Point-by-point response to reviewer's comments
Manuscript NHESS-2024-3

"Social sensing a volcanic eruption: application to Kīlauea 2018"

By James Hickey, James Young, Michelle Spruce, Ravi Pandit, Hywel Williams, Rudy Arthur, Wendy Stovall, and Matthew Head

---

First, we would like to extend our gratitude to the reviewer for their thoughtful critique of our manuscript, we thoroughly appreciate the time and effort that goes into this work. We have acknowledged all of the suggestions and believe the manuscript has been further improved by these revisions. We hope that you will now find it suitable for publication in NHESS. Our point-by-point responses to the reviewers' comments are detailed below in red text beneath the original comment. If any of the points remain unclear, we would be happy to revisit them and provide further clarification.

Yours sincerely,

On behalf of the authors,

James Hickey
(Corresponding Author)

Response to Reviewer 2

This piece presents a unique study analysing the content and sentiment of Tweets posted on Twitter before, during and after the 2018 eruption of Kilauea. The study scraped over 160,000 tweets from Twitter with reference to the eruption which were then filtered for relevance and classified by geolocation, content and sentiment. From the remaining tweets, the authors were able to identify some interesting trends in the changes in public sentiment linked to key activities in the eruption timeline (e.g. a more negative sentiment after tourists were injured on a boat). The implications of using something like social sensing during live volcanic crises would be significant in informing crisis and risk management on the ground and for informing our crisis communications work. Overall, the paper is a unique contribution to the academic literature and, after revision, would suit publication in NHESS.

This study provides some interesting insight into how Twitter traffic and public sentiments change throughout out the crisis, but currently the paper lacks the finer detail on the methodology and is a light touch on the analysis of the results. I would love to see more time spent on looking at a broader range of potential factors affecting the data set, and where possible a deeper content analysis on some of the trends that are clearly evident in the dataset (e.g. the influence of local culture on the content and sentiment analysis).

Here, I present some general and specific comments on the manuscript and hope the authors find them useful to strengthen the paper.

General comments:
The content analysis, although useful, feels quite limited. It would be useful to know why you didn't choose to delve deeper into the content of the tweets – e.g. references to religion and beliefs, references to political sentiments in the context of the eruption, and sentiments towards disaster response. Interestingly Figure 4.c clearly shows 'Pele' as a highly used term in tweets which would suggest that

there is more to the tweet content than just positive and negative sentiments. It's a link to local cultural beliefs and values in a location of indigenous population. Analysis of how and why this type of terminology entered into social discourse during an eruption is extremely useful for local authorities to inform their risk communication work and presents a potential use case of this methodology during volcanic crises. I think it would be great to try and delve a little deeper into some of the more qualitative content of the tweets and your dataset to draw the true value of the analysis.

We agree that further exploration of the content of the tweets, including the references to religion / culture / beliefs would be interesting and could provide useful information for further researchers and practitioners. However, we chose not to delve deeper into these aspects as they were not the intended aim of the study. As this is the first large-scale application of social sensing to volcanology, we focused our aim on the higher-order, broader picture of how feasible it is to use social sensing during a volcanic crisis, including whether social sensing can track and quantify changes in societal actions and emotional responses during an eruptive crisis, and whether those changes are coincident with different stages of the eruption. Additional focuses on what we consider, for the sake of the current study, to be lower-order specifics were outside and beyond our scope, but would be ideal for a range of follow-up studies. We have added text to clarify these opportunities to the manuscript, and are also making the tweet text openly available.

Specific comments:

Abstract

- This could do with a little more work to better reflect the article. Maybe include a mention of some of the specific linkages and trends e.g. an increase in negative sentiment was noted after specific incidents such as the tourist injured on a boat.
- L20 - What do you mean by 'societal actions'? I think making more clear references as per the previous suggestion and you can remove some of this non-specific language.

We agree the abstract was too vague. We have added more explicit sentences to better reflect the article, and clarified our use of 'social actions', while also revising the abstract in response to reviewer 1.

Introduction

- L31 – On this point of a no-mechanism to track information, it's also that this is really challenging because evaluations need resources, time and people and often our project funding ends and little is done.

  This point has been added to the manuscript.

- L71 – For a non-technical audience, perhaps define 'laze'.

  Definition has been added.

Methods and Results

This first section and section 2.1 are relatively weak. Some points could be expanded upon:

- Why Twitter over other social media platforms? Is it just the availability of the data or is there an advantage to using Twitter posts over other social media content?

  Yes, it is due to the availability of data, and we have added this point to the manuscript text. We are also currently working on similar Facebook data for a follow-up study.

- Are you able to get any demographic information to accompany the tweet data other than the location for a subset?

The only demographic information present for users is what they give in their public profile, which is limited to a few words, can change over time, and can also be untrue. Twitter/X terms and conditions explicitly restrict any kind of profiling or labelling of users based on demographics. One could possibly attempt to profile users and extract demographic information based on tweet content, or user profile pictures, but it is likely to be very highly uncertain and would have complicated ethics, and requirements for anonymisation and protecting privacy. Furthermore, GDPR says you cannot store "protected characteristics" unless you have a clear need, which we do not currently have as it is not directly relevant to our research questions; we are interested in the observations, regardless of who observes it. We have added text to the Methods to clarify we do not attempt to infer any demographic information, as well as some text to the discussion to explain how the lack of demographic information may affect our interpretations.

- P6 L132 - Why did you specifically use this number of tweets? It seems it accounts for ~4% of the total data set. What is the minimum number needed to effectively train machine learning models? Was this initial filtering done by just one person or was there a verification of multiple people across a sample?

  We used this number as it is a similar proportion to what has been successfully used in previously published natural hazard social sensing studies. This filtering step was carried out by a team of 5 human coders who also conducted inter-coder reliability checks. Text has been added to this section to clarify these points in the manuscript.

- P8 170 – you mention here about VADER and that it has a dictionary and lexicon – but I'm unfamiliar with the technique and would like to know what words it classifies how and how it was developed. How are the sentiment values weighted? A table might be useful to the reader.

  We have now provided the citation to the original VADER development paper (Hutto and Gilbert, 2014) which we previously forgot to include but provides the full details on the implementation of VADER. We do not deem it necessary to repeat such details in our current manuscript.

- P8 section on Content Analysis – how did you treat tweets that might have crossed multiple categories?

  Where tweets could have crossed multiple categories, they were assigned to the category they were deemed to represent most strongly in order to simplify the data analysis. This text has been added to the manuscript, as well as a note suggesting that future workers may wish to adapt the technique in the future to allow tweets to be placed in one or more categories.

- P10 L246 – these positive peaks suggest that something else is influencing the sentiments of the tweets. Were there any political or social interventions related to the eruption or response that could have triggered this? This is important to understand in more contextual detail because clearly the data are influenced by other external factors that should be controlled for.

  Our apologies, the original text here was not sufficient. Where we said "…and the eruption" we meant the physical eruption, as well as political and social interventions. So, to answer your question, 'no' there were no political or social interventions related to the eruption or response that we could identify that could have triggered the positive peaks. We have edited the text in the manuscript to make this point clearer.

- Your analysis looks at tweets based on the island and those not, but could you have extended this to include a category of diaspora? E.g. tweets from accounts linked to those on the island? The diaspora is well known to be an effective source of information during volcanic crises. Can you determine if the tweets on the island were truly residents vs. tourists?

In theory, yes, one could have analysed Tweets from accounts linked to those on the island, but it would require manually identifying those accounts to isolate their tweets, which would be a very challenging and time-consuming task.
We can not determine in an automated way if tweets on the island were from residents or tourists, and have added text to the discussion to make this clear and improve a point we previously made along the same lines. Distinguishing between residents and tourists could probably be worked out manually with further data scraping and exploration of user accounts, but it would be very time-consuming to do on an individual basis.

- What margin for error is there in your data set for the sentiment analysis linked to things like misinterpretation of tweets? Can you determine something from the manual validations of tweets you did with multiple assessors?

  There is potential for misinterpretation of some tweet text to introduce errors into the sentiment analysis, but with a big data approach these potential errors average out; consequently, VADER has been successfully applied in previous state-of-the-art social sensing studies. Text has been added to the manuscript to clarify this point. We can not determine manual validation of sentiment, as sentiment was not assessed when manually reading the tweets for relevancy or content.

Discussion
The discussion feels a little disappointing and I wanted a bit more depth to the analysis for example, how exactly could this method be used in real-time crises to inform decision-making. What are the advantages over more traditional methods (e.g. qualitative interviews) to assess sentiment? What is the application outside of crises, can this be used in a broader sense to help shape communication strategies prior to eruptions (or other disasters)?

We have added detail to all parts of the existing discussion, as well as added new aspects to the discussion (also in response to comments from reviewer 1). In particular, we now discuss in greater detail the mechanisms and opportunities for real-time social-sensing during volcanic eruptions and the advantages over more traditional qualitative interview approaches. Discussing the use of social sensing outside of crises to help shape communication strategies is outside the scope of this project and publication, so we have not added this.

P17 L329 – seems like quite a jump to link negative tweeting to mental health – this needs much more fleshing out in the text to enable the reader to understand why this link may exist. But also, this seems contrary to what you say about the word clouds in Figure 4 “Perhaps non-intuitively, tweets originating from Hawai'i are amongst the most positive, especially for regions with more than 30 geo-located tweets, driven largely by messages of hope and support (Fig. 4c).” (P11 L264).

The link between mental health and analyses of big data from social media is demonstrated in the references provided. We have fleshed out the text in the manuscript to make this point clearer.

We do not believe our point to be contradictory. We state that Hawaiian tweets are generally rated as being overall more positive with the sentiment analysis than other geographic regions on an absolute scale, but they still record decreases at the start of the eruption and around particular events (Figure 3b). Additionally, this point made in the discussion is also in reference to the global data set (i.e., not only related to the Hawaiian tweets), so we have better clarified that in the updated manuscript.

L338 – the language around ‘social actions’ and ‘sharing mitigation actions’ is very vague! Can you link tweets to communities actively adopting risk reduction practices e.g. family evacuation plans, stockpiling essential items, planning evacuation routes etc.).

We do agree that the text here was too vague and have now edited the text in the revised manuscript to be more explicit, to also link back to figures as evidence for our statements, and in line with comments from the other reviewer.
No, unfortunately our geospatial analyses in the majority of cases are not accurate enough to be able to pinpoint down to particular communities.

L344 – “Our analyses lend further weight to this finding…” how does your analysis do this? Be explicit.

This sentence now reads “Our analyses lend further weight to this finding by showing that Twitter posts were used to share warnings, advice, and observations of the eruption to social networks…”.