Author’s response to reviews

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

Authors:

Jia Rong (jia.rong@vu.edu.au)
Sandra Michalska (sandra.michalska@live.vu.edu.au)
Sudha Subramani (sudha.subramani1@vu.edu.au)
Jiahua Du (jiahua.du@live.vu.edu.au)
Hua Wang (hua.wang@vu.edu.au)

Version: 1 Date: 10 Jul 2019

Author’s response to reviews:

BMC Medical Informatics and Decision Making

Melbourne, 9th July 2019

Dear Editor

We greatly appreciate your feedback on how to improve the quality of our manuscript ‘Deep learning for pollen allergy surveillance from Twitter in Australia’. Please find below the point-by-point responses to any concerns raised by the reviewers. The required modifications have been highlighted in the article as suggested. Thank you once again for your constructive comments and opportunity to resubmit again.

Best regards

Sandra Michalska
Editor Comments:

1. Individual users can be reidentified from the manuscript. Consider obfuscating the text to avoid re-identification if this is aligned with the ethical approval/waiver of the university.

Thank you. Our previous work in similar domain (health-related knowledge extraction from Social Media [1]) did not entail the ethical approval requirement, thus the approach followed.

Still, we sought the advice from the Associate Professor Nicola Reavley, Ethics Committee Member, Melbourne School of Population and Global Health. Prof. Reavley published a study on mental health assessment using Twitter data. The posts were openly released without the ethical approval [2]. Also, in [7] the authors observed ‘Twitter data are often relegated to nonhuman participants research, because collected tweets are publicly available’ (the necessary measures to reduce the risk of users’ re-identification were yet recommended).

We further identified the studies of Social Media data mining in public health context, which also did not report any ethical considerations. Please see [3], [4], and [5]. Still, we absolutely agree with your concern regarding the potential privacy issues of the users. Hence, the number of studies has been reviewed from the perspective of any mitigation steps implemented to avoid users’ re-identification. In [6], the authors stated that ‘it is common practice in social computing research to analyse publicly available data without the posters’ consent or knowledge’, and ‘to reduce risk of participants’ data being resurfaced, we changed quotes slightly and used paraphrasing to obfuscate posters’ data, a common method in ethical social computing work’.

As a result of your suggestion and further research, all of the posts included in the manuscript have been obfuscated (i.e. paraphrased, shortened). The obfuscation was mostly applied to the words/phrases irrelevant from Pollen Allergy surveillance point of view. The original form of the symptoms/treatments mentions was retained though (any modification/removal could reduce the value of study, which focal point is the actual language used on Social Media platforms to report the Hay Fever instances).

Also, we believe that Pollen Allergy can be considered less of a sensitive topic in comparison with the multitude of Social Media-based studies on mental health assessment, violence victims’ identification, or abuse incidents mining that use social media as a data source. In the examples mentioned, potential users’ re-identification could carry an undesirable impact on social functioning of the sufferers/victims. Still, the obfuscation was applied to alleviate any potential privacy concerns. The information about obfuscation has been included in Table 5 footnotes of the manuscript.


[3] Doan et al., ‘Extracting health-related causality from twitter messages using natural language processing’, BMC Medical Informatics and Decision Making (2019) https://doi.org/10.1186/s12911-019-0785-0 (Ethics approval and consent to participate: This research is not human research and did not require IRB approval.)


2. Consider that readers will not necessarily be familiar with all of the NLP and deep learning approaches, so more detail may be required - this might be in the manuscript text or in an additional supplementary file if it extends over several pages. Always important to explain how parameters were selected and to report which alternatives were tested but not reported because they didn't work - that is a hallmark of good machine learning study reporting.

We greatly appreciate your feedback. We further expanded the details of the approach followed (as also suggested by the reviewers), in particular the data extraction process, pre-processing steps, and the parameters selected for classifier training. The relevant sections have been highlighted in manuscript (Data extraction, Training and Testing sections).

Given the computational cost of running Deep Learning classifiers, we followed either default or parameters reported in previous works, as noted in the relevant section. In terms of the various
alternatives exploration, e.g. the number of dimensions for the embeddings vectors development, or default versus domain-specific embeddings schema, the results have been reported appropriately and highlighted in the corresponding table (Table 3).

We further followed the guidelines outlined in the paper above [7] regarding the methodological considerations of Twitter research in health domain. We hope the corrections made have considerably improved the quality of our manuscript. Thank you for guidelines.

Reviewer reports:

Jason Colditz (Reviewer 1): The present study seeks to develop and evaluate deep learning classifiers for tracking hayfever in Twitter posts. The target data were composed of geotagged tweets surrounding distinct locations in Australia. Classifiers relied on lowercased tweet text that was tokenized and vectorized with word embeddings from both GloVe and "HF" approaches. HF word embeddings were based on lexical corpora scraped from Twitter, Reddit, and YouTube platforms. Several ML approaches and embedding strategies were compared for accuracy against double-coded data. Aggregate trends in data were validated against weather trends that predict hayfever in the population. Results indicated that the GRU classifier with GloVe embeddings functioned best (0.879 accuracy), and that the classified Twitter trends were visually concordant with weather-related hayfever risk factors.

General feedback:

Overall, I found the paper to be interesting and informative about recent machine learning and NLP methods for analyzing Twitter data. These approaches seem generally appropriate for the research aims and scope, and classifier performance was admirable. However, as a stated goal is to illustrate inner workings of the process and aide in reproducibility of what is commonly considered a "black-box" approach, clarification on a few additional facets would help to achieve this objective. Many of the following questions/concerns could be briefly clarified or justified in just a sentence or two. Clarification on these points will aide in further understanding the scope of work and will bolster reproducibility of the methods, so that they can be appropriately applied among other contexts or health domains. Given some additional clarifications on methods, this paper is poised to make a valuable contribution to the health informatics literature.

We very much appreciate your feedback. We have further elaborated on the point regarding the inner workings of the ‘black box’ approach illustration. We also have given a specific example on how it was achieved. The following passage has been added into the Conclusions section:
The in-depth investigation of predictive probabilities and embeddings weights on the real-world example has provided an insight into the internal workings of the classifier. For instance, the top similar terms associated with HF-related keywords were produced to demonstrate why the selected approach worked, i.e. the vector for 'antihistamines' included a wide range of specific medications' brands, proving suitable for the emerging treatments discovery - valuable information for the robust Pollen Allergy Surveillance System development.

I pose a few specific questions/concerns related to Twitter data collection and formatting, which relate back to a recent paper that discusses these topics in greater depth (Colditz et al., "Toward Real-Time Infoveillance of Twitter Health Messages" [http://doi.org/10.2105/AJPH.2018.304497]). This will be a helpful reference to consult for contextualizing some of the feedback in comments #1-5, below.

In fact, we have found the paper suggested exceptionally useful when revising, structuring and filling any potential gaps in our existing methodological approach reporting. Thank you.

Specific Feedback:

1. The authors broadly state that target data were obtained from the Twitter API. Additional detail would be welcome for how exactly the Twitter API was accessed. For example, a Twitter API endpoint could have been accessed using a particular 3rd party tool (what tool?) or a custom script (is the script open source and/or built on an existing API wrapper library?). Overall, what did the client-side data collection framework look like? Is this particular data collection approach readily available to others? This clarification will aide in reproducibility of the work.

We absolutely agree. We have used the publicly available library and specified the parameters set. The data extraction procedure with all the details required for the results reproducibility and/or follow-up approach has been clarified in the manuscript in Methods section (please see below).
‘Custom script was used to extract the data using R programming language and ‘TwitteR’ package. The posts were captured retrospectively at regular time intervals, and the parameters were as follows:

- Search terms: ‘hayfever’ OR ‘hay fever’;

- Maximum number of tweets: n=1,000 (never reached due to limited number of posts meeting the specified criteria);

- Since/until dates: s=01/06/2018, u=31/12/2018 following the weekly pattern, i.e. 01-07/06, 08-14/06, and so on;

- Geo-coordinates: Alice Springs (-23.698, 133.880), Sydney (-33.868, 151.209), Melbourne (-37.813, 144.963), and Brisbane (-27.469, 153.025).

2. Which particular Twitter API endpoint was engaged (e.g., Sample Stream, Filtered Streams, Search)? Please clarify this in the data collection narrative. This will have some relevance for understanding the comprehensiveness and timeliness of tweets that were captured for primary analysis. Given that "the posts were extracted weekly" it would appear that the Search API was used to capture retrospective data at regular intervals, though this isn't entirely clear in the paper. Are there additional considerations or limitations that should be noted for this particular data collection strategy?

Given the very specific criteria and limited number of data meeting those criteria (number of Twitter users in Australia, search query, location, etc.), all the relevant posts were extracted. Thus, no particular sampling or filtering approach was applied. In terms of the retrospective data collection, the point was clarified in the manuscript as suggested.

3. Clarifying the data collection approach (#1) and API endpoint (#2) will also help to contextualize why "posts from remaining 23 days were not captured due to technical issues." Some missing data are to be expected in this type of work and I applaud the authors for being transparent about exact dates that missing data occurred. However, I think that readers may also be interested to understand underlying causes for why gaps in data collection seemed to happen
at fairly regular intervals (i.e., end-of-week from Jun-Oct, but not so much thereafter). These technical issues might be briefly clarified as a note under Figure 2. This serves to enhance study reproducibility and may present pragmatic considerations for continued improvement of Twitter data collection procedures (which appear to have improved over the course of this study). These considerations might also be briefly covered in the discussion narrative.

The technical issues mentioned lie solely on the part of the authors. ‘Until date’ parameter excludes the tweets posted on that date, of which the authors have been unaware (therefore the end-of-week gaps). It has been realised and corrected from October onwards. The clarification note has been added and highlighted in the manuscript.

4. I anticipate that the search keywords for the target data (i.e., "hayfever", "hay fever") would have strong precision (specificity), but are lacking in recall (sensitivity). For example, inclusion of search terms like "allergies" or "sneezing" could have substantially enhanced search recall. As the present study used lexical embeddings, it might have been beneficial to include a wider scope of search terms, resulting in a corpus that accounts for broader contexts of allergy-related posts. Some additional justification for excluding other potentially informative search terms is warranted. While I understand that this is currently framed as a "case study" (See #10 for why this is not necessarily an appropriate study classification), a rationale for the scope of search terms is nonetheless important. Related limitations should also be discussed.

Initially, the list of potential symptoms and overall Hay Fever-related keywords was defined to extract as much of potentially relevant posts as possible. The terms included: ‘allergy’, ‘spring’, ‘pollen’, ‘sneeze’, ‘watery eyes’, ‘red eyes’, ‘itchy throat’, ‘runny nose’, etc. After preliminary data exploration, only the fraction of tweets was indicative of Hay Fever. The numerous false positives were extracted given the common symptoms for various conditions. For instance, keywords ‘sneezing’ and ‘allergy’ were highly associated with ‘cat’, ‘dust’, and ‘peanut’ terms. Similarly, ‘runny nose’ and ‘itchy throat’ queries were commonly mentioned with Cold and Flu condition.

Unless Hay Fever was specifically indicated, the wide range of HF-related keywords introduced an excessive noise to the dataset. Thus, we prioritised precision over recall in the classifier development approach, its further enhancement with lexical embeddings, and finally the potential scope extension in the future work.

The appropriate paragraph explaining the search query strategy has been included in the manuscript.
'The high precision was prioritised over the high recall, thus the very narrow scope of the search terms. After preliminary data exploration, wider list of search queries introduced an excessive noise to the dataset. For instance, the generic term 'allergy' included other popular allergy types (i.e. Cats, Peanuts), and the specific symptoms such as 'sneezing', 'runny nose', 'watery eyes' frequently referred to the other common conditions (i.e. Cold, Flu).

5. Were emoji or other non-word features included in the present analysis? If so, how and why were they included? If not, how and why were they excluded? This might be clarified in the "Training and testing" section, where other approaches to text formatting are discussed. This concern relates primarily to the targeted Twitter data, but also to the data used in developing embeddings. For example, there are differences in emoji use among different social media platforms. Some of these features may be highly informative for Twitter data, but are less prevalent and/or rendered differently on YouTube and Reddit (where HF embedding data were also collected). In the discussion section, for the example of "hay fever sob", I would guess that "sob" is a word representation of the sob emoji (which is commonly used on Twitter). Punctuation handling is another area where additional clarification would be welcome. In some NLP embedding models, inclusion of punctuation can be helpful.

Though, punctuation can sometimes be problematic in deep learning models, depending on how they are pre-processed. It would be interesting to know more about why and how the authors decided to pre-process emoji and punctuation in this lexical framework.

We greatly appreciate your feedback and interest in the topic. In terms of the ‘hay fever sob’ expression, ‘sob’ is an actual term used, which algorithm was expected to identify as potential Hay Fever symptom (watery eyes). As the natural language used is the focal point of the current study, any additional examination of special characters, i.e. emojis, will serve as an interesting avenue for our future work within the health surveillance from Social Media.

Further clarification on pre-processing procedure in the current study has been added in Training and Testing section of the manuscript (please see below).

‘Pre-processing performed was minimal, and included removal of URLs, non-alphanumeric characters and lowercasing. In terms of emojis, their numerical representation was retained, following the punctuation removal. No excessive pre-processing was applied as models perform the operations on sequence of words in order they appear. Words are preserved in their original form without stemming/lemmatising due to their context-dependent representation, e.g. 'allergy', 'allergic', 'allergen'. Also, Sarker et al. suggested that stop words can play a positive effect on
classifier performance. Analogical pre-processing steps were implemented for the embeddings development.’

6. The authors clarify the keyword parameters, but not other characteristics of the data collected for developing HF embeddings. For example, what were the web crawling methods (e.g., API scripting, 3rd party tool; see also comment #1)? What proportion of the roughly 22k posts and/or comments came from each of Twitter, YouTube, or Reddit platforms? Does this estimate count each comment individually in the 22k, or are comments grouped with parent post content? Do the authors have a sense of post/comment length or other data characteristics that might differ among platforms? What were the data collection timeframes and/or other restrictions on data collection scope? Clarifying these and similar data collection aspects will help to elucidate any potential influence that drawing data from multiple platforms may have had on the embedding schema. I realize that some of these clarifications might be seen as trivial, because the HF embeddings did not ultimately improve upon GloVe embeddings. However, additional clarifications to enhance reproducibility and to inform future lexical data considerations will be of value to readers. This needn't be an exhaustive detailing of data characteristics for each platform, but readily available clarifications will add helpful context. These clarifications may warrant inclusion of a small table to report differences among platforms and the associated data characteristics that informed development of this particular embedding schema.

We understand the requirement for methods transparency and results reproducibility, thus have further detailed the particular embeddings schema development approach.

The relevant table specifying the source, number of posts, mean of post length, standard deviation of post length, min date and max date has been appropriately added into the manuscript.

‘The following web crawling methods were applied based on the data sources used: (i) Twitter - TwitteR R package, (ii) Reddit - RedditExtractoR R package, and (iii) YouTube - NVivo. To enhance results reproducibility and inform future research, the details of the particular embeddings development schema implemented have been presented in Table 1.’
7. There is a lack of information about how the HF embeddings were developed. For example, the Python word2vec library is a common approach, and there are similar libraries in R, and likely other approaches out there. Explaining the software tools and approaches used in embedding development is critical information. It also isn't clear what the window size (i.e., number of consecutive words) was for developing embeddings.

We further specified ‘Gensim library for Python that provides access to Word2Vec training algorithms was used, with the window size set to 5’ (added to the manuscript).

8. It isn't clear which pre-trained GloVe embeddings were used in this study. The footnote reference "[1]" points to the Stanford NLP site, which has several available pre-trained vector files. I would assume that the Twitter-based vector file was used, but this should be clarified in the paper.

The appropriate annotation has been made regarding the embeddings file used for the embeddings development. The Common Crawl has been selected due to the largest number of both tokens (840B) and dimensions (300). Still, the future exploration of other embeddings options forms an opportunity for our current study extension.

9. For the deep learning procedures, how were training vs. test sets allocated? How many tweets were in each set what is the justification (based on power analysis or past research norms) for the current allocation strategy?

The 5-fold cross-validation was followed, with 80:20 training (3,318 posts) and testing (830 posts) split as in [1].

10. The authors refer to this as a "case study" in the abstract and also within the text. This classification doesn't seem appropriate, especially as the paper presents a technical approach to assessing population trends. In a clinical research sense, "case study" entails a particular study design with an N-of-1 participant evaluation. This designation doesn't seem appropriate nor helpful for framing the current study. The approach might be more appropriately called a "validation study" of deep learning procedures that are presented in similar work. Even if the underlying procedures aren't entirely novel (i.e., have been used in similar studies), presentation of a thorough methods narrative and data considerations (if reviewer concerns are addressed) seem to be a valuable contribution to the current literature. The additional use of weather data would certainly entail a validation component to this work.

We used the term ‘case study’ more in general rather than clinical research terminology, i.e. case study as the particular approach application (Deep Learning) to the problem of interest (Hay Fever surveillance). We understand that the population health estimation involves the participants sampling, based on specific criteria. In our case, the community of Twitter users can be considered as such sample to generalise over the wider population. Still, the criterion might not seem strong enough from the clinical research perspective when applied to condition prevalence evaluation. To avoid any potential confusion and to closer reflect the nature of the experiments conducted, the ‘case study’ terms have been removed from the manuscript. We greatly appreciate your suggestion.

11. The Results: "Weather correlation" section could be strengthened by reporting correlation coefficients, rather than simply referencing the included graph visualizations. As the graphs don't appear to be highly concordant between Evaporation/Humidity vs. tweets, some normalization procedures may be warranted for calculating inferential statistics. As a reader, I would want to know if these trends are objectively concordant (i.e., statistical inference reporting).

The relevant Pearson’s coefficients along with the p-values have been added and highlighted in the manuscript (Weather correlation section). The normalisation procedure has also been performed for calculating the inferential statistics. The full list of coefficients for all weather variables has been included in the Appendix 1.
12. Given the many technical acronyms used in this paper, a comprehensive index of acronyms and perhaps brief contextualizations (e.g., GloVe: "Global Vectors for word representation" embedding framework, GRU: "Gated Recurrent Unit" machine learning approach) would be a helpful reference point for readers.

Following your suggestion, the comprehensive list of acronyms has been compiled.

Abbreviations
ADR – Adverse Drug Reactions
AIHW – Australian Institute of Health and Welfare
AR – Allergic Rhinitis
ASCIA – Australasian Society of Clinical Immunology and Allergy
CNN – Convolutional Neural Network
DL – Deep Learning
ERP – Estimated Resident Population
GloVe – Global Vectors for word representation
GRU – Gated Recurrent Unit
HCP – Health-Care Professional
HF – Hay Fever
LSTM – Long-Short Term Memory
ML – Machine Learning
NLP – Natural Language Processing
RNN – Recurrent Neural Network
WHO – World Health Organisation
13. The "Ethics approval" section simply states "Not applicable." Was this research protocol reviewed by an Institutional Review Board or similar ethics panel? This would seem important as the study tracks health status among Twitter users, which has confidentiality and privacy implications. If the protocol was reviewed and deemed exempt from further review or IRB oversight (e.g., not human subjects research), then it would be useful to clarify this aspect. "Not applicable" makes it sound like the research was not sent for critical ethics review.

Thank you. Our previous work in similar domain (health-related knowledge extraction from Social Media [1]) did not entail the ethical approval requirement, thus the approach followed.

Still, we sought the advice from the Associate Professor Nicola Reavley, Ethics Committee Member, Melbourne School of Population and Global Health. Prof. Reavley published a study on mental health assessment using Twitter data. The posts were openly released without the ethical approval [2]. Also, in [7] the authors observed ‘Twitter data are often relegated to nonhuman participants research, because collected tweets are publicly available’ (the necessary measures to reduce the risk of users’ re-identification were yet recommended).

We further identified the studies of Social Media data mining in public health context, which also did not report any ethical considerations. Please see [3], [4], and [5]. Still, we absolutely agree with your concern regarding the potential privacy issues of the users. Hence, the number of studies has been reviewed from the perspective of any mitigation steps implemented to avoid users’ re-identification. In [6], the authors stated that ‘it is common practice in social computing research to analyse publicly available data without the posters’ consent or knowledge’, and ‘to reduce risk of participants’ data being resurfaced, we changed quotes slightly and used paraphrasing to obfuscate posters’ data, a common method in ethical social computing work’.

As a result of your suggestion and further research, all of the posts included in the manuscript have been obfuscated (i.e. paraphrased, shortened). The obfuscation was mostly applied to the words/phrases irrelevant from Pollen Allergy surveillance point of view. The original form of the symptoms/treatments mentions was retained though (any modification/removal could reduce the value of study, which focal point is the actual language used on Social Media platforms to report the Hay Fever instances).

Also, we believe that Pollen Allergy can be considered less of a sensitive topic in comparison with the multitude of Social Media-based studies on mental health assessment, violence victims’ identification, or abuse incidents mining that use social media as a data source. In the examples mentioned, potential users’ re-identification could carry an undesirable impact on social functioning of the sufferers/victims. Still, the obfuscation was applied to alleviate any potential privacy concerns. The information about obfuscation has been included in Table 4 footnotes of the manuscript.


[3] Doan et al., ‘Extracting health-related causality from twitter messages using natural language processing’, BMC Medical Informatics and Decision Making (2019) https://doi.org/10.1186/s12911-019-0785-0 (Ethics approval and consent to participate: This research is not human research and did not require IRB approval.)


Reviewer reports

Hansi Zhang, MS (Reviewer 2): 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.

Thank you very much for your suggestion. As the study lies at the intersection of multiple research areas (i.e., allergy surveillance, text classification, social media, deep learning), we reviewed the relevant areas in separate paragraphs. The approach followed the work [1].

Overall, Deep Learning validation to Hay Fever detection from User-Generated Content has not been investigated before. Previous works that applied Deep Learning, but in different context, have been highlighted in the manuscript accordingly. The promising results obtained served as a motivation for the approach evaluation on equally challenging Pollen Allergy detection from Social Media.

Also, to emphasise our contributions, the following paragraph has been added at the end of the Background section.

‘Contributions

The main contributions of the study can be stated as follows:

• We introduce Deep Learning application in the context of Pollen Allergy surveillance from Social Media in place currently dominant conventional Machine Learning classifiers;

• We focus on challenging informal vocabulary, which leads to condition under/over-estimation if unaddressed in place of the traditional keyword/lexicon-based approaches;

• We propose the fine-grained classification into 4 classes in place of the most common binary classifiers, i.e. Hay Fever-related/Hay Fever-non-related;

• We enrich the data with an extensive list of weather variables for potential patterns identification, where previous studies focus mainly on Temperature, and Pollen Rate.’
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.

Initially, we were exploring multiple options for relevant posts extraction, emphasising either high precision (‘hay fever’ OR ‘hayfever’), or high recall (‘sneezing’, ‘runny nose’, ‘red eyes’, ‘itchy throat’, and so on). As a result of search queries extension, a large proportion of false positives was collected. The Hay Fever symptoms were often mentioned in the context of other conditions such as Cold, Flu, Cats/Peanuts Allergies, etc.

Due to the informal nature of conversations on Social Media platforms, we have excluded the medical term ‘Allergic Rhinitis’ from the search keywords list. Official Allergic Rhinitis expression also occurs most frequently in non-relevant posts (e.g. advertisements/news), which only adds noise to the dataset. Thus, we have prioritised the high-precision approach, including word embeddings to capture as much relevant data as possible, while accounting for lexical variety.

Further clarification on the keywords selection approach has been added and highlighted in the manuscript.

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?

Thank you. The Table with datasets’ specifications for embeddings development has been added in the manuscript (Table 1).

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).
Please find below the relevant information:

Dataset 1: 4,148 (100.0%) (Alice Springs + 2,000mi radius) - Text Classification

Dataset 2: 1928 (60.4%) Melbourne, 1040 (32.6%) posts Sydney, 222 (7.0%) posts, and Brisbane - Weather Correlation

Geo-location is available for Dataset 2.

The manuscript has been updated accordingly (Methods – Data annotation section).

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.

We absolutely agree, thus the obfuscation of the raw text has been applied to reduce the risk of potential users’ re-identification. The posts have been either shortened or paraphrased. We have ensured that the relevant words/phrases remained intact (e.g. symptoms/treatments) not to diminish the value of the study (natural language exploration in the context of Hay Fever detection). The information about the obfuscation implemented has been added in Table 5 footnote.

Also, please find the examples of the posts before/after obfuscation in the Table below.

<table>
<thead>
<tr>
<th>Original post</th>
<th>Obfuscated post</th>
</tr>
</thead>
<tbody>
<tr>
<td>'I think I’m finally getting hay fever sob pls no', 'I’m getting hayfever sob no thank you'</td>
<td>'I’m getting hayfever sob no thank you'</td>
</tr>
<tr>
<td>'But now my hay fever is in overdrive. Sniff sniff.', 'My hay fever is in overdrive today. Can’t stop sniffing'</td>
<td>'My hay fever is in overdrive today. Can’t stop sniffing'</td>
</tr>
<tr>
<td>'I have terrible hayfever. Terrible smells have no effect on during November-January', 'I have an awful hay fever. Bad smells do not affect me during pollen season'</td>
<td>'I have an awful hay fever. Bad smells do not affect me during pollen season'</td>
</tr>
<tr>
<td>'Hay fever season soon. Time to stock up on antihistamine', 'Time to stock up on antihistamine as hay fever season fast approaching'</td>
<td>'Time to stock up on antihistamine as hay fever season fast approaching'</td>
</tr>
</tbody>
</table>
6) For the 4 categories of your annotation classes. How do you consider information posted by doctors, nurses, and other professionals?

As the study aims to accurately estimate the prevalence of Hay Fever within population, the content posted by healthcare professionals have been classified as Non-relevant (unless the doctors/nurses were referring to their own symptoms/treatments).

Please see the example of post of informative/advisory nature, excluded from the prevalence estimation:

A useful @username article discussing use of antihistamines in allergic rhinitis (hay fever). it also inc. evidence of isotonic nasal salines helping reduce antihistamine reliance, improve symptoms and quality of life. (non-informative (marketing))

We hope this clarifies the point raised.

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.

Thank you for your suggestion. Further clarification on pre-processing procedures has been added in the Training and Testing section of the manuscript. Please find below:

‘Pre-processing performed was minimal, and included removal of URLs, non-alphanumeric characters and lowercasing. In terms of emojis, their numerical representation was retained, following the punctuation removal. No excessive pre-processing was applied as models perform the operations on sequence of words in order they appear. Words are preserved in their original form without stemming/lemmatising due to their context-dependent representation, e.g. 'allergy', 'allergic', 'allergen'. Also, Sarker et al. suggested that stop words can play a positive effect on classifier performance. Analogical pre-processing steps were implemented for the embeddings development.’

We have assumed that tweets extracted using English keywords would be mostly in English. Thus, no additional treatment of non-English characters was applied (unless the non-alphanumeric characters were included).
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.

The abovementioned sentence referred to the embeddings development, for which the dataset of 22k posts was used (considered as moderate size). We have replaced the term ‘training’ with ‘development’ to avoid any potential confusion with the actual classifier training (dataset of 4k posts).

Please see the change below:

‘The application of HF word embeddings did not improve the performance of the classifier, what can be attributed to relatively moderate dataset size of (20k posts) used for their development.’

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,"

The indices have been moved to the front for consistency, thank you for your suggestion.

10) The weather correlations are visualized in figures. What about also providing the correlation using Pearson correlation coefficient?

The relevant Pearson’s coefficients along with the p-values have been added and highlighted in the manuscript (Weather correlation section). The full list of coefficients for all weather variables has been included in the Appendix 1.