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

Muhammad Hammad Memon[1]

1 Southwest University of Science and Technology

Potential competing interests: No potential competing interests to declare.

The article titled "Comparative Analysis of Machine and Deep Learning Techniques for Text Classification with Emphasis on Data Preprocessing" by Dr. Saikat Gochhait:

Overview

The article presents a detailed comparative analysis of machine learning and deep learning techniques for text classification, focusing significantly on data preprocessing methods. It employs several algorithms, including Naive Bayes, Gradient Boosting, Support Vector Machine, and deep learning architectures like BiLSTM combined with CRFs. The primary dataset used is the Titanic Disaster Dataset, which is an interesting choice for demonstrating text classification methods, given its structured nature.

Strengths

1. **Comprehensive Methodology:** The study's methodology is robust, employing a variety of classical and modern techniques to address text classification.
2. **High Accuracy:** The reported accuracy of the BiLSTM model (98.5%) is commendable and suggests a well-tuned model.
3. **Detailed Experimental Setup:** The article provides a clear description of the experimental setup, preprocessing steps, and model evaluation metrics, which enhances the reproducibility of the research.

Areas for Improvement

1. **Dataset Relevance:** The use of the Titanic Disaster Dataset for text classification is unconventional since it is primarily a structured dataset used for survival analysis. The connection between this dataset and text classification tasks is not clearly established, which might confuse readers about the applicability of the methods to actual textual data.
2. **Comparative Analysis Depth:** While the paper presents a comparison of various models, the discussion could be deepened to include more on why certain models performed better than others beyond just their architectural advantages. Insights into feature importance or model decisions could be valuable.

3. **Literature Review:** The related work section could be expanded to include more recent studies, particularly those that employ similar datasets or methodologies. This would position the paper more firmly within the current research landscape.
4. **Methodological Justifications:** Some methodological choices, such as the specific configurations of deep learning models and their layers, are not fully justified with experimental or literature-backed reasoning. Providing more details on these choices could enhance the academic rigor of the paper.

Suggestions for Revision

1. **Clarify Dataset Usage:** More information should be provided on how the Titanic Disaster Dataset was used for text classification. If textual data was derived from this dataset, the process and rationale should be explicitly stated.
2. **Expand Comparisons:** Include additional metrics like precision, recall, F1-score, and confusion matrices for all models discussed to provide a more rounded view of their performance.
3. **Enhance Theoretical Background:** A deeper theoretical exploration of the models used, particularly the BiLSTM and CRFs, would help in understanding their specific contributions to the accuracy improvements noted.
4. **Update Literature Review:** Incorporating a broader range of recent publications would make the study more relevant and might provide additional comparisons or benchmarks for the models used.

Conclusion

The article is well-constructed with a clear focus on data preprocessing in text classification. However, clarifying the use of the dataset, expanding the comparative analysis, and enhancing the literature review could significantly improve its impact and usefulness in the field of text classification.