Review of: "Comparative Analysis of Machine and Deep Learning Techniques for Text Classification with Emphasis on Data Preprocessing"

Scythia Marrow

Potential competing interests: No potential competing interests to declare.

This article is a comparison of different standard, well-established machine learning algorithms on the Titanic and Titanic2 datasets. All of the algorithms presented have been evaluated on these datasets before, so this paper does not present any novel information. However, that in and of itself should not prevent publication, as replication studies are important. This is the core thesis of the work and the strongest part of the article.

The article suffers from several egregious errors and cannot be recommended for publication.

The Titanic datasets are not text classification datasets. They contain no textual data. None. The article should not be talking about text classification because no text algorithms are evaluated. BERT is not relevant at all. Don't talk about BERT. This is fatal flaw number 1.

The article only explains the BiLSTM algorithm in detail, which is misleading as it suggests that BiLSTM is the author's work even when it is not. Passing others' work off as your own is plagiarism and is not acceptable. The authors did not create the BiLSTM and so should either clearly explain ALL algorithms they are comparing or none of them. The authors should make it abundantly clear that they are not performing novel work and are simply doing a replication study. This is fatal flaw number 2.

The authors talk about the importance of preprocessing, but we are not given a reason for why. Algorithms are only evaluated with preprocessing steps; no comparison is made between algorithms trained on raw data and algorithms trained with preprocessed data. This is another massive problem and is fatal flaw number 3.

In addition to the fatal flaws, the article also contains numerous stylistic, grammatical, and structural errors. It reads as if two different articles have been smashed together without care. I would not be surprised if they had, as a way to get around a plagiarism filter.

To salvage an actual paper out of this, the authors would need to:

1) Produce an actual, somewhat novel scientific thesis. “BiLSTM is SOTA on Titanic2” is a thesis, yes, but it is not novel in

any way. “Preprocessing data is good for <X> class of algorithms and not necessary for <Y> class of algorithms" would be a better thesis that is actually somewhat novel and could be investigated using the authors' current methods.

2) Support said thesis clearly and concisely. In the case of the somewhat novel preprocessing thesis, they would need to compare algorithms with and without preprocessing to see if there is any variation across algorithms.

3) (Optional) If there is no support for that somewhat novel thesis in your data, that's fine. Clearly indicate that you found no evidence for it and indicate which studies you just replicated. That's still publishable. What you have now is not.