Nat. Hazards Earth Syst. Sci. Discuss.,
https://doi.org/10.5194/nhess-2020-335-RC1, 2020

Natural Hazards and Earth System Sciences
Discussions

Interactive comment on "Online Urban Waterlogging Monitoring Based on Recurrent Neural Network for Classification of Microblogging Text" *by* Hui Liu et al.

Anonymous Referee #1

Received and published: 2 December 2020

The article is well-structured although it might be good to do another language edit. I was interested to see that RNN networks work so well for Chinese text. While the methods applied in the article are not very new, this is indeed – to the best of my knowledge – the first dataset related to waterlogging specifically. However, there are a number of things that the authors should address.

L 19. Would be nice to go in a bit more detail here. Do the authors mean to say predict or detect? Online media is especially good at detecting events, but are very limited for predicting events.

L. 94. Is this location attached to every post? The localization for the waterlogging

seems to be done based on the text itself. Could the authors use the coordinates attached to the post, to deal with duplicate location names at a later stage. This could make the application described at the end of the paper more useful as less or no manual work is required depending on your response to my comment regarding L. 125. See for example: de Bruijn, J.A., de Moel, H., Jongman, B. et al. TAGGS: Grouping Tweets to Improve Global Geoparsing for Disaster Response. J geovis spat anal 2, 2 (2018). https://doi.org/10.1007/s41651-017-0010-6

L. 97. It would be nice for the non-Chinese reader to include some information on communities. For example, how many communities are there on average per city? If only 307 cities are included, could posts also be referring to cities that are not included in the city database and therefore not detected as duplicate?

L. 105. Are there more ambiguous cases, such as the threat of waterlogging, severe rainfall but no waterlogging etc.? Was there any ambiguity with real floods? If so, the authors should in more detail discuss how these posts tagged as relevant or not.

L. 125. If I correctly understand the authors 1) create a dataset with keywords "æůź" and "çğŕæřť", 2) manually select only those that are relevant to waterlogging and 3) select those that also mention a community. The positive samples are those that both mention a community and waterlogging, while the negative posts are only those that are not related to waterlogging. This means that the tweets that are related to waterlogging but do not mention a location are neither in the positive nor in the positive set.

Therefore, 1) A lot of manual work is required should the classifiers be applied to another dataset. This makes the automated classifier hard to apply "in the wild". 2) The classifier could in fact be learning location names rather than other features because location names are artificially inflated in the positive class.

L. 139. How were stop words determined?

L. 216. Recurrent neural networks and LSTMs are a relatively older neural network for text classification. Did the authors consider algorithms such as BERT and XML-RoBERTa. Usually these algorithms perform much better.

L. 402. The waterlogging posts seem to be manually located – or at least take manual work before deduplication – this should be mentioned in the discussion.

L. 410. The authors mention that a higher GDP indicates more waterlogging. However, provinces in the east also have a much higher population, and internet penetration. These factors need to be considered when making these claims.

L. 415. The labels in Figure 8 are Chinese.

L. 417. Could the authors validate the results, for example by comparing the results to significant rainfall events?

L. 428. Could the authors elaborate what needs to be done to apply the methods in this study in real-time (as the application shown in 4.5 only makes sense when applied in real-time)?

Interactive comment on Nat. Hazards Earth Syst. Sci. Discuss., https://doi.org/10.5194/nhess-2020-335, 2020.