SynthNID: Synthetic Data to Improve End-to-end Bangla Document Key Information Extraction

Syed Mostofa Monsur*, **Shariar Kabir***, **Sakib Chowdhury***
Celloscope Ltd.
{mostofa.monsur,shariar.kabir,sakib.chowdhury}@cellosco.pe

Abstract

End-to-end Document Key Information Extraction models require a lot of compute and labeled data to perform well on real datasets. This is particularly challenging for low-resource languages like Bangla where domain-specific multimodal document datasets are scarcely available. In this paper, we have introduced *SynthNID*, a system to generate domain-specific document image data for training OCR-less end-to-end Key Information Extraction systems. We show the generated data improves the performance of the extraction model on real datasets and the system is easily extendable to generate other types of scanned documents for a wide range of document understanding tasks. The code for generating synthetic data is available at https://github.com/dv66/synthnid

1 Introduction

Document Key Information Extraction (KIE) is a very crucial task to extract structured or semi-structured information from printed documents and images (Luo et al., 2023; Shi et al., 2023; Lee et al., 2022). Numerous user-facing applications nowadays require scanning and extracting information as key-value pairs from raw images of invoices, receipts, ID cards etc. Previously, these types of information extraction systems required a mandatory OCR engine and rule-based approaches in the pipeline. However OCR-based systems become error-prone easily because of the lack of contextual understanding (Kim et al., 2022). Also, hand-picked rule-based extraction systems cannot handle all possible transformations and variations that can be found in scanned documents.

Recently, a surge of pre-trained models has been witnessed in the area of document understanding. These models overcome the problems of OCR-powered Key Information Extraction systems by completely removing the need to apply character-level recognition and information aggregation. These models are essentially vision transformers which are pre-trained on huge datasets of scanned documents across multiple languages (Kim et al., 2022). These pre-trained models achieve state-of-the-art in various document understanding tasks on the particular languages they were pre-trained on. It is possible to fine-tune these models for downstream extraction tasks in other languages with sub-optimal performances. In the case of low-resource languages like Bangla, it is a severe issue because there are almost no datasets for the Document Key Information Extraction task in Bangla. Although a number of works are found in the literature regarding Bangla OCR and/or text detection systems (Safir et al., 2021; Hossain et al., 2022; Rabby et al., 2019; Alam et al., 2020), none of them perform end-to-end Key Information Retrieval which is essential to retrieve necessary fields from the scanned document.

Collecting annotated data for Document Key Information Extraction is also quite expensive and time-consuming. To reduce labelling effort and cost of labelling, synthetic data is often used alongside real data to train models for various Document Understanding tasks (Gupta et al., 2016). Unfortunately, general-purpose synthetic image generators do not focus on Key Information Extraction only but on general-purpose document understanding tasks. Most of the state-of-the-art generators support only English or rich-resource corpus thus low-resource languages like Bangla are completely ignored. Also, the lack of availability of high-quality Bangla corpus for Key Information Extraction tasks is another reason for the absence of end-to-end models in this area. The end-to-end Key Information Extraction model requires huge datasets with millions of samples (Kim et al., 2022) and hundreds of GPU hours which also contributes to this issue. One option is to fine-tune the pre-trained multi-

* Equal contribution

Proceedings of the First Workshop on Bangla Language Processing (BLP-2023), pages 117–123

lingual end-to-end models on the target language e.g. Bangla to perform Key Information Extraction on real documents like the National ID card of Bangladesh, License Plates etc. This approach is particularly useful when the whole document image is used as an information source and the target output is a structured data format like JSON. Because then an extra field extraction or linking stage is no longer required to add to the extraction pipeline.

In this work, we propose a system *SynthNID*, to generate domain-specific synthetic data which improves the performance of end-to-end document key information extraction tasks when fine-tuned alongside real data. Our primary focus for this work was to extract key values from the National ID Card of Bangladesh which contains a mixture of English and Bangla text in the document but using our approach a wide range of scanned documents can also be generated for Key Information Extraction tasks. We demonstrate the effectiveness of the generated data by fine-tuning end-to-end Key Information Extraction models. Our synthetic data increases the model's performance in extracting key-value pairs from real datasets.

2 Related Work

Document Key Information Extraction is a widely studied task in the literature. Most popular models incorporate the output of an OCR engine and learn to parse them from scanned documents (Hwang et al., 2021b,a). (Hwang et al., 2019) and (Majumder et al., 2020) have applied Document Key Information Extraction to various real-world applications. Most of these approaches introduce a learning framework where the text is detected separately using an off-the-shelf OCR engine and a sequence model takes the input from the previous stage considering the text content and locality of the information. Despite the convenience of end-to-end models for KIE tasks, only a few are available in the literature like OCR-free document understanding transformer (Kim et al., 2022). This model takes the whole document image as an input and applies a visual attention mechanism to learn the output sequence which is essentially a key-value structure like JSON. However, the end-to-end model variants are only pre-trained in Chinese, Japanese, Korean and English.

Although there are a number of works present in literature regarding Bangla OCR and a few in Document Understanding, almost none of them address end-to-end Key Information Extraction on Bangla scanned documents or multilingual scanned documents where Bangla is present. *bbOCR* is a scalable document OCR that employs a Bangla text recognition model using synthetic datasets (Zulkarnain et al., 2023). *BaDLAD* is a large multi-domain Bangla Document Layout Analysis dataset which contains more than 33k manually labeled documents from a wide range of sources including books, magazines, newspapers etc. (Shihab et al., 2023). (Ataullha et al., 2023) improves Document Layout Analysis performance leveraging Mask R-CNN architectures. They show competitive results in segmenting Bangla Documents.

Most of the existing works in Bangla Document Understanding have tackled problems like text extraction or layout analysis by segmenting the image components, whereas none of them considered extracting key information in a structured format which can be easily used by independent user-facing applications. In this effort, we have addressed this gap in the existing literature and aimed to solve the issue by new approaches to generate Bangla synthetic documents for domain-specific KIE tasks.

3 Datasets

3.1 Synthetic Dataset Generation

Our synthetic data generation system depends on named entities and random background images. We have collected a dataset of Bangladeshi Bangla first names, middle names and last names for males and females. We developed an empty layout of the 'overlay' which is proportionally similar to the national ID card of Bangladesh. For this work, we have skipped the image and signature part of the ID card because we are only interested in the text here. On the Bangladeshi national ID card's front side, the person's name, mother's name, father's name, date of birth, and identification number these fields are dynamic. Other texts don't change mostly. Our tool picks random names from a dataset of names, generates proper names, and fills the dynamic slots on the overlay. The name dataset was collected manually by labeling named entities from publicly available Bangla corpus. Identification number and date of birth fields are randomly generated according to their standard formats and inserted. Then we pick a random background image from an image store and put the data-filled overlay on the back-

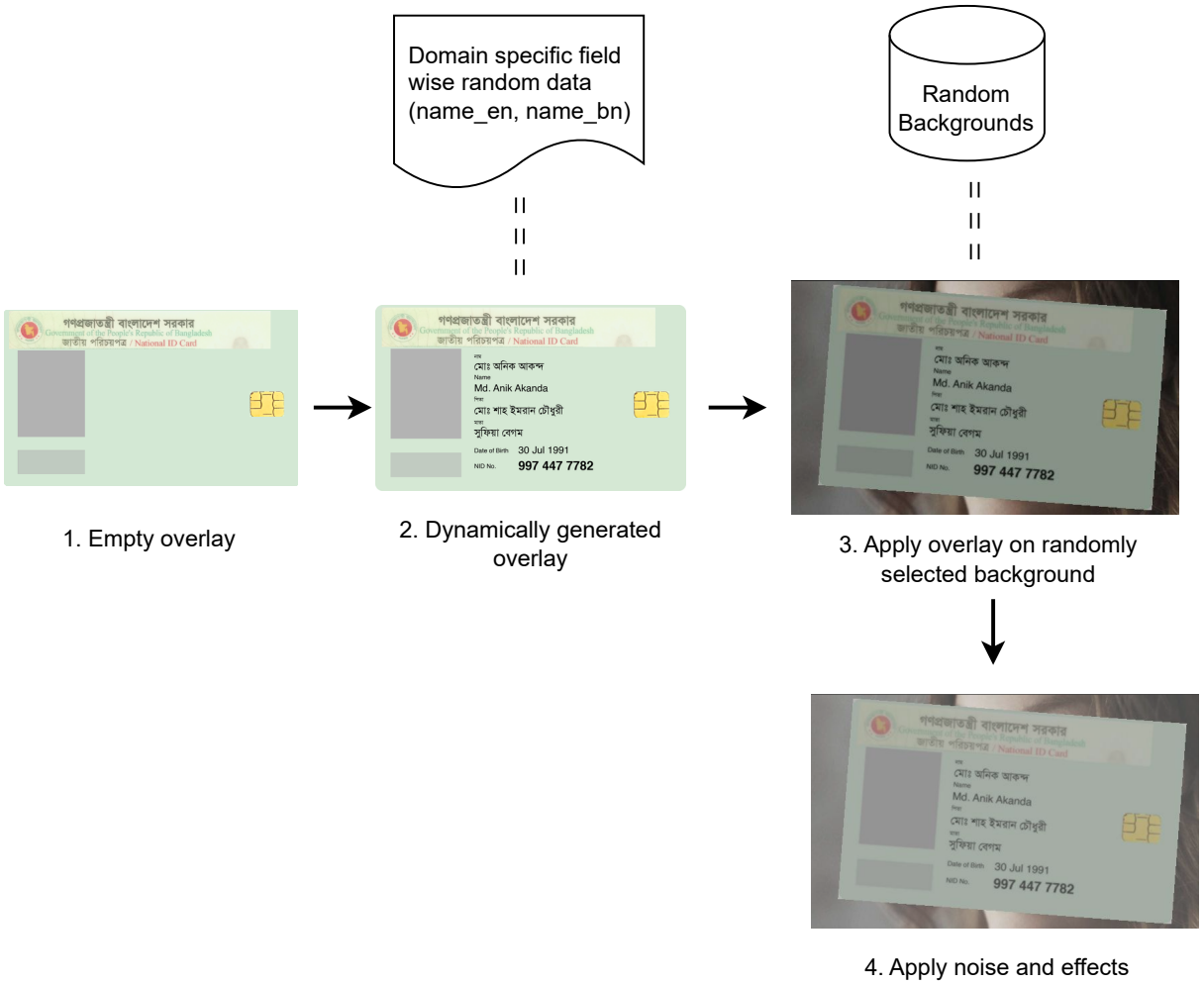

Figure 1: Steps for preparing synthetic NID data for KIE fine-tuning.

ground image. While doing this, we ensured that the overlay covered from 75% to 90% of the background after applying a rotational transformation. Finally, we apply random blur noise effects using (Jung et al., 2020).

With simple customization, this tool can be used to generate a wide range of Bangla-scanned synthetic documents for downstream document understanding tasks.

3.2 Real Dataset

In our experiments, we have used a real dataset of 11,390 images. From this, 10,890 are used for the training and validation and 500 for testing. The end users were provided with a mobile application for data collection. The mobile application allows end users to capture an image using a guiding rectangle box. Users had the option to review the captured image and, if necessary, retake it to ensure data quality. Although most of the real data was of good quality, there were some unavoidable noisy, faded, and rotated or tilted images which made the real data more challenging for the model than synthetic data.

4 Experiments

For our end-to-end model, we used the OCR-free document transformer (Kim et al., 2022) which is the current state-of-the-art in KIE. The model outperforms various other models like *LayoutLM* (Xu et al., 2019), *LayoutLMv2* (Xu et al., 2020), *LayoutXLM* (Xu et al., 2021), *SPADE* (Hwang et al., 2021b), *WYVERN* (Hwang et al., 2020) in document Key Information Extraction tasks. The model is essentially a vision transformer (Dosovitskiy et al., 2020) which is pre-trained on a huge dataset containing real and synthetic data 13M in total. The real dataset IIT-CDIP (Lewis et al., 2006) contains around 11M samples of complex scanned English

Training Dataset	Performance					
	Real Test Data Acc.			Synthetic Test Data Acc.		
	Bangla Fields	English Fields	Overall	Bangla Fields	English Fields	Overall
synth:real-50K:0K	25.96%	31.10%	25.02%	90.71%	94.55%	92.37%
synth:real-0K:10K	76.55%	82.92%	79.28%	80.08%	92.8%	85.91%
synth:real-2K:10K	78.76%	83.95%	81.14%	83.57%	96.54%	89.54%
synth:real-5K:10K	80.56%	83.79%	82.01%	85.58%	98.52%	91.55%
synth:real-10K:10K	**81.53%**	83.73%	82.5%	85.79%	98.79%	91.59%
synth:real-50K:10K	81.35%	**84.6%**	**82.74%**	**89.06%**	**99.39%**	**93.72%**

Table 1: Performance of the models trained on different splits of the dataset in terms of TED

documents. They created a synthetic dataset generation system *SynthDoG* which generates synthetic document samples in Chinese, Japanese, Korean, and English 0.5M per language.

4.1 Fine-tuning Process

We fine-tune the Donut-base model (Kim et al., 2022) for the Key Information Extraction (KIE) task in a mixed scheme where we use splits of both synthetic and real data in our fine-tuned training set. The input resolution is set to 345×575 pixels and the max length in the decoder is set to 100. For the English and Bangla multilingual tokenizer, we use the Banglabert-large model (Bhattacharjee et al., 2022). We finetune the model in early stopping setup for a maximum epoch of 30, using Pytorch-Lightning module (Falcon, 2019) and with one NVIDIA RTX 3070 GPU. We use the Adam (Kingma and Ba, 2014) optimizer, training and validation batch size is set to 512 and 8 respectively and the learning rate is set to 3×10^{-5}. We use a number of 1,000 training samples per epoch. For the evaluation of the models, we use the *tree edit distance* (TED) metric (Zhang and Shasha, 1989), by representing the extracted field values of the NID as a tree.

4.2 Performance on Split Datasets

We evaluated 6 different models trained on different splits of the dataset containing different mixes of the real with synthetic data. Our dataset contains a total of 10,890 (10K) real and 50,000 (50K) synthetic NID images. The real dataset contains more than one images from a user (but in a slightly different orientation). An NID contains 3 Bangla fields (name_bn, father_name and mother_name) and 3 English fields (name_en, dob and nid_no). For establishing a baseline the first two models were trained on only synthetic and real data respectively. For the rest of the 4 models, we used a mix of the 10K real data with different quantities of synthetic data for training. For the evaluation of the models, we used an unseen test set containing 500 synthetic and 500 real data across all the tests.

The results are shown in Table 1. The performance over the Bangla fields (name_bn, father_name and mother_name) and English fields (name_en, dob and nid_no) are shown separately along with the overall performance. For the models' performance over each field please refer to Appendix A. We found that, although the first model trained purely on the 50K synthetic data performs well over the synthetic data it performs poorly over real data. This suggests even with our different approaches to make the synthetic data represent real data the model was not able to learn how to work with real data. In the second model where we used all of the 10K real data with no synthetic data, we see a significant improvement in the performance over real data. However, the performance was poor over the Bangla fields where the second model only achieved an accuracy of 76.55% over the Bangla fields of real data.

We start to see performance improvement in the third case where the model was trained on all of the 10K real data along with 2K synthetic data. In fact, the model outperformed the second model where the train set contained only real data. This proves that a mix of synthetic with real data indeed improves the performance of the model. The performance was more prominent in the case of the Bangla fields where the accuracy improved from 76.55% to 78.76% over real data. The performance improves consistently over the Bangla fields as well as the English fields as more synthetic data is mixed in the train set. Most of the improvement

was found in the extraction of Bangla fields with the addition of synthetic data and we were able to achieve a best of 81.53% accuracy over the Bangla fields of the real data in the fifth model. In the sixth model where all 50K synthetic data is mixed with the 10K real data, we start to see a slight drop in the accuracy of the Bangla field extraction from real data. This suggests a case of diminishing returns in the performance over real data when there is more synthetic data than real data in the mix.

5 Ethical Considerations

While developing our system we prioritized end users' privacy protection. The app was developed and used to collect data inside the organization. Informed consent was obtained from the app users and stringent data anonymization measures were applied while using real data for testing the models' performances. While generating the synthetic data, every field is generated in a completely random strategy. Our ethical framework was aimed to develop high-quality Bangla KIE models while protecting user privacy, maintaining transparency, and ensuring responsible data handling thus strengthening our commitment to conduct ethical AI research.

6 Conclusion

In this paper, we have introduced a scheme to generate high-quality domain-specific synthetic data for the Key Information Extraction task on Bangla scanned documents. We have shown the synthetic data generated using our approach enhances the performance of end-to-end KIE models. In future, we will investigate the areas where effective labelling strategies can be employed to learn good models with a low amount of data using active learning techniques.

References

Samiul Alam, Tahsin Reasat, Asif Shahriyar Sushmit, Sadi Mohammad Siddiquee, Fuad Rahman, Mahady Hasan, and Ahmed Imtiaz Humayun. 2020. A large multi-target dataset of common bengali handwritten graphemes.

Md Ataullha, Mahedi Hassan Rabby, Mushfiqur Rahman, and Tahsina Bintay Azam. 2023. Bengali document layout analysis with detectron2.

Abhik Bhattacharjee, Tahmid Hasan, Wasi Ahmad, Kazi Samin Mubasshir, Md Saiful Islam, Anindya Iqbal, M. Sohel Rahman, and Rifat Shahriyar. 2022. BanglaBERT: Language model pretraining and benchmarks for low-resource language understanding evaluation in Bangla. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 1318–1327, Seattle, United States. Association for Computational Linguistics.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. 2020. An image is worth 16x16 words: Transformers for image recognition at scale.

William Falcon. 2019. Pytorch lightning.

Ankush Gupta, Andrea Vedaldi, and Andrew Zisserman. 2016. Synthetic data for text localisation in natural images.

Md. Ismail Hossain, Mohammed Rakib, Sabbir Mollah, Fuad Rahman, and Nabeel Mohammed. 2022. Lila-boti : Leveraging isolated letter accumulations by ordering teacher insights for bangla handwriting recognition.

Alyssa Hwang, William R. Frey, and Kathleen McKeown. 2020. Towards augmenting lexical resources for slang and African American English. In *Proceedings of the 7th Workshop on NLP for Similar Languages, Varieties and Dialects*, pages 160–172, Barcelona, Spain (Online). International Committee on Computational Linguistics (ICCL).

Wonseok Hwang, Seonghyeon Kim, Minjoon Seo, Jinyeong Yim, Seunghyun Park, Sungrae Park, Junyeop Lee, Bado Lee, and Hwalsuk Lee. 2019. Post-{ocr} parsing: building simple and robust parser via {bio} tagging. In *Workshop on Document Intelligence at NeurIPS 2019*.

Wonseok Hwang, Hyunji Lee, Jinyeong Yim, Geewook Kim, and Minjoon Seo. 2021a. Cost-effective end-to-end information extraction for semi-structured document images. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 3375–3383, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Wonseok Hwang, Jinyeong Yim, Seunghyun Park, Sohee Yang, and Minjoon Seo. 2021b. Spatial dependency parsing for semi-structured document information extraction. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 330–343, Online. Association for Computational Linguistics.

Alexander B. Jung, Kentaro Wada, Jon Crall, Satoshi Tanaka, Jake Graving, Christoph Reinders, Sarthak Yadav, Joy Banerjee, Gábor Vecsei, Adam Kraft, Zheng Rui, Jirka Borovec, Christian Vallentin, Semen Zhydenko, Kilian Pfeiffer, Ben Cook, Ismael Fernández, François-Michel De Rainville, Chi-Hung

Weng, Abner Ayala-Acevedo, Raphael Meudec, Matias Laporte, et al. 2020. imgaug. https://github.com/aleju/imgaug. Online; accessed 01-Feb-2020.

Geewook Kim, Teakgyu Hong, Moonbin Yim, JeongYeon Nam, Jinyoung Park, Jinyeong Yim, Wonseok Hwang, Sangdoo Yun, Dongyoon Han, and Seunghyun Park. 2022. Ocr-free document understanding transformer. In *European Conference on Computer Vision (ECCV)*.

Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Chen-Yu Lee, Chun-Liang Li, Timothy Dozat, Vincent Perot, Guolong Su, Nan Hua, Joshua Ainslie, Renshen Wang, Yasuhisa Fujii, and Tomas Pfister. 2022. FormNet: Structural encoding beyond sequential modeling in form document information extraction. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3735–3754, Dublin, Ireland. Association for Computational Linguistics.

D. Lewis, G. Agam, S. Argamon, O. Frieder, D. Grossman, and J. Heard. 2006. Building a test collection for complex document information processing. In *Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '06, page 665–666, New York, NY, USA. Association for Computing Machinery.

Chuwei Luo, Changxu Cheng, Qi Zheng, and Cong Yao. 2023. Geolayoutlm: Geometric pre-training for visual information extraction.

Bodhisattwa Prasad Majumder, Navneet Potti, Sandeep Tata, James Bradley Wendt, Qi Zhao, and Marc Najork. 2020. Representation learning for information extraction from form-like documents. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6495–6504, Online. Association for Computational Linguistics.

AKM Shahariar Azad Rabby, Sadeka Haque, Md. Sanzidul Islam, Sheikh Abujar, and Syed Akhter Hossain. 2019. Ekush: A multipurpose and multitype comprehensive database for online off-line bangla handwritten characters. In *Recent Trends in Image Processing and Pattern Recognition*, pages 149–158, Singapore. Springer Singapore.

Farisa Benta Safir, Abu Quwsar Ohi, M. F. Mridha, Muhammad Mostafa Monowar, and Md. Abdul Hamid. 2021. End-to-end optical character recognition for bengali handwritten words.

Dengliang Shi, Siliang Liu, Jintao Du, and Huijia Zhu. 2023. Layoutgcn: A lightweight architecture for visually rich document understanding. In *Document Analysis and Recognition - ICDAR 2023*, pages 149–165, Cham. Springer Nature Switzerland.

Md. Istiak Hossain Shihab, Md. Rakibul Hasan, Mahfuzur Rahman Emon, Syed Mobassir Hossen, Md. Nazmuddoha Ansary, Intesur Ahmed, Fazle Rabbi Rakib, Shahriar Elahi Dhruvo, Souhardya Saha Dip, Akib Hasan Pavel, Marsia Haque Meghla, Md. Rezwanul Haque, Sayma Sultana Chowdhury, Farig Sadeque, Tahsin Reasat, Ahmed Imtiaz Humayun, and Asif Shahriyar Sushmit. 2023. Badlad: A large multi-domain bengali document layout analysis dataset.

Yang Xu, Yiheng Xu, Tengchao Lv, Lei Cui, Furu Wei, Guoxin Wang, Yijuan Lu, Dinei Florencio, Cha Zhang, Wanxiang Che, Min Zhang, and Lidong Zhou. 2020. Layoutlmv2: Multi-modal pre-training for visually-rich document understanding.

Yiheng Xu, Minghao Li, Lei Cui, Shaohan Huang, Furu Wei, and Ming Zhou. 2019. Layoutlm: Pre-training of text and layout for document image understanding.

Yiheng Xu, Tengchao Lv, Lei Cui, Guoxin Wang, Yijuan Lu, Dinei Florencio, Cha Zhang, and Furu Wei. 2021. Layoutxlm: Multimodal pre-training for multilingual visually-rich document understanding.

K Zhang and D Shasha. 1989. Simple fast algorithms for the editing distance between trees and related problems. *SIAM Journal on Computing*, 18(6):1245–1262.

Imam Mohammad Zulkarnain, Shayekh Bin Islam, Md. Zami Al Zunaed Farabe, Md. Mehedi Hasan Shawon, Jawaril Munshad Abedin, Beig Rajibul Hasan, Marsia Haque, Istiak Shihab, Syed Mobassir, MD. Nazmuddoha Ansary, Asif Sushmit, and Farig Sadeque. 2023. bbocr: An open-source multi-domain ocr pipeline for bengali documents.

A Appendix

A.1 Models' Performance Variation Across Bangla Fields

Table 2 and 3 shows the performance of the models over the three Bangla fields: name_bn, father_name and mother_name across the real and synthetic test sets respectively. In the NID card of Bangladesh, all of the three fields consist of only Bangla letters. Although the models perform a little worse than English fields across the Bangla field which is due to the absence of Bangla data in the pre-trained base Donut model, we can see a steady increase in performance as more synthetic data is used in the fine-tuning process. The performance improvement was almost equal across the real (Table 2) and synthetic test data (Table 3).

A.2 Models' Performance Variation Across English Fields.

Table 4 and 5 shows the performance of the models over the 3 English fields: name_en, dob and nid_no

Training Dataset	Performance		
	name_bn	father_name	mother_name
synth:real-50K:0K	27.67%	30.03%	28.92%
synth:real-0K:10K	77.23%	72.9%	81.2%
synth:real-2K:10K	78.98%	75.69%	82.7%
synth:real-5K:10K	81.23%	77.2%	84.59%
synth:real-10K:10K	82.02%	79.07%	83.98%
synth:real-50K:10K	81.9%	79.98%	83.55%

Table 2: Performance of the models over Bangla fields in terms of TED across real test data

Training Dataset	Performance		
	name_bn	father_name	mother_name
synth:real-50K:0K	90.63%	90.57%	91.4%
synth:real-0K:10K	82.1%	78.08%	81.05%
synth:real-2K:10K	84.35%	82.54%	84.68%
synth:real-5K:10K	85.57%	84.94%	86.92%
synth:real-10K:10K	85.9%	85.92%	87.86%
synth:real-50K:10K	89.9%	89.31%	90.25%

Table 3: Performance of the models over Bangla fields in terms of TED across synthetic data

across the real and synthetic test sets respectively. In the NID card of Bangladesh, name_en consists of only English letters, dob consists of a mix of English letters and numbers (DD Month YYYY) and

Training Dataset	Performance		
	name_en	dob	nid_no
synth:real-50K:0K	24.33%	52.42%	37.72%
synth:real-0K:10K	78.84%	96.02%	77.87%
synth:real-2K:10K	79.44%	96.67%	79.97%
synth:real-5K:10K	82.11%	96.37%	76.09%
synth:real-10K:10K	81.71%	95.88%	76.75%
synth:real-50K:10K	81.49%	95.85%	79.85%

Table 4: Performance of the models over English fields in terms of TED across real test data

Training Dataset	Performance		
	name_en	dob	nid_no
synth:real-50K:0K	99.41%	99.82%	81.82%
synth:real-0K:10K	89.54%	97.23%	93.09%
synth:real-2K:10K	92.19%	99.93%	99.36%
synth:real-5K:10K	96.74%	99.98%	99.69%
synth:real-10K:10K	97.74%	99.5%	99.73%
synth:real-50K:10K	99.17%	99.45%	99.91%

Table 5: Performance of the models over English fields in terms of TED across synthetic test data

nid_no consists of only English numbers. The base Donut model being pre-trained on English data performs better across the English fields, except for the nid_no field of real test data where the model seems to struggle more than any other fields. Like before we see a steady increase in performance as more synthetic data is used in the fine-tuning process with equal performance improvement across the real (Table 4) and synthetic test data (Table 5).