ response to reviewers, and highlighted text is the changes in manuscript and is also copied here to give reviewers easy access to the changes.

**Reviewer 3**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Yes | Can be improved | Must be improved | Not applicable |
| Does the introduction provide sufficient background and include all relevant references? | ( ) | (x) | ( ) | ( ) |
| Is the research design appropriate? | ( ) | (x) | ( ) | ( ) |
| Are the methods adequately described? | ( ) | (x) | ( ) | ( ) |
| Are the results clearly presented? | ( ) | (x) | ( ) | ( ) |
| Are the conclusions supported by the results? | ( ) | (x) | ( ) | ( ) |

Comments and Suggestions for Authors

This paper presents a study on the verification of the reliability of social media in the context of natural disasters (in particular, floods). The study took advantage from both manual and automatic analysis. The results of the study highlight some problems related to the coverage of geolocated tweets and the reliability of some automated image analysis tools. The topic of the paper is interesting and the problem of filtering social media streams to extract information of value for first responders is still an open issue. I think that the the paper has some value to researchers interested in designing automatic or semi-automatic systems that analyze social media streams searching for disaster-related items. However, there are some small issues that the paper should address.

First of all, it is important to formalize the steps that were taken during the manual analysis of tweet contents:

Thank you for the great suggestion. We added a paragraph describing the steps (copied below).

**(Line 165-175)** The text analysis involves a few consecutive steps. The first and foremost step is to search for evidence information from reference documents, especially weather warning/alert messages from the National Weather Services and the state and local emergency management agencies. Events, names of damaged roads, streets, and the posted time of each tweet were manually identified (De Choudhury, Diakopoulos et al. 2012) from relevant tweets and then used as keywords to search for related information in reference documents. If no such information can be found, the topic, posted time, location of tweets can be documented for further use. The next step is to holistically assess if the documented unverified tweets have any association with other tweets based on topic, posted time, or location. Finally, news information may also be a complementary reference source if available. If evidence can be found from reference documents or enough tweets from multiple Twitter users presented facts that fit the hazard context in the relevant tweets, the studied tweets can be considered reliable.

When is it necessary to search the information in news (by the way, this may be affected that news can appear later than in the tweet stream)?

Thanks for the question. Yes, the usefulness of news information depends on the availability of news. So, it should be used as a complementary source but not the primary reference source.

(**Line 172-175)** Finally, news information may also be a complementary reference source if available. If evidence can be found from reference documents or enough tweets from multiple Twitter users presented facts that fit the hazard context in the relevant tweets, the studied tweets can be considered reliable.

How to search the information?

Thanks for the question. Given the mixed nature of our reference sources, e.g., National Weather Services, state & local emergency management agencies, we primarily relied on Google search to seek out official reference sources, which were then used in the analysis.

Looking for the exact content of the tweet plus the position?

Searching for exact content of tweet plus the position generally returns nothing. Furthermore, different combination of words is used in different reference documents, which also posed the possibility of missing useful information. Hence, we used a combination of keywords reflecting either events, location, or time to search for related information from reference documents.

What are the sources that one could use for news and how to select them (for instance, if the event is occurring in Colorado, how to find all local Colorado news sources?)

The major official sources are the agencies involved in emergency management activities during a flood event, which included the National Weather Services, Colorado Division of Homeland Security & Emergency Management, Boulder County Emergency Management office.

With regard to the geolocation issue, that is, very few tweets are geographically tagged compared to the overall number of relevant tweets, it is a known problem: https://codete.com/blog/observing-world-tweeting-tendencies-in-real-time-part-2/ The numbers provided by the authors seem in line with most observations. However, it seems to me that given the selection process (l.108) potential relevant tweets were discarded although they contained geographical information under the form of toponyms.

It would be useful to indicate how many tweets with place names were discarded over the total, and eventually to look for some examples where they would have been useful for the analysis.

Thank you so much for this suggestion. We used those keywords (location names /incidents) to extract more non-geotagged tweets and assessed their reliability. As you suggested, this method adds another 2472 reliable tweets and 752 images. We added this section in the manuscript and copied below:

(Line 313-328) 4.3. Extracting extra tweets using verified keywords

Section 4.1 and 4.2 identified 584 reliable tweets and 60 reliable images, which accounts for 11% and 1% of 5202 geo-tagged tweets, respectively, and 0.05% and 0.01% of all 1,195,183 purchased tweets, respectively. To make better use of this data source, we selected a group of keywords/locations (e.g., Highway 36) from the verified reliable tweets discussed in section 4.1 and 4.2 and used the keywords to extract more tweets that do not possess any geo-location information. We believe doing this would yield a larger volume of relevant tweets that were discarded due to lack of geoinformation. Without geolocation, it is possible that those tweets may be sent from outside the study area, but the time frame (September 9th to 18th, 2013) and keywords (a. location names: Colorado, Boulder, etc., and b. hazard event/impacts: flooding, rain etc.) used to download those tweets from Twitter database significantly decreased this possibility.

The keywords we put together were mostly from Table 1 and they include West of Broadway, Broadway, Arapahoe Ave, Marine St, 28th St, Colorado Ave, Boulder Creek, Highway 36/US-36, and Skunk Creek. Using these keywords, we found 2472 non-repetitive relevant and reliable tweets and 752 reliable images, which account for 0.2% and 0.06% of all 1,195,183 raw tweets, respectively. This is a big improvement than using geo-tagged tweets alone for this research workflow.

About the tweet selection, it is not clear to me if (l.107-108) the tweets had to contain both a place name and a keyword or just one of them.

Thank you for the comment. We added an explanation about the criteria used to extract tweets (copied below).

**(Line 127-128)** Any tweet that contain either the location name or hazard event/impact was included. Any tweet that contained either the location name or hazard event/impact was included in the analysis.

l.148 how many tweets are necessary to confirm an information?

Thanks for the question. It is difficult to define a set number to confirm a piece of information and this really depends on the content of the tweets. For instance, if there is direct evidence of high flood risk for a flooded region, Boulder Creek, can be found from NOAA weather warning messages, one tweet can be considered reliable. If no such evidence can be directly found anywhere, two or more tweets will be necessary to verify a piece of information.

l.274 can you quantify in how many cases the image detection was better than web detection (or vice versa)?

Thank for this question. The web detection always performed better than the image detection approach because the web detection approach searched all available content on the web pertaining to the event as opposed to the image detection approach that only focused on image content. This explanation is included in the manuscript (copied below).

**(Line 312-313)** Among the 34 images detected by GCV API to be relevant to 2013 Colorado flood, web detection outperformed the results from image detection, one example can be found in Figure 7.

l.307 are the 11% and 19% overlapping?

Yes, most of the images identified by AI approach were also identified by manual approach.

In conclusion I suggest the authors to give some more insights to some of the findings and to outline a possible workflow for the textual analysis.

Thank you for this suggestion. We added more explanation about text analysis and results.

3.3 Analytics and techniques

(**Line 160- 164**) Specifically, the bag-of-words model was applied to geo-tagged tweets to extract assumed relevant tweets. The bag-of-words extraction used topic-specific search terms, top frequency words and high-frequency hashtags, to measure the relevance of a document (i.e., tweets) to the search terms and extract the assumed relevant documents. The relevance of these tweets was determined first following which their reliability was evaluated.

(**Line 165 -175**) The text analysis involves a few consecutive steps. The first and foremost step is to search for evidence information from reference documents, especially weather warning/alert messages from the National Weather Services and the state and local emergency management agencies. Events, names of damaged roads, streets, and the posted time of each tweet were manually identified (De Choudhury, Diakopoulos et al. 2012) from relevant tweets and then used as keywords to search for related information in reference documents. If no such information can be found, the topic, posted time, location of tweets can be documented for further use. The next step is to holistically assess if the documented unverified tweets have any association with other tweets based on topic, posted time, or location. Finally, news information may also be a complementary reference source if available. If evidence can be found from reference documents or enough tweets from multiple Twitter users presented facts that fit the hazard context in the relevant tweets, the studied tweets can be considered reliable.

We also updated section 4.1(Evaluation of text content) and 4.2 (Evaluation of image content). And we also added a new section 4.3 (Extracting more tweets using verified keywords) based on your previous comments.

**References**

Bonney, R. (1996). "'Citizen Science: A Lab Tradition' [in] Living Bird: For the Study and Conservation of Birds." Living Bird: For the Study and Conservation of Birds **15**(4): 7-15.

Chen, S.-H. and Y.-H. Chen (2017). A content-based image retrieval method based on the google cloud vision api and wordnet. Asian conference on intelligent information and database systems, Springer.

d'Andrea, C. and A. Mintz (2019). "Studying the Live Cross-Platform Circulation of Images With Computer Vision API: An Experiment Based on a Sports Media Event." International Journal of Communication **13**: 21.

Dale, S. (2015). "Heuristics and biases: The science of decision-making." Business Information Review **32**(2): 93-99.

De Choudhury, M., N. Diakopoulos and M. Naaman (2012). Unfolding the event landscape on twitter: classification and exploration of user categories. Proceedings of the ACM 2012 conference on Computer Supported Cooperative Work.

Greengard, S. (2011). "Following the crowd." Communications of the ACM **54**(2): 20-22.

Hawkins, J. B., J. S. Brownstein, G. Tuli, T. Runels, K. Broecker, E. O. Nsoesie, D. J. McIver, R. Rozenblum, A. Wright, F. T. Bourgeois and F. Greaves (2016). "Measuring patient-perceived quality of care in US hospitals using Twitter." BMJ Quality &amp; Safety **25**(6): 404-413.

Hetmank, L. (2013). "Components and Functions of Crowdsourcing Systems-A Systematic Literature Review." Wirtschaftsinformatik **4**: 2013.

Horita, F. E. A., L. C. Degrossi, L. F. G. de Assis, A. Zipf and J. P. de Albuquerque (2013). "The use of volunteered geographic information (VGI) and crowdsourcing in disaster management: a systematic literature review."

Klonner, C., S. Marx, T. Usón, J. Porto de Albuquerque and B. Höfle (2016). "Volunteered Geographic Information in Natural Hazard Analysis: A Systematic Literature Review of Current Approaches with a Focus on Preparedness and Mitigation." ISPRS International Journal of Geo-Information **5**(7): 103.

Liu, X., B. Kar, C. Zhang and D. M. Cochran (2018). "Assessing relevance of tweets for risk communication." International Journal of Digital Earth **12**(7): 781-801.

Moon, G. (2014). "Why People Share: The Psychology of Social Sharing." from <https://coschedule.com/blog/why-people-share/>.

Oh, S. and S. Y. Syn (2015). "Motivations for sharing information and social support in social media: A comparative analysis of Facebook, Twitter, Delicious, YouTube, and Flickr." Journal of the Association for Information Science and Technology **66**(10): 2045-2060.

Onsrud, H., G. Camara, J. Campbell and N. S. Chakravarthy (2004). Public commons of geographic data: Research and development challenges. International Conference on Geographic Information Science, Springer.

Onsrud, H. and J. Campbell (2007). "Big opportunities in access to" Small Science" data." Data Science Journal **6**: OD58-OD66.

Wiggins, A. and K. Crowston (2011). From conservation to crowdsourcing: A typology of citizen science. 2011 44th Hawaii international conference on system sciences, IEEE.