Reviewers report

Review 1

RECOMMENDATION: WEAK ACCEPT

There are so many problems with this manuscript, one hardly knows where to begin...

Clearly, the authors are not very familiar with academic writing. There are hardly any details given on the corpus collection, the annotation method is flawed and the classification process is not really described. What labels are you using? What are ”M & S” and ”H & M”? Is the data available to the community? Even the references are inadequate.

In conclusion, this manuscript was likely submitted by a student before their supervisor had the opportunity to approve it. I suggest the authors revise the paper thoroughly and seek the assistance of senior colleagues before considering a re-submission.

Review 2

RECOMMENDATION: STRONG REJECT

This paper addresses the classification of conversation transcripts using neural networks. Overall, the presentation of the corpus is lacking in details. The description of methods and positioning of the work could also be improved.

Strengths:

- the authors developed a corpus of transcribed support group conversations (although there are serious concerns about the ethics of the process, see below).

Weaknesses:

- What is the research question question addressed in this work? The authors should consider articulating the question, and explaining how the proposed dataset and method are relevant to their question.
- the authors make a number of unsupported claims, such as ”rule based systems are more expensive than machine learning”. Please consider providing evidence to substantiate claims.
- the description of the dataset could be improved by providing additional information: how many ”texts” were collected? What is the size of the corpus? What kind of annotations were done? What is the distribution of annotations overall? More importantly, the overall protocol seems questionable, in particular the collection of human subject data without consent and the unnecessary introduction of identifying information are not consistent with current legislation (e.g. HIPAA or GDPR).
- A number of studies in the paper are referenced inappropriately. For example, Fort et al. do not support the authors’ claim that ”AMT should be used because it is inexpensive.”
- There are no technical details on the classification method used. Which libraries/tools/parameters were used?
- The evaluation measure used is not defined.
- Results are reported with 9 figures after the decimal point; this seems unnecessary and indicates that differences are likely not significant. In addition, the neural network training process is known to be non-deterministic and performance difference between the lowest performing run and the best performing run of a single system can, in some cases be as high as 2 points as evidenced by Reimers and Gurevych (EMNLP 2017).
- the reference list seems to rely heavily on the work of a limited set of authors, published in non peer-reviewed venues.

Other comments:

- the authors could consider using a more specific title
- the abstract does not convey a clear idea of what was done: What is the task addressed? What is the method used? What are the results obtained?
- the first sentence of the paper directly refers to the authors’ previous work, therefore voiding the anonymity of the submission.

Overall, both the experimental protocol and presentation of the work raise important issues that need to be addressed.

Review 3

RECOMMENDATION: STRONG ACCEPT

The authors present a study that classified support group meeting transcripts with a number of deep learning methods, including convolutional neural networks, recurrent neural networks, and combinations of the methods. The approaches are interesting and clearly successful. However, it would be nice to have a discussion of the merits of the different methods used. For example, RNNs are known to be better than CNNs for taking context into account. It would be useful to have a direct comparison of RNN and CNN. If the authors could include this particular result, it would be a great improvement on the paper.

Review 4

RECOMMENDATION: REJECT

This paper uses deep learning for the classification of patient authored text. I have major concerns regarding the methods used to collect the corpus, conduct the experiments and write the report.

Major comments:

- Was this study approved by an IRB or equivalent? This is mandatory for work conducted on patient data
- Link to the code seems unrelated to the methods described
- the baseline system is not described
- Section 4 seems unrelated to the rest of the paper, and wording is identical to a previously published paper

Minor comments:

- a more appropriate reference for word2vec would be: *Mikolov T et al. Distributed Representations of Words and Phrases and their Compositionality. NIPS 2013: 3111-3119*
- in section 4.1, ”peformance” should read ”performance”.