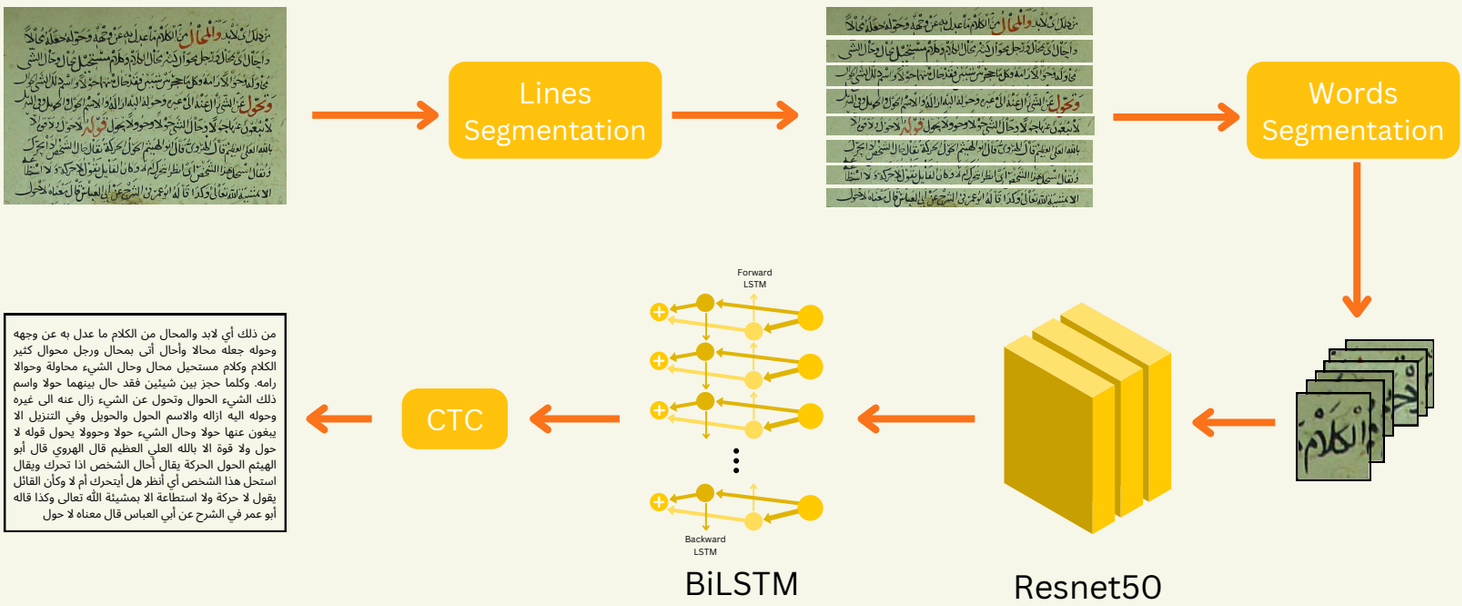
Automatic OCR System for Arabic Historical Manuscripts Using Deep Learning and image processing techniques

Mohammed A. Al-shabrawi Abdulshakoor R. Bantan Abdullah A. Al-zahrani Azzam A. Al-sharif

Abstract

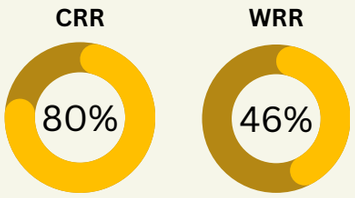
Historical Arabic manuscripts are important for preserving and understanding past civilizations. They provide insights into various aspects of ancient societies. However, these manuscripts are often in non-editable formats, making them challenging for scholars and researchers. A text recognition system is proposed to automate text extraction from these manuscripts to address this issue. The system uses the Seam carving algorithm to segment pages into lines and words, and the CNN-RNN-CTC model for word recognition. The system achieved a character error rate (CER) of 51% for line images and a CER of 20% and word error rate (WER) of 54% for word images, surpassing all existing systems.

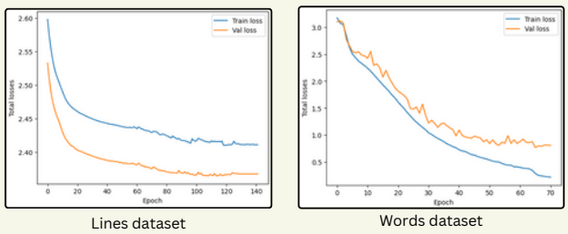
Methodology



Results

Our model outperforms other systems in terms of Word Recognition Rate (WRR) and Character Recognition Rate (CRR) in word datasets. This achievement can be attributed to the implementation of advanced techniques, as outlined in this paper, which differentiate our system from others.





Lines dataset Words dataset

Tools

Conclusion

The authors conclude that there is a lack of available datasets specifically for Arabic historical manuscripts, and the process of manually labeling the data is time-consuming. They suggest finding faster ways to label the data or exploring unsupervised techniques that don't require extensive labeling. In some cases, it is challenging to separate words using image processing or vision models because Arabic words often overlap. Even native speakers may struggle to visually segment the words accurately. Therefore, the authors propose the use of NLP models to improve word segmentation.