Predicting Numerical Ratings from E-commerce Reviews Using Advanced Machine Learning Models

Abstract: This study explores the application of advanced machine learning techniques for sentiment analysis, aimed at predicting numerical customer ratings from e-commerce reviews. The "Amazon Product Review Dataset" from the University of California, San Diego, was utilized to evaluate the performance of Convolutional Neural Networks (CNN), Long Short-Term Memory networks (LSTM), and Transformer models. Data preprocessing involved text cleaning and handling of missing values, followed by model training and evaluation using metrics such as Mean Absolute Error (MAE) and Root Mean Square Error (RMSE). The findings indicate that the LSTM model excels in capturing long-range dependencies in textual data, yielding more accurate sentiment predictions. The CNN model also shows strong performance by effectively identifying local patterns in text. Despite the potential of the Transformer model, it underperformed compared to expectations, suggesting a need for further tuning. Additionally, specific words and features that significantly influence customer ratings were analyzed, offering insights into consumer satisfaction and preferences. For instance, terms like "comfort" and "price" have a notable impact on review scores. This research highlights the capabilities and limitations of each model, contributing to the development of more nuanced sentiment analysis tools to help businesses better understand and enhance consumer satisfaction. Future work will focus on refining these models, exploring additional features, and incorporating sophisticated attention mechanisms to improve prediction accuracy and model interpretability.

Keywords: sentiment analysis, e-commerce reviews, machine learning, CNN, LSTM, Transformer, Amazon Product Review Dataset, MAE, RMSE.