Author's response to reviews

Title: Improved de-identification of physician notes through integrative modeling of both public and private medical text

Authors:

Andrew J McMurry (mcmurry.andy@gmail.com)
Britt Fitch (britt.fitch@gmail.com)
Guergana Savova (guergana.savova@childrens.harvard.edu)
Isaac S Kohane (isaac.kohane@harvard.edu)
Ben Y Reis (benreis@gmail.com)

Version: 4 Date: 17 September 2013

Author's response to reviews: see over
September 16, 2013

Dear editorial office of BMC informatics and medical decision making:

Please see below our response to reviewers for the submitted manuscript titled “Improved de-identification of physician notes through integrative modeling of both public and private medical text”. We have carefully considered the reviewers comments and addressed in detail below.

Kind Regards,
Andrew McMurry, MS
Reviewer's report
Title: Improved de-identification of physician notes through integrative modeling of both public and private medical text

Version: 3 Date: 19 August 2013

Reviewer: Murat Kantarcioğlu

Reviewer's report:
In this paper, authors discuss their machine learning based techniques for removing identifying information in medical notes. One novel aspect of this work is that they leverage the existing medical literature to model non-identifying medical terms. After this step, they use standard natural language processing and machine learning techniques to build models to classify terms in medical notes as PHI or non-PHI.

Minor Revisions: Authors addressed most of comments but "the scored reported in this paper would have placed first overall in recall and last overall in precision.". It is not than in many cases you can increase recall by lowering precision and vice versa. One solution to address this issue is to use area under the curve measure (ROC AUC). It would be good to use AUC to compare against all the existing techniques.

Response In an effort to produce a fair comparison of performance the authors have used the same reporting as found in papers reviewing the original i2b2 challenge (precision, recall, f-measure). The authors further agree that in some cases one could 'increase recall by lowering precision and vice versa', our statements regarding standings of precision and recall are intended to give the reader information regarding our theoretical placement among the challenge participants given our personal IRB-driven motivation.

Level of interest: An article whose findings are important to those with closely related research interests.

Quality of written English: Acceptable

Statistical review: No, the manuscript does not need to be seen by a statistician.

Declaration of competing interests: I declare that I have no competing interests.
Reviewer's report

Reviewer's report
Title: Improved de-identification of physician notes through integrative modeling of both public and private medical text

Version: 3 Date: 10 August 2013

Reviewer: Joel Martin

Reviewer's report:
I believe that the authors have adequately addressed my concerns. This paper will be useful for other researchers because it carefully considers multiple sources of general features.

Discretionary Revisions
I think the authors do a better job of motivating the focus on Recall. It did seem unnecessary to mention F10, because it is for all practical purposes the same as Recall.

I found the new discussion about bigrams and trigrams to be a little dissatisfying. I agree that unigrams are probably more robust and that you are already doing very well. However, what if bigrams and trigrams raise the precision without affecting recall? How much time would it take to try that? My guess is that it would only take a few hours to two days to try it.

Response
Thank you Dr. Martin for your suggestions on N-grams. We reviewed the code and determined it would be considerably more effort because of the highly parallel and non-sequential processing of the tokens. Nonetheless, we spent considerable time to fit a sequential token analysis of the tokens as you suggested using a standard CRF (conditional random field). Unfortunately the sequential model never had strong tendencies to over-fit the training data. After careful review of the misclassifications it became clear that the sequential model was nearly perfect when the test instances and the train instances were nearly identical. However, when the test data was characteristically different from the train data the model was unstable. Your suggestion was a good one and if we ever are able to find a larger annotated cohort we will try it again. At this point we have invested more time and resources than were budgeted for that analysis. It would be interesting to revisit the idea in future work.

Level of interest: An article whose findings are important to those with closely related research interests

Quality of written English: Acceptable

Statistical review: No, the manuscript does not need to be seen by a statistician.

Declaration of competing interests: I declare that I have no competing interests.
Reviewer’s report

Reviewer's report
Title: Improved de-identification of physician notes through integrative modeling of both public and private medical text

Version: 3 Date: 10 August 2013
Reviewer: Oscar Ferrandez

Reviewer’s report:
The revised version of the paper is very well-written and structured, and comments from the reviewers have been addressed thoroughly.

- Major Compulsory Revisions
  The authors’ response and changes were acceptable.

- Minor Essential Revisions
  The authors’ response and changes were acceptable.

- Discretionary Revisions
  The authors’ response and changes were acceptable.

Level of interest: An article of importance in its field

Quality of written English: Acceptable

Statistical review: No, the manuscript does not need to be seen by a statistician.

Declaration of competing interests: I declare that I have no competing interests