

E-Document Classification Using Deep Learning

M.W.A.S.P. Wijesuriya¹ and R. Sakuntharaj²

¹*Department of Computing, Faculty of Science, Eastern University, Sri Lanka*

²*Department of ICT, Faculty of Technological Studies, University of Vavuniya, Sri Lanka*

Abstract

The proliferation of digital content in the form of electronic documents necessitates efficient classification methods to manage and analyze this growing body of information. This research explores the application of advanced deep learning mechanisms, specifically Convolutional Neural Networks (CNN), Long Short-Term Memory Networks (LSTM), and Bidirectional LSTM (BLSTM), for the automated classification of electronic documents. The study emphasizes leveraging these sophisticated neural network architectures to categorize electronic documents into predefined classifications, such as news articles, academic papers, and social media content. To achieve this, the methodology involves a comprehensive preprocessing pipeline that includes tokenization and embedding techniques, transforming raw textual data into a format amenable for deep learning models. Subsequently, we design and implement CNN, LSTM, and BLSTM architectures utilizing Tensor Flow and Keras frameworks, training them on specifically labeled datasets representative of electronic document types. Model performance is thoroughly evaluated using metrics such as accuracy, precision, recall, and F1-score. Experimental results highlight that the implemented deep learning models exhibit commendable performance in accurately categorizing electronic documents across various domains. Notably, CNNs excel in capturing local patterns and features, while LSTMs and BLSTMs effectively analyze the sequential structure of document content to capture long-range dependencies. Furthermore, the study systematically investigates the impact of varying design configurations and hyper parameter settings on the classification accuracy, identifying optimal conditions for model performance. The outcomes of this research significantly advance document classification methodologies and have pertinent implications for information retrieval, content organization, and automated decision-making processes. The proposed framework not only enhances the capability to process extensive volumes of electronic documents but also bolsters knowledge discovery and informed decision-making across diverse sectors.

Keywords: Deep Learning, Document Classification, LSTM, Neural Networks, Text Processing