Reviewer’s report

Title: Deep learning for pollen allergy surveillance from Twitter in Australia.

Version: 0 Date: 27 May 2019

Reviewer: hansi zhang

Reviewer's report:

The paper introduces a framework for real-time surveillance of pollen allergy from social media data using a deep learning based method. The data mainly comes from Twitter, Youtube and Reddit. The data was grouped based on the availability of the geo-location. One group is for classifier training. Another group is for correlation detection with the weather information. For the classifier, the author compared 4 difference deep learning models with difference dimensions. The results showed that GloVe/300 outperform the others. The weather correlation looks really high based on the figures and the correlation is well explained. The idea of using deep learning for surveillance from social media is not new, but the experiments on Australia "pollen allergy" are much needed. Overall, the paper is well-written, well-organized, and well-motivated. However, there are a few concerns as detailed below as well.

1) The background section is well-motivated. It helps me to understand the motivation of the paper. However, there is no summery for the background, each small section are separated. Please provide a summery paragraph to describe what you did for this project, compared with other similar research (e.g, deep learning for surveillance of social media data), what is your main contributions.

2) In the Method section, the keywords used for search is 'hay fever’ OR 'hayfever' OR 'pollen allergy'. What about 'allergic rhinitis' which is also know as 'hay fever’? Since this keywords is highly relevant, you might be able to enlarge your corpus. The posts relevant to allergic rhinitis may provide additional information regarding symptoms, treatment, etc.

3) The data is collected from 3 sources. I'm interested in how many posts for each datasets. Please provide a table to provide more details about your data sources. Are most of the posts from Twitter?
4) I noticed in the annotation process, you annotated 4,148 posts. Is this the datasets 1 that the geo-information is available? Please clarify how many post in each dataset (dataset 1 and dataset 2).

5) In table 1, author lists some example tweets. For social media studies using data that could be potentially sensitive, I think it's reasonable to anonymize tweets. It is exceptionally easy to identify Twitter users just by searching for tweets with particular text, and as this study is partially about consumer's tweets regarding HPV it seems prudent to make it more difficult to access the raw tweets. Only remove the user name is not enough.

6) For the 4 categories of your annotation classes. How do you consider information posted by doctors, nurses, and other professionals?

7) For the processing of the posts, do you only consider low case? How about the URLs in the posts, non-english? If you have processed these, please clarify in the manuscript.

8) In the first part of the discussion section, the author said "what can be attributed to relatively moderate training dataset size of (20k posts)". Do you mean 4K posts for training the classifier? If yes, please correct the number.

9) Typo: "whole Australia (1) and its major cities (2) were created." The numbers should be moved to the front (e.g., (1) Australia) to keep consistent with the previous example (e.g., "(1) Alice Springs (radius=2; 000mi), and (2) Sydney,"")

10) The weather correlations are visualized in figures. What about also providing the correlation using Pearson correlation coefficient?
Are the methods appropriate and well described?
If not, please specify what is required in your comments to the authors.
Yes

Does the work include the necessary controls?
If not, please specify which controls are required in your comments to the authors.
Yes

Are the conclusions drawn adequately supported by the data shown?
If not, please explain in your comments to the authors.
Yes

Are you able to assess any statistics in the manuscript or would you recommend an additional statistical review?
If an additional statistical review is recommended, please specify what aspects require further assessment in your comments to the editors.
Not relevant to this manuscript

Quality of written English
Please indicate the quality of language in the manuscript:
Acceptable
Declaration of competing interests

Please complete a declaration of competing interests, considering the following questions:

1. Have you in the past five years received reimbursements, fees, funding, or salary from an organisation that may in any way gain or lose financially from the publication of this manuscript, either now or in the future?

2. Do you hold any stocks or shares in an organisation that may in any way gain or lose financially from the publication of this manuscript, either now or in the future?

3. Do you hold or are you currently applying for any patents relating to the content of the manuscript?

4. Have you received reimbursements, fees, funding, or salary from an organization that holds or has applied for patents relating to the content of the manuscript?

5. Do you have any other financial competing interests?

6. Do you have any non-financial competing interests in relation to this paper?

If you can answer no to all of the above, write 'I declare that I have no competing interests' below.

If your reply is yes to any, please give details below.

I declare that I have no competing interests
I agree to the open peer review policy of the journal. I understand that my name will be included on my report to the authors and, if the manuscript is accepted for publication, my named report including any attachments I upload will be posted on the website along with the authors' responses. I agree for my report to be made available under an Open Access Creative Commons CC-BY license (http://creativecommons.org/licenses/by/4.0/). I understand that any comments which I do not wish to be included in my named report can be included as confidential comments to the editors, which will not be published.

I agree to the open peer review policy of the journal