Reviewer's report

Title: Automated Systems to Identify Relevant Documents in Product Risk Management

Version: 1 Date: 28 August 2011

Reviewer: Courtney Corley

Reviewer's report:

1. The author's present a methodology to automatically classify health product literature into "useful" or "not useful" in determining risk. Two systems are developed, one based on general predictors the other on product specific ones (TNF-alpha). Generally, the hypothesis is understandable; however, the problem statement is not well formed. For example, have what are the specific health risks and how is literature determined to be useful. Granted, the authors do present sufficient motivation in the background section, literature review, and provide a complex algorithm of how usefulness is decided; though the specific problem to be solved is not adequately described in the text. The author's vocabulary is challenging to parse and understand. Product risk management is a broad field that encompasses more than health products, what is "primary literature?"

2. There are significant weaknesses and not novel methods in the research. Term normalization (morphological reduction), removing stem words, key-term extraction, term weighting schemes are not novel and should be given minimal treatment in the text, similarly, Figure 3 is not needed - also be careful not to confuse word occurrence with term phrase occurrence. TF-IDF and binary schemes are not strictly frequency measures, they are a measure of term weight.

Technically, the terms are not predictors but features to the machine learning algorithm used to build a model. Why do you provide the general terms but not the specific terms? Do you extract general terms from the general data set or both? What about specific terms, do they come from the specific data? If not how do you determine they are relevant to TNF-alpha risk assessment. 730 general terms are reported, however, that seems too low for a corpus of nearly 4200 abstracts.

Only the term weights of the predictor terms were chosen as features in building the supervised model. There are many other features that might be useful to this task and could significantly improve performance of both the general and specific systems. Syntactic (e.g., parse trees, part-of-speech), semantic (e.g., UMLS), or other rich features.

What is the motivation for using the Gaussian kernel, oftentimes radial kernels produce the best results on text analysis. The explanation of SVM is very good, if you choose to explain this classifier - then describe each of the techniques list in
the tables. A short two or three sentence description is sufficient.

3. The data collection method is not sound, it is extremely subjective and not clear which data annotation are automated and which manually. How is the usefulness algorithm developed, who developed it, what are the underlying assumptions, has it been validated, is it specific to TNF-alpha? Please report how many annotators reviewed the articles, what was their agreement (e.g., Kappa statistic). The number of abstracts in the specific data-set is an order of magnitude larger than in the general data-set. The variance between them is too large, specific and general is because of the small "N" of the generalizability data set, I suggest increasing the generalizable data set. Other comments on the method appear in remarks above.

4. The manuscript adheres to the relevant standards for reporting and data deposition.

5. The authors should compare the various weighting schemes across each of the supervised methods. That is, TF, IDF, TF-IDF for logistic, k-means, etc. too. It is difficult to assess the system performance in the general automated system. What is meant by the case where the general system was not used? The terminology becomes confusing, do you use the general terms or specific terms for the generalized system? After reading the article many times, I do not understand what is meant by the first 10% or 20% of articles presented to reviewers.

The section "specific automated system" in results, does not make sense, the document selection technique is confusing. Why is the training set size only two for 1:1 useful/non-useful articles? The ROC AUC results are interpretable with such a small "N." What is the justification for such small N?

Performance statistics are listed throughout the article for the general system but not the specific system.

6. The author's do an OK job presenting the potential limits of the work. However, there are many other limitations, based on this reviewer's previous comments.

7. The authors clearly acknowledge any work upon which they are building, both published and unpublished.

8. The title accurately convey what has been found. However, the abstract is not representative of the work. See comments below.

9. The manuscript needs to be edited and reviewed by a professional technical editor to compensate for English not being the first language of the authors.

Abstract

10. What is the problem statement/hypothesis, what is the basic premise?
11. What are health products, old and newer? provide an example of each, and the types of risks this type of system would mitigate.
12. the sentence ending "useful for risk assessment work." is awkward and not clear.
13. What is meant by work efficiency?
14. How do you define useful? provide a sentence or two on what is considered useful
15. How many specific automated systems were created? The text only describes one, for TNF-alpha. What are the four other drug classes? This is not captured in the text.
16. Are evaluator's people or automated systems
17. what is 10% and 20% of articles, does not make sense why this would happen
18. Provide the performance results for the specific systems too, (not just general).
19. How was the specific system better? it is a strong statement not backed up by data.

Background
20. What are other types of product risk management, and have similar systems been developed?
21. How were the 700 abstracts containing valuable information determined? Who determined it?

Methods
22. How many annotators reviewed the text and what was the Kappa statistic?
23. Who developed the algorithm, is it reported elsewhere since it is in the appendix? What are the underlying assumptions.
24. were the 693 articles determined to be useful, automatically or by a human annotator?

Text Mining
25. How were the specific and general predictors selected? By human or automatically, is the algorithm similar to choosing useful abstracts?

Specific automated system
26. The last two sentences do not make sense, need to re-write for clarity

Measuring prediction performance.
27. Need to better describe that AUC is from a ROC curve

Results
28. The description of Figure 5 is not clear, what is the case where general
automated system was not used. What is the line graph, confidence, how was confidence calculated,

Discussion
29. The first paragraph of the "general automated system" is not clear what you mean. the terms general and specific and generalized are confusing.
30. PG 17 (first full paragraph). How and why is this done,
31. the subsection title "potential application of automated..." is not representative of its content.

- Minor Essential Revisions
32. PG 20. typo - text is superscript after "[39]"

Level of interest: An article of limited interest

Quality of written English: Not suitable for publication unless extensively edited

Statistical review: Yes, and I have assessed the statistics in my report.

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