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

Angel Sanchez[1]

1 Universidad Rey Juan Carlos

Potential competing interests: No potential competing interests to declare.

This manuscript describes a comparative analysis of several Machine Learning and Deep Learning techniques for a considered Text Classification problem. The “Titanic Disaster Dataset” was used for the experiments with the aim to predict the patterns and relationships among the demographic characteristics of the passengers (i.e., gender, age, ticket fare, ...) and their respective survival outcomes.

An important part of the proposed methodology for improving the prediction results is data preprocessing performed before the experiments. It consisted of data cleaning, missing data value imputation, among other tasks.

After that, six different classification models were tested: Naïve Bayes (NB), Gradient Boosting (GB), Support Vector Machines (SVM), Conditional Random Fields (CRF), Recurrent Weighted Average (RWA), and Bidirectional Long Short-Term Memories (biLSTM), respectively. Experimental results show that the BiLSTM model outperformed the other text prediction techniques.

In general, the paper is well organized and also well written.

For the sake of manuscript legibility, I suggest the author include a table including the names and acronyms of all classification techniques mentioned along the text.

Additionally, the author should also include some more explanations in the manuscript about the following aspects of the proposed solution:

1) Provide specific examples of how the data preprocessing steps improved the model's performance.

2) Point out additional details on the comparative performance metrics of each model, beyond just stating that BiLSTM outperformed the other approaches.