Initial Research and Literature Review

Title	Methodology	Strengths	Limitations
TrOCR: Transformer-based Optical Character Recognition with Pre-trained Models (Minghao Li, et al. 2021)	•Transformer Architecture •Uses pre-trained CN and NLP models •Splits image into sequence of patches that are used as inputs	•SOTA Results •Uses pre-trained CN and NLP models, which take advantage of large-scale unlabeled data for image understanding and language modeling, with no need for an external language model •Does not require CNN for backbone, so image-specific biases are avoided	•Requires huge amount of data •Not suitable for low-resource languages (e.g Urdu), due to the nature of transformer architecture as it requires large amounts of data
LayoutLM: Pre-training of Text and Layout for Document Image Understanding (ACM, 2019)(Yiheng Xu, et al. 2019)	•Uses both text and document layouts for training •Joint training in textual and layout information •BERT is used as the backbone, and adds two new input embeddings: Positional and image embedding •Positional embeddings to capture relationship among tokens within a document	•SOTA Results •Takes into account both textual and layout information, which is beneficial for a great number of real-world document image understanding tasks such as information extraction from scanned documents	•Only works for English (will not work for Urdu, or multilingual use cases) •Needs a separate model for text extraction and localization, alongside the LayoutLM model itself, which is quite computationally heavy

GAP

- No work has been done for mainstream work on multilingual free form document digitization

- No work has been done especially for Urdu

- Some work has been done but limited to Numbers only