Dear editor,

We would like to thank the editor and the reviewers for the time spent reviewing our paper and for the comments they have provided which have, in our opinion, helped us improve the paper and clarify some points that were not well explained in the previous version.

Our point-to-point answers to the reviewers' comments are shown as follows.

Please do not hesitate to contact us should you require any further information.

Comments to the Author

In this study, authors take a machine learning based approach in the emerging field of entomology to tackle the challenge of surveillance in public health for early detection of infectious diseases epidemics.

They have used Twitter data for this purpose. Overall the manuscript is written very sound and provides clear and easy to read descriptions of concepts such as Infodemiology and Opinion Mining which are essential for telling the work later on. Also, the reason for using Twitter is well justified. So I appreciated the efforts of the authors on this regard.

Although i find the article generally easy to read, innovative and well written, i have five major comments, followed up by a few minor points/suggestions for the authors:

We would like to thank the reviewer for this encouraging comment.

In the sections 3.2 and 3.3 there are various figures and tables that shows the behaviour of RF and SMO models with respect to using different corpus and different char-grams and word-grams. Most of these figures the text is merely describing what the graph is already showing. I would expect more context and explanatory information for the reader to relate to and interpret the findings than merely describing the graphs.

These sections have been extended including more explanatory information about the tables and figures.

The graphs showing for various word/char - gram for different classifiers are certainly not surprising as from any machine learning tasks that there is learning in the certain phase and plateau in the second phase and decrease with overfitted models. The important finding is common findings and where two Machine learning modalities actually differ and why?

We agree with the reviewer in this point and we have added a brief discussion about the results of different classifiers when the tables are explained.

Next is about validation. The manuscript describes a methodology but not clearly validates it. Since they are going for binary classification where a random classifier has %50 accuracy, some instances %67 and a maximum of ~%90 is certainly a good starting point but not very impressive yet. But this is partially due to the complexity of the phenomena the authors chosen.

As it is stated by the reviewer, the application domain of our classifiers is a very complex domain. Besides, we performed different sets of experiments with different corpora size and that is the reason why the accuracy varies through the experiments. However, it remains stable with the same corpus size in the experiment. With all the tests in different corpora we tried to discuss how the results changes with the corpus size and quality. The better results are

obtained with the corpora that have been manually tagged by more users (that is corpus-3 and corpus-4). Besides, it can be observed that including more tweets the two models get worse results that means that it is possible that other models such as the ones based on psycholinguistic features could obtain better results. This discussion has been added to the section 4.

For this I would suggest a much more direct approach and pick a topic that they can validate. For example weather: apply the methodology and categorize the weather attributes. Since the weather data for any regions is available, they can first predict the weather via tweets and check with the actual weather reports across certain time periods where & when the tweets are generated their accuracy. This way we have a good idea that this is a sound methodology that we can also trust for infectious diseases.

We would like to thank the reviewer for this comment. We agree that our study can be applied to different domains such as the weather classification and it would be very interesting to show these results, but the authors think that this study does not fit with the paper very well and the Special Issue where the paper was submitted "Intelligent Health Services Based on Biomedical Smart Sensors". However, we have included as future work this study.

The following point is I understand the difficulty in coming up with a good corpus size and content for machine learning purposes but i believe the article should provide more details about practical applicability of such an approach in practice. After reading the article I have more questions in my mind than answered in the text

In the introduction section we have added more information about the motivation of infodemiology systems and the experiment performed in this paper.

Is the corpus size or word/char - gram is dependent on the language? If yes can we still learn from one study to actually transfer what we learn to another language?

This question has been answered now in the discussion section with the following comments: (Section 2.2)

BoW model […] it is not truly language-independent. Language independence means that, once an algorithm is developed, it can trivially be extended to another language as long as we have available a large amount of training data. But in the BoW model, this is only true for non-agglutinative and western languages, such as Spanish, English, Italian or Portuguese, where words usually have a single inflectional morpheme to denote multiple grammatical, syntactic, or semantic features. In agglutinative languages, on the other hand, complex words are formed by stringing together morphemes without changing them in spelling or phonetics which hinders the frequency of the word-grams and it suffers from the curse of dimensionality, because it tends to create over-fitted models due to the vocabulary size. This problem is aggravated in agglutinative languages because it produces a large number of word forms that can be obtained from a single root.

Char-grams presents some advantages over the word-grams model: First, char-grams capture (1) lexical information, such as the use of punctuation symbols, and (2) morphological information, such as prefixes or suffixes. Second, char-grams are robust against grammatical errors, because a misspelled word and its correct version should share common char-grams. Third, char-grams can deal with agglutinative languages because it is capable of extract the individual phonemes from compound words.

How could such a first line information system be used to inform public health authorities? For example why the authors might consider to make an online tool that can analyze the Twitter activity from the last week and generate graphs about the upcoming weeks/months. Later they can follow this up with the actual disease prevelance data for example such data usually exists for flu seasonally.

In the Conclusions and Future Work section we have included more information about Collaborative Health Project, that partially funds this project. As the reviewer suggests, this project consists in the creation from an infodemiology system. One of the modules of this project is a dashboard which will be capable of generate graphs and other valuable information for health-care professionals.

The fifth comment is can the authors give guidelines for creating corpus that could be applied in other domains in a clear to implement point-by-point way so the manuscript would be of interest for a larger public.

We would like to thank the reviewer for this comment. We have included in Section 3 information about the tool we have used to compile and tag the corpus. This section includes a discussion about the configuration parameters that are used (latitude, longitude, radius and keywords) in the study. In addition, we have included a more detailed description of the process concerning how the volunteers performed the task and how the authors monitored the process.

Minor comments:

Line 161 ..."that distinctly "separates" the classes in a data-set"...

This typo has been fixed.

I would mention Twitter in the abstract/ title since actually this is the data source and an important part of the study.

This information has been now included in the abstract.

Again the abstract i would give the number of tweets analyzed, used for classifief etc. to give an idea about the content of the article.

This information has been now included in the abstract.