**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 2**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | 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

The manuscript titled "*Extracting reliable Twitter data for flood risk* *communication using manual assessment and Google* *Vision API from text and images*" its into the scope of International Journal of Geo-Information journal. It has a borderline novel idea, but nevertheless, the manuscript needs major improvement **in results and discussion**. My comments are following:

I failed to find connection between text and image analysis in the results or discussion?

Thank for raising the issue. We have addressed your concern in the discussion and is also copied here:

**(Line 338-342)** In this study we implemented text and image content analysis to extract and evaluate tweet reliability because research on using both image and text analysis was relatively rare in Twitter based flood research while majority research focus only on Twitter text. Another reason for using this approach is that information extracted from text and images is oftentimes complementary; so including both would extract more information than only using only text or images.

What was the purpose of manually determining readability of tweets? In reality it would not be possible because that would need huge effort which in crisis situation everything must be done in almost real time. If the purpose was to test Google Cloud Vision API for crisis situations, I could be done much simpler, without GIS data or NOAA or reference documents. Manually extracting all images related to that particular disaster could be done by hand and labeling them through GCV API would give you much bigger dataset for analysis.

Thanks for the question. The manual approach was used to validate and complement the automatic API recognition approach. The result from GCV API shows that the algorithm was not able to detect the same amount of information as humans do due to insufficient flood-related training images used to support the algorithm.

We have referenced studies that have indicated the advantages of manual validation by researchers or Amazon Turk workers (De Choudhury, Diakopoulos et al. 2012, Hawkins, Brownstein et al. 2016).

The concern about the impossibility to implement this workflow for a single entity or agency is legitimate and we acknowledge that. However, the literature has documented dozens of examples where interested common citizens are participating disaster response activities by voluntarily providing technical support or processed information to add to humanitarian efforts (Horita, Degrossi et al. 2013). Several citizen science efforts by FEMA and USGS are underway (discussed below), which allow citizens to be involved in emergency response activities. It is possible to implement human-in-the loop approach to integrate local knowledge for extraction of reliable and relevant information from tweets for decision-making.

**(Line 368-379)** First, given the time-consuming and expensive nature of the proposed manual approach, its implementation may require a team of specialists to dedicate huge efforts to extract relevant and reliable risk information in an emergency setting. One promising phenomenon to counterbalance this limitation, though, is the emergence of volunteered citizen scientists who involved themselves in disaster response activities by voluntarily providing technical support or processed information to facilitate humanitarian efforts in recent disasters (Horita, Degrossi et al. 2013, Klonner, Marx et al. 2016). For instance, CitizenScience.gov, citizen science efforts by the United States Geological Survey (<https://www.usgs.gov/topic/citizen-science>) and FEMA’s crowdsourcing and citizen science efforts have allowed citizens to participate during emergency management and response efforts to complement the activities underway by the decision-makers. With the involvement of these digital humanitarians, we believe that the workflow outlined in our research can partially or fully be adopted in disaster responses.

Text analysis of tweets was done manually, described in line 144. There are lot of research papers that deal with this issue. If failed to find any novelty in manual classification of tweets.

Thanks for this great question and this helps us better explain why manual classification is necessary. Although the manual method may not be novel, current AI approach still a supervised approach that requires manual involvement or human-in-the loop to extract training images and data before a learner could be implemented. The manual approach helped us identify names of damaged roads, streets, and tweet posted time **were used as keywords** to search for related information from our reference document. This is a common practice in determining search keywords (De Choudhury, Diakopoulos et al. 2012).

**(Line 167-169)** 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.

The main aim of the paper was to present novel framework to asses reliability of tweets. I failed to find where is this framework described? Using only Google Cloud Vision API is not a novel framework.

Thank for identifying the issue. We have made changes to discuss the “research workflow” implemented to assess the reliability of tweets for risk communication process. The workflow discusses the steps implemented to evaluate reliability of relevant tweets by leveraging AI as well as human-intelligence (a major component of citizen science based research).

Given that neural networks for image processing still require human involvement to generate significant training samples, the workflow discussed here in could be used to develop supervised and/or semi-supervised algorithms to process crowdsourced images to extract disaster relation information. Although we used the Google Cloud Vision API, it could be replaced with a neural network. This explanation has been added in the introduction and discussion sections (and copied below for your reference).

**(Line 102-108)** Using this definition, a workflow was developed to assess reliability of extracted risk information from relevant tweets that were obtained for the 2013 Colorado flood event. Using the workflow, we examined the tweet text and images leveraging human intelligence and Google Cloud Vision API (GCV API). The relevant tweets were extracted via several data mining techniques and can be found from previous publication(Liu, Kar et al. 2018). GCV API allowed automatic identification of image content, labeling of images, matching other online information by leveraging pre-trained machine learning models, and has been widely used by other research (Chen and Chen 2017, d'Andrea and Mintz 2019).

**(Line 356-366)** While leveraging human intelligence to validate textual content of tweets is not novel in Twitter text mining research, it brought a unique contribution to the flood research. Specifically, it allowed identification of different scenarios and process information beyond plain text (e.g., associate events in different images or associate events based on their proximity to events in the surrounding areas by pinpointing them on maps), which is impossible for current AI approaches to achieve. Given that the current neural networks (E.g., ResNet, UNet) used for disaster situations require human intelligence to collect and label significant amount of training images, our manual approach complemented the AI approach. The GCV API could be replaced with other AI algorithms. However, our research workflow can be repurposed to be used by researchers interested in designing automatic or semi-automatic systems to extract reliable and relevant data and information from social media streams searching for disaster response.

Figure 1 was never described in manuscript.

Thanks for pointing this out. We added explanation of Figure 1.

**(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.

In the results, section 4.2.2, you presented several figures of characteristic location of floods and also the results of GCV API classification. What was the purpose of that? I don't see the connection between those results and those described in section 4.2.1.?

Thank you for this comment. We explained the reasons for having these two separate sections.

4.2.1

**(Line 276-278)** This section illustrates the method of organizing images based on the locations, time, and the photo taker, which could help include as many images as possible into categories and then allow comparison of images within each category to elucidate the themes or topics of those categories.

4.2.2

**(Line 304-305)** This section illustrates an AI approach to detecting flood related features using GCV API and presents the result of the automatic detection.

Figure 1, what is Bag-of-Words extraction?

Thanks for the question. We added the explanation of Bag-of-Words extraction.

**(Line 160 -163)** 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.

Line 19 unnecessary description of API, to many parenthesis

Thank for identifying the error. It has been fixed.

You said that you purchased historical tweets from Twetter Inc. and 5202 tweets were geotagged. You did not mention the accuracy between coordinates and location names, did you check if they correspond? Coordinates could help in the situations like described on Line 181.

Thanks for the question. We extracted location names. One type of location name included the location service people use to pinpoint their location, and 339 tweets have this type of location, e.g., “I'm at Atmel Corporation (Colorado Springs, CO) <http://t.co/Vt4xWwudOu>”. In this location type, the url will show location on Foursquare map, and matches exactly the coordinates of the tweet. The second type of location is either very broad (e.g., Colorado Spring) or too specific to even be searched for (e.g., Ideal Market - @wholefoods) and 1131 tweets belong to the second type. Based on the two reasons, we did not compare tweet coordinates and location names within tweets.

Line 182, text "*However, as shown in symbol #2 in Figure 1, ..*" , I don understand where is symbol #2 ? Maybe you meant Figure 2? This error can be found in whole text. Please revise Figure numbers.

Thank for identifying this error, which has been fixed in relevant paragraphs and high-lighted.

Line 217 - 220. "*Consequently, keywords that were verified to be related to important incidents/places, such as Highway 36, could be used to extract tweets that were beyond the spatial limit of the study area or even do not possess any geo-location information. This approach would yield a larger volume of relevant tweets.*" Why you haven't done that approach in this research? It would drastically improve you model without too much effort. If you already considered that that tweet is reliable, finding similar tweets with similar keywords (like Highway 36), in close timeframe, should not be too much effort.

Thank you 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 324-338) 4.3. Extracting extra tweets using verified keywords

Sections 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 sections 4.1 and 4.2 and used the keywords to extract more tweets that do not possess any geo-location information. We believe that 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 used were from Table 1, which included: 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 additional 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.

Line 271. Can you please provide little more information about difference between GCV API image detection and web detection.

Thank for this question. We have incorporated the explanation about these two detection methods (highlighted below).

**(Line 307-311)** Image detection results include image annotations by detecting the features within images, and web detection uses the image content and its metadata to crawl the web and detect relevant information from the internet. Accuracy of image detection is based on the availability of training data and the detection algorithm. Accuracy of web detection is based on image content, metadata, and the availability of related information on the web.

**References**

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.

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.

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.

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.