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 both reviewers and editors for their thoughtful and highly extensive critiques 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 1

This manuscript applies social sensing as a novel approach to understanding broad patterns in public reactions to Twitter posts ("tweets") related to and published during the 2018 eruption of Hawai'i's Kīlauea volcano. Specifically, this paper investigates temporal trends in topics of tweets published within Hawai'i and compares these to temporal patterns in user sentiment obtained through the VADER sentiment analysis program. The stated aim of this paper is to test 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. Another, broader goal--based on the abstract--is to identify and explain how the observed temporal trends in syn-eruptive tweet content and tweet sentiment scores reflect patterns in volcanic activity, civil protection actions and socioenomic pressures, and (possibly) identify a correlation between the posting of tweets containing warning or risk information and the resulting actions taken and/or sentiments felt by members of local communities in Hawai'i dealing with the eruption.

As a scientist who has conducted and published smaller scale qualitative and mixed methods analyses of social media and other communications during the 2018 Kīlauea eruption, I share the goals of the current manuscript's authors, and am excited to see how these authors employed a Twitter API and VADER to analyze and interpret the content and sentiment of over 100,000 tweets. I also commend the authors for presenting the results of this large dataset in concise and easily understandable figures, and for their choice of reader-friendly sequential color scales in several of these figures.

However, some of the main inferences and conclusions need to be explained or illustrated more clearly before the scientific quality of this manuscript is sufficient for publication. Although I consider these to be "minor" revisions, they are important enough that I strongly encourage the authors to incorporate

them before publication. These changes are summarized as follows (and explained in more detail in the Specific Comments):

- Stating more explicitly the patterns and relationships identified from the analyzed Twitter data, as well as the limitations in those patterns and relationships.
- Defining more clearly in the body of the manuscript several terms used in the Abstract, Discussion and Conclusions (see Specific Comments).
- Either justifying more clearly the inferred correlation between the "damage & disruption" cumulative tweet counts and field-based damage counts, or softening this claim to be a "weak" or "slight" similarity (see Specific Comments for more detail).
- Providing more detailed explanations of how the methodology of this study can be applied to real-time tracking methods, and also to tracking misinformation.
- Clearly defining what is "local"--specifically, whether "Hawai'i" refers to the Big Island or to the State of Hawaii.
- Providing more citations--I indicate the places where they are needed and also provide some examples of publications worth citing.

Detailed responses are provided below where these points are repeated in greater detail.

Finally, I have one major stylistic recommendation that I repeat in each of the relevant figures: changing some of the timeseries plots to be colorblind friendly.

Specific Comments (for Abstract and Conclusions, without line numbers)

Changes regarding statements made in the Abstract:

- The statements in the final two sentences of the Abstract should state more explicitly what can and cannot be inferred from the temporal patterns presented in the Results. The current language was too vague for me to understand without rereading the Results text and revisiting the accompanying figures. If you have the word and page count allowance to do so, I recommend adding one or two sentences that explain
    - What societal actions are taken (main categories or the most common actions),
    - What patterns are observed in volcanic activity, civil protection actions, and socioeconomic pressures,
    - Any observed correlations between one or more of these attributes (e.g., volcanic activity and societal actions)

  (Alternatively, if your Abstract word count is unable to incorporate these additions, consider adding them to the Conclusion)

We agree the abstract was too vague. We have added sentences describing the items suggested above, and also revised the abstract in response to reviewer 2.

- Even after rereading/revisiting the Results, there are several terms or inferences that need to be further defined or explained in the body of the manuscript:
    - It is not clear to me what "socioeconomic pressures" are reflected in your data and results (see "Specific Comments" for more information). This term is used several times in the paper but without being clearly explained or defined, apart from three citations in the Introduction.
    - Similarly, what are "community response actions"?

A full response to these comments is provided below where the same points are made in more detail ('Specific Comments' section), but in short, we have added further explicit mentions and explanations of where we infer socioeconomic pressures and community response actions, and added a definition of the term 'socioeconomic pressures' to the introduction.

- Finally, "hazard and risk information" is a broad term that, while sufficient for the abstract, should be further defined in the context of your analysis. My understanding from your Results section is that "hazard and risk information" includes warnings of possible hazards, advice for responding to hazards (particularly ash), and emergency response assistance announcements. If that is correct, be sure to relate these terms back to "hazard and risk information."

Yes, that is correct. To better clarify these points, we have adapted the main paragraph in the discussion section where we address how hazard and risk information is reacted to, and made sure to relate warnings of possible hazards, advice for responding to hazards (particularly ash), and emergency response assistance announcements back to "hazard and risk information".

- Also, if there are specific tweets discussing the risk posed by particular hazards, this should be explained in the main text as well, since hazard and risk are not interchangeable terms. (In fact, the proper usage of "risk" is a subject of ongoing debate in the hazard/risk communication research community, so tread carefully.)

We agree that risk and hazard are separate concepts, but in any case we did not find any tweets that explicitly link hazard to risk so this is a moot point.

Changes regarding statements made in the Conclusion:

- In the second-to-last sentence, the statement "Our work generally shows how hazard and risk information is discussed and reacted to on Twitter," should be expanded to explicitly state what kind of hazard and risk information is discussed (as recommended in my final comments on the Abstract).
  - Moreover, the types of reactions should be explicitly stated, as well as the observations/results that provide evidence of these reactions.

We have added text to this noted sentence to explicitly state what kind of hazard and risk information is discussed. However, we believe a conclusions section should be concise and summarise broad points, therefore we choose not to provide further text and detail to additionally re-state the observations/results that provide evidence of reactions to hazard and risk information. This information is presented and explained in the discussion section and does not need to be repeated in the conclusions section.

- The statement "which informs our understanding of community response actions and the efficacy of warnings and other official risk reduction communications" needs further elaboration. Specifically, explain
  - The main types of community response actions you are referring to (as recommended for the Abstract)
  - How you are evaluating the efficacy of warnings and other official risk reduction communications (and also keeping in mind the previous comment about defining what "risk" information is discussed in the tweets)

The elaboration of this statement is already provided at an earlier point in the conclusions paragraph where we mention “We find evidence of social action around sharing official warnings in the eruption’s lead up and early stages and sharing official mitigation actions later during the eruption. Such evidence is a positive outcome for volcano monitoring and emergency management organizations that are responsible for the official messaging”. Therefore, we do not deem it necessary to repeat such information again at the end of the conclusions in what is intended to be a broad summary sentence. We also note that the full details are also provided in the results and discussions section.

Specific Comments (with line numbers--also included in annotated pdf)

Line 25: Consider stating what distance range is defined as "near" a volcano.

Clarification (<100 km) added to the text.

Line 32: Include reference(s) on emotional state or reaction of affected populations.

Reference added.

Line 47: "from inaccessible locations"--explain how exactly these locations are inaccessible: physically, geographically, technologically? Presumably these are locations that allow individuals to access their social media accounts. Would be good to elaborate on this.

The text has been edited and the term “from inaccessible locations” no longer remains in the manuscript.

Lines 53-54: "strong positive correlation between social media activity and damage losses"--Does this mean higher social media activity with greater damage losses?

Yes, that is what is implied with the use of the term “positive correlation”. The text has been edited to make that clearer.

Lines 54-55: Does "negative correlation" mean more negative sentiment with higher damage losses?

Yes, that is what is implied with the use of the term “negative correlation”. The text has been edited to make that clearer.

Line 71: Consider adding a parenthetical definition for "laze" if your target audience is not limited to volcanologists.

Definition added.

Line 80: "driven by rock fall into the lowering lava lake"--this was an interesting phenomenon that is worth citing a source or two for! (Especially because the original hypothesized explanation--lava falling below the water table--was later disproven)

Reference added.

Line 83: "significant additional lower magnitude seismicity"-- What defines "significant" seismicity that is lower magnitude? I ask because this phrase may read more easily if you define it, e.g., "additional lower magnitude seismicity (M_ to M_)," . . . Or, if you still want to convey significant but un-felt seismicity, you might consider rephrasing: "collapses were associated with felt ~M5 earthquakes and additional unfelt but significant seismicity (M_ to M_),"

The text has been modified as suggested.

Lines 97-98: "increasing two-way dialogue and the speed and reach of official communications"-- I would encourage you to also cite Goldman et al. (2024), since two-way dialogues and reach of USGS Volcanoes' social media are significant components of this study, which were not captured in the 2023 paper. Full citation below:
Goldman, R.T., McBride, S.K., Stovall, W.K., & Damby, D.E. (2024). USGS and social media user dialogue and sentiment during the 2018 eruption of Kīlauea Volcano, Hawai’i. Frontiers in Communication, 9:986974. https://doi.org/10.3389/fcomm.2024.986974

This paper was not published at the time of submitting our manuscript; it has now been referenced in this line.

Line 121: 'Kilauea'--Does this include lowercase kilauea and spelling with kahakō (Kīlauea), if applicable? Would be good to clarify either way.

The search term is case-insensitive, but does not include spelling with kahakō (i.e., Kīlauea). This text has been added to the manuscript.

Line 138: "Source removal"-- Perhaps rename this as "External source removal" or "Removal based on Source" since it doesn't seem like you're removing the source attribute itself.

It has been renamed to “Source filter” in the updated manuscript, to better reflect the filtering process we are describing.

Line 142: "Username removal"-- Consider rephrasing to better describe the process. For example, "Removal based on Username".

It has been renamed to “Username filter” in the updated manuscript, to better reflect the filtering process we are describing.

Line 155: "F1 Score" in Table 1-- I would recommend you define F1 score, perhaps in your description of the Machine Learning Relevance Filter.

This definition has been added.

Lines 168-169: I would encourage you to cite Goldman et al. (2024)--full citation provided in an earlier comment--and any other studies that have used VADER for short-form social media sentiment analysis.

References added.

Line 170: I would also encourage you to cite the original study on VADER:
Hutto, C., and Gilbert, E. (2014). VADER: a parsimonious rule-based model for sentiment analysis of social media text. ICWSM 8, 216–225. doi: 10.1609/icwsm.v8i1.14550)

Reference added.

Line 172: "including the use of intensifiers, negations, and punctuation"-- You should also mention emoticons and slang, and cite Hutto and Gilbert (2014) here.

Reference and text added.

Lines 185-186: Provide citations for the process of inter-coder reliability checks.

A citation has been added.

Lines 186-188: Provide citation(s) that explain Fleiss Kappa agreement score and support your judgement that the score range was sufficient to progress.

Citations have been added.

Line 220: This is a particularly strong paragraph due to citing other sources, and explaining the significance and possibly reasons for the contrast between the high percentage or relevant volcano

tweets and low relevance of posts in other social sensing studies natural hazards. Use this as guidance for adding citations in the other portions of your main text as indicated in my comments.

No modifications necessary.

Lines 234-235: cite Hutto & Gilbert (2014).

Reference added.

Line 241: I recommend citing sources that discuss one or both of these explanations.

Reference added.

Line 243: Consider citing a source or two that also captures these sentiments ("personal shock and upset")

We are not aware of any such suitable reference to add here, and this line reflects our current research findings, so we have not added references here.

Lines 243-244: Are you able to cite an examples of this increased media attention and circulation of news articles on Twitter? You do this further down when describing dramatized/sensationalized accounts of the eruption, so it would be good to see some citations up here, as well!

We are not aware of any such suitable reference to add here, and this line reflects our current research findings, so we have not added references here.

Line 260: Define whether "Hawai'i" is the State of Hawaii or just the Big Island.

We have clarified in the text that we mean the State of Hawaii.

Lines 263-264: In addition to Calabrò et al. (2020), I recommend also citing Goldman et al. (2023), since interview participants from that study consistently stated that news media outlets provided sensationalized eruption coverage.

The Goldman et al (2023) study mentions state and national news as being too sensational, but this sentence is focussing on international news, so we do not believe citing Goldman et al (2023) would be appropriate here.

Line 281: Consider adding a parenthetical definition of paroxysm if your target audience is not exclusively volcanologists. I would also recommend you cite source(s) for the occurrence of this relatively significant event.

We have edited the text for additional, non-expert, clarity, and added a reference.

Line 288: Are you able to cite the news article reporting on the destruction of homes?

Citation added.

Line 303: I would be careful in how you define correlation here. To me, the shape of the observation, support & concern, and damage & disruption curves are much more linear than either field-based damage assessment curve, but particularly more linear than the "number of buildings in contact with lava" curve. Moreover, the increases are staggered in time, with the tweet curves increasing roughly

two weeks before the damage assessment curves do. I'm not confident there is an actual correlation here.

We now provide a quantitative correlation coefficient value; more detail in response below.

Lines 306-307: I would cite Neal et al. (2019), *Science*, and any other relevant publications describing this event and the ensuing change in lava flow impacts.

References added.

Lines 308-309: As implied in my earlier comment, I am not personally convinced there are correlations, or at least those strong enough for you to consider them favorable. I would urge you to either provide a stronger argument and evidence that this is the case, or soften your claim from "favorable" correlations to "weak" or "minor" correlations.
For example, can you point to other studies that compare cumulative count curves and clearly distinguish between (strongly) correlated data and uncorrelated (or weakly) correlated data? Is there a correlation coefficient or other metric you can provide to quantify the strength of your correlation? I do think your inclusion of the field-based damage assessments are informative and worth presenting, but would suggest you strengthen your argument, or otherwise soften your claim.

We now provide quantitative evidence for our noted correlations through calculated Pearson's Linear correlation coefficient (*r*) values. Our *r* values for the correlations noted in the text are 0.96 and 0.97. Values of the correlation coefficient can range from –1 to +1. A value of –1 indicates perfect negative correlation, while a value of +1 indicates perfect positive correlation. A value of 0 indicates no correlation. Our high, positive r values support our claims of 'favourable' correlations.

Line 329: "socioeconomic pressures"-- This needs to be explained more in the Results sections. I can infer that there are socioeconomic pressures from the word clouds in Figure 4 and the "damage & disruption" tweet counts and field-based damage assessment data in Figure 5, but there are missing explanations of how these data indicate socioeconomic pressure.
Some questions for you to consider:

- What socioeconomic pressures are indicated in the results presented in Figures 4 and 5?
- Are these pressures reflected in the sentiment analysis?
- If so, can you quantify the correlation and identify whether it is strong or weak?
- If not, are there other patterns in your data that point to these pressures? Are these patterns clearly illustrated in the corresponding figure(s)?

There is unfortunately no way to quantify the correlation between the socioeconomic pressures and the sentiment analysis. In the revised manuscript text, socioeconomic pressures are now explicitly mentioned in the results when describing: (i) temporal patterns in sentiment (e.g., from destruction of Vacationland) in section 3.2; (ii) tweets with negative sentiment originating within Hawaii (e.g., loss of homes, damage to property, and closure of the National Park) in section 3.2; and (iii) damage and destruction tweets (e.g., loss of homes, damage to property, and closure of the National Park) in section 3.3. In the revised text we also now point to the corresponding figures where the patterns / pressures are indicated, and have also added a definition of the term 'socioeconomic pressures' to the introduction.

Lines 332-333: This point--"there is no guarantee those individuals most affected, for example losing property or livelihoods, contributed to the data collection"--is worth exploring further. At a minimum, you should point to a few previously published studies that explore the impacts of this eruption (or other eruptions) on individuals' sentiments or well-being. Then, you should probably explain how your study can be built upon in order to address the uncertainty arising from this anonymised big data approach and thus provide an even more concrete correlation between social media sentiment and on-the-ground impacts to individuals.

We do not believe references to work exploring “the impacts of this eruption (or other eruptions) on individuals' sentiments or well-being” would be appropriate for a sentence talking about whether or not our dataset contained data from those most impacted by the eruption since we are more focussed on a limitation of the data collection approach and wish to keep our focus on this point.

We already state “using social sensing in parallel with traditional structured interviews of affected individuals will allow further verification and quality control of the social sensing approach, and allow researchers and practitioners to benefit from the respective advantages of both methodologies” in the last sentence of the discussion section, so do not feel it necessary to repeat this point here as well.

Line 334: I would advise clarifying how these news headlines would have contributed to negative sentiment, as you do in the Results. If it is due to sensationalizing, I would state that again here, and also recommend citing Goldman et al. (2023), *Volcanica*.

We have clarified in the text that we mean due to sensationalising, and added the suggested reference.

Line 344: "Our analyses lend further weight to this finding"-- How? You should explain which correlations illustrate the positive impact of sharing warnings and mitigation actions on user sentiment, and indicate which figures show these correlations.

This sentence has been edited for additional clarity and explanation, also in line with feedback from reviewer 2. The noted line/sentence was not discussing user sentiment, so we are not able to distinguish what this reviewer comment additionally means with respect to ‘user sentiment’, but we do now also refer back to Figure 5.

I've also noted in your Conclusions section that the manuscript does not currently provide a correlation between the occurrence/timing of warnings and risk reduction communications (on the one hand), and community response actions and affect (if any) on user sentiment on the other. Put another way, it is not clear to me that a link has been established between warning/risk reduction communications and community response or sentiment.

We believe the original manuscript text and figures indicated on several occasions where there were temporal correlations between warning / risk reduction communications and community response actions (e.g., section 3.3 and Fig 5). However, we have now also amplified these points in a number of places to further demonstrate these points.

Lines 350-352: It's not clear to me what point you are aiming to get across in this paragraph. Are you advocating for the incorporation of social sensing in more scientific studies of crowd-sourced observations? How does that improve upon the approach of Wadsworth et al. (2022)? Or, what is/are the main weaknesses of the Wadworth et al. approach that social sensing addresses?
If you address the above questions, and provide a more natural segue into your next paragraph on the broader implications of automated social sensing data collection and analysis, this paragraph will be much stronger.

We agree this paragraph was relatively weak. The text in this paragraph has now been edited to provide greater clarity along the lines suggested by the review comment. We also edited the text to provide a better segue into the following paragraph.

Line 353: Or else what? Elaborate on the consequence of having insufficient metadata.

This comment is no longer relevant after the editing in relation to the comment above.

Line 356: "in real-time"-- You should cite some studies that have already utilized real-time social sensing.

References have been added.

Lines 358-359: Goldman et al. (2024), Frontiers in Communication, would be another relevant source to cite for tracking the spread of misinformation on social media during an eruption event.

Reference added to reviewer's paper.

Lines 359-360: I'm not sure what you mean by "irrelevant" data, and also what "this approach" refers to. Please clarify. (See also my technical correction for this sentence).

The sentence has been re-worded along the lines suggested in the technical corrections, and also to provide extra clarity.

Lines 361-362: I recommend you cite studies that have studied posts in different languages and/or across different social media platforms. For the latter, here are two publications:
Hughes, A. L. et al. (2014) Online public communications by police and fire services during the 2012 Hurricane Sandy. In Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, 1505–1514. https://doi.org/10.1145/2556288.2557227
Ruan, T., Kong, Q.,McBride, S. K., Sethjiwala, A., and Lv, Q. (2022). Cross-platform analysis of public responses to the 2019 Ridgecrest earthquake sequence on Twitter and Reddit. Sci. Rep. 12:1634. doi: 10.1038/s41598-022-05359-9

Instead of adding further references here at the end of the discussion, we have edited the text in the introduction where we provide background information to specify that past social sensing studies have used different social media networks.

Line 363: Provide citations of this phenomenon, particularly if the term "signal" is defined.

There are no citations to add since we are discussing avenues for potential future work, and similar work does not exist for us to be able to reference.

Line 364: I would suggest using the term "non-local" instead of "external," since external can mean information external to a particular organization (e.g., an official volcano monitoring agency), regardless of its locality.

We prefer to stick with the 'external' wording, since one may also be interested in information external to a particular organisation depending on the context of the study.

Line 366: Regarding the use of traditional structured interviews, I would cite the following publications:
Donovan, A., J. R. Eiser, and R. S. J. Sparks (2014). "Scientists' views about lay perceptions of volcanic hazard and risk". Journal of Applied Volcanology 3(1). issn: 2191-5040. doi: 10.1186/s13617-014-0015-5.
Goldman et al. (2023), Volcanica. (Full citation already included in manuscript)
Haynes, K., J. Barclay, and N. Pidgeon (2008). "The issue oftrust and its influence on risk communication during a volcaniccrisis". Bulletin of Volcanology 70(5), pages 605–621. issn: 1432-0819. doi: 10.1007/s00445-007-0156-z.
Naismith, A., M. T. Armijos, E. A. Barrios Escobar,W. Chigna, and I. M. Watson (2020). "Fireside tales: understanding experiences of previous eruptions among other factors that influence the decision to evacuate from eruptive activity of Volcán de Fuego". Volcanica 3(2), pages 205–226. issn: 2610-3540. doi: 10.30909/vol.03.02.205226.

This sentence is talking about the combination of traditional interviews with a social sensing method, so we do not believe references related to only traditional interviews or analysis of social media content are appropriate or warranted.

Lines 367-368: Regarding the benefit of complementing qualitiative interviews with quantitative social sensing methods, I would cite the following publications:
Creswell, J. W. (2009). Research Design: Qualitative, Quantitative, and Mixed Methods Approaches, 3rd Edn. Thousand Oaks, CA: Sage Publications, Inc.
Goldman, R.T., McBride, S.K., Stovall, W.K., & Damby, D.E. (2024). USGS and social media user dialogue and sentiment during the 2018 eruption of Kīlauea Volcano, Hawai'i. Frontiers in Communication, 9:986974. https://doi.org/10.3389/fcomm.2024.986974
Graham, O., Thomas, T., Hicks, A., Edwards, S., Juman, A., Ramroop, A., et al. (2023). Facts, Faith and Facebook: science communication during the 2020–2021 La Soufrière, St. Vincent volcanic eruption. SP 539, SP539-2022–289. doi: 10.1144/SP539-2022-289
Ruan, T., Kong, Q.,McBride, S. K., Sethjiwala, A., and Lv, Q. (2022). Cross-platform analysis of public responses to the 2019 Ridgecrest earthquake sequence on Twitter and Reddit. Sci. Rep. 12:1634. doi: 10.1038/s41598-022-05359-9

We have added the reference to Creswell. As the sentence is talking about the combination of traditional interviews with a social sensing method, we do not believe references related to only traditional interviews or analysis of social media content are appropriate or warranted.

Line 380: "similar temporal trend"-- This language makes more sense than the stronger claim of "favorable correlation" that I critiqued in your Results section.

No modifications necessary.

Line 382: See overarching Conclusions comment near start of "Specific Comments" section of interactive comments regarding "efficacy of warnings and other official risk reduction communications."

Response provided above where first mentioned by the reviewer.

Lines 383-384 (final clause of final sentence of Conclusion): This point needs to be explained more in the Discussion, particularly the transition to real-time data collection and monitoring misinformation.

We have revised the final paragraph of the discussion to provide considerably more explanation around the possibility for real-time data collection and analysis.

Line 399: Clarify that readers must have a Zenodo account in order to access.

The data will be freely available after publication.

Technical Corrections (for figures--also included in annotated pdf)

Figure 1(a): The isopachs are hard to see, especially the 50 cm since the color is nearly identical to the lava flows there. Maybe consider making the isopachs dashed black lines and distinguishing them by different dash-lengths, or perhaps just the thickness labels alone? This would allow the isopachs to be legible in grayscale, as well.

The isopachs have been changed to a blue sequence to remove the clash with the lava flow colours, and maintain distinction if ever viewed in greyscale, while still maintaining good distinction in (the most likely) colour view.

Figure 1(b): I really like this figure overall, but can't help noticing how small the event text is in panel (b). May be worth providing a table listing each event and the corresponding aviation color code in the supplement?

We do not believe it to be necessary to duplicate the information in an additional table so opt not to implement this suggested revision.

Figure 2(a)-2(b): I would suggest assigning a different color than green to the "all" tweets lines in panels (a) and (b). Perhaps black or dark gray, both of which would stand out better to colorblind readers (or users with a grayscale copy of your manuscript).
I would also suggest you make the legend and axis tickmark label fonts slightly larger, at least as large as your axis labels and significant event labels.

These changes have been implemented.

Line 215 (Figure 2 caption): I don't see the term "bigram" defined anywhere--I would recommend you do so in the Methods.

The definition has been added to the figure caption.

Lines 215-216 (Figure 2 caption): Do you have frequency values for the most common and least common bigrams shown in panels c) and d)? It would give a sense of scale and also be useful to compare with the daily tweet frequencies in panels a) and b).

The frequency values for the most and least common bigrams have been added to the caption.

Figure 3: I would suggest changing the color of the timeseries in panel (a) from green to a different color (e.g., yellow-orange) to provide contrast with the timeseries of panel (b) that is colorblind friendly, while maintaining contrast with your red mean value lines.
(That being said, given that you have two separate panels for each timeseries, this suggestion should take lower precedence than my color adjustment suggestions for your other figures, where the timeseries overlap or three or more timeseries are being compared).
As with Figure 2, I would also suggest increasing the font size of your axis tickmark labels.

The font sizes have been increased and the non-colourblind friendly green-red contrast has been removed from the figure.

Figure 4(a)-4(b): Given the larger overall size of this figure, I think your tickmark labels are a good size! However, I would suggest making the colorbar legend label and tickmark numbers slightly larger to match the grid axis labels, especially since you have the Log-10 subscript.

This change has been implemented.

Line 270 (Figure 4 caption): I would explicitly define the scores for positive and negative sentiment score groups.

Definitions added to caption.

Line 271 (Figure 4 caption): As with Figure 2, I would recommend you define the max and min frequency counts for the largest and smallest words, respectively, in each wordcloud.

We have clarified the caption to confirm that larger bigrams in the wordclouds here indicate a greater relative degree of occurrence, i.e., large bigrams contain words which are more common in positive

tweets but uncommon in negative tweets (or vice versa). Therefore, absolute max and min values are arbitrary and unnecessary.

Figure 5: I would suggest choosing a legend color scheme akin to a sequential gradation, such as the thermal color legend used in Figure 1, the cyan to magenta gradient in Figure 4 (a)-(b), or the red-to-brown/black gradient used in your word clouds in Figure 2 panels (c)-(d).
This would benefit colorblind readers or readers with a grayscale version of your manuscript.
(Link with other examples of sequential color gradients, if helpful): https://matplotlib.org/stable/users/explain/colors/colormaps.html

We do not believe a sequential colour scale is appropriate for these data as they are separate, discrete data series.

Are you able to correlate the earliest syn-eruption peaks in panels (a), (b), (c), and (e) with specific events or types of tweets? If so, I would also label those. If not, are these peaks attributable to the start of the eruption itself? It may be worth reiterating in the figure caption if that is the case.

These peaks are only attributable to the start of the eruption, which has now been additionally clarified in the revised figure caption.

As with Figures 2-3, I would suggest making the tickmark labels a larger font, as well as the font for each of your five timeseries categories (observation, warning, etc.). The size of your "Daily Tweets" and event labels are good. Same suggestion for panels (f)-(g) as with (a)-(e): larger font for the axis tickmark labels.

This change has been implemented.

I like your usage of dashed lines in panel (g) to distinguish between lines--you might consider assigning different dash marks or other symbols to your timeseries lines as an alternative or complementary strategy to the sequential gradations I've suggested for this and other figures.

We prefer to keep the dashed line distinction only for the difference between our social sensing data (solid lines) and the independent field-based damage data (dashed lines).

Lines 296-297 (Figure 5 caption): Is the normalization time period for panel (g) identical to the gray "watch" period in the preceding panels? Consider clarifying this.

The exact time period for normalisation has been added to the caption.

Line 297 ("building damage data," Figure 5 caption): Is this the same as "contact with lava"? I would advise clarifying this point in the text, since in my mind contact with lava can range in severity from minor exterior damage to complete destruction of a building.

The text has been edited to provide better clarity.

Line 306: Consider adding a vertical line in Figure 5(g) indicating June 3, to help illustrate the contrasting rates of tweet accumulation before and after this date.

This change has been implemented.

Table 2: You might consider a light gray shading background for the bold-face rows and columns as an additional way to create contrast between these and the non-bolded table cells.

Shading has been added as suggested in the comment.

Figure A1: I would suggest choosing a legend color scheme akin to a sequential gradation, such as the thermal color legend used in Figure 1, the cyan to magenta gradient in Figure 4 (a)-(b), or the red-to-brown/black gradient used in your word clouds in Figure 2 panels (c)-(d). This would benefit colorblind readers or readers with a grayscale version of your manuscript.
(Link with other examples of sequential color gradients, if helpful): https://matplotlib.org/stable/users/explain/colors/colormaps.html

This change has been implemented.

Figure A2: I would also suggest replacing the green "not news" line with a different color, such as blue, to aid red-green colorblind readers.

This change has been implemented.

For both Figures A1 and A2, I would also suggest larger axis tickmark labels.

This change has been implemented.

Technical Corrections (with line numbers)

Line 311: Consider replacing the highlighted text with this grammatical/stylistic edit: "highlighted a high proportion of these were related"

This suggestion has been implemented.

Line 315: capitalize "Volcanoes" in "USGS Volcanoes"

Correction made.

Lines 335-336: I think the clear message gets lost in how this sentence is structured, which currently reads more like a dependent clause. Is the clear message that Hawaiian tweets with a negative sentiment score show a harmful effect on societal mood? Is the message that this harmful effect is the result of localized eruption impacts? I recommend you rephrase to make the meaning clearer.

This sentence has been rephrased.

Lines 359-360: This sentence may read easier with less qualifying language and without the double negative. (e.g., "This approach may be facilitated through collecting highly relevant data within online volcanic conversation.")

The sentence has been re-worded along the lines suggested.

Line 361: You may want to tighten the wording of this sentence. Example: ". . . if improved geolocation information are available, and to compare the insights provided by different languages, social media networks, or messaging applications."

The sentence has been re-worded using the suggested text.

Line 364: I might add: "bias our understanding of events away from their perceptions by local communities"

The comparison to the local scale is provided in the latter part of the sentence already, so we prefer to keep the original wording.

Line 371: delete "very"

Word deleted.

Line 373: add a comma after "eruption"

Comma added.

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…".