A Comprehensive Analysis of LayoutLM and Donut for Document Classification

Merxhan Bajrami[1,2], Eftim Zdravevski[1], Petre Lameski[1] and Biljana Stojkoska[1]
[1] *Faculty of Computer Science and Engineering, Ss. Cyril and Methodius University, Skopje, N. Macedonia*
[2] *Laigo GmbH, Eckenerstraße 65, Friedrichshafen, 88046, Germany*

merdjan.bajrami@laigo.ai, eftim.zdravevski@finki.ukim.mk,
petre.lameski@finki.ukim.mk, biljana.stojkoska@finki.ukim.mk

***Abstract*—Document classification is important in everyday life as it allows for efficient management and organization of vast amounts of digital documents, saving time and resources. This task is essential for businesses, organizations, and individuals who handle large volumes of data and need to quickly retrieve and analyze specific information. AI-based document classification can help organizations better manage and organize their digital assets, improve information retrieval, and make better business decisions based on the insights derived from the classified documents. This paper compares the performance of two transformer-based models, LayoutLM and Donut, for image classification tasks on two different datasets. LayoutLM was trained using pre-trained weights from Microsoft, while Donut used pre-trained weights from Huggingface. Both models were fine-tuned for 100 epochs with early stopping technique, using the Adam optimizer and Cross Entropy Loss. Our results show that LayoutLM performs better than Donut on the first dataset, achieving an overall accuracy of 0.88, while Donut achieved an accuracy of 0.74. Our study demonstrates the importance of carefully selecting and evaluating different models for document classification tasks, based on the specific char- acteristics of the dataset and the task requirements. Additionally, we provide insights into the strengths and weaknesses of both LayoutLM and Donut models for document classification on different datasets.**

***Index Terms*—document classification, layout analysis, OCR, intelligent document processing**

I. Introduction

Document classification is a crucial task in information retrieval and management, especially in Western countries where vast amounts of digital documents are generated every day. Document classification involves automatically categorizing documents into predefined classes or categories based on their content. It can be used in a variety of applications, such as email spam filtering, news categorization, legal document classification, and financial data analysis.

One of the main reasons for the importance of document classification in Western countries is the sheer volume of data that is generated. According to a study by IDC [2], unstructured data, such as documents, images, and videos, constitutes up to 80% of all data generated by businesses. It is therefore essential for businesses to be able to classify this data effectively to derive value from it. Another study by AIIM found that document classification can improve productivity by up to 50% and reduce errors by up to 35% [3].

Automated document classification has numerous benefits, such as improving the efficiency of document processing, helping organizations to better manage and organize their digital assets, and making better business decisions based on insights derived from the classified documents. Additionally, automated document classification can help to reduce errors and inconsistencies that may arise due to the subjective interpretation of human classifiers.

In this paper, we aim to investigate two different state-of-the-art algorithms for document classification. LayoutLM (Xu et al., 2019) is a pre-trained model for document image analysis that combines object detection and recognition with language modeling. LayoutLM is specifically designed for text classification on documents with complex layouts, which are present in many real-world document classification tasks.

Donut (Kim et al., 2021) is an OCR-free visual document understanding model that aims to solve the issues of using OCR engines in visual document understanding. Donut is a simple yet effective end-to-end sequence model that achieves state-of-the-art performance on various document understanding tasks.

Overall, the related work shows that deep learning models, transfer learning, and pre-trained language models have significantly improved document classification tasks. LayoutLM and Donut are recent models that have shown promising results on document classification tasks, with LayoutLM being more suitable for documents with complex layouts and Donut being an OCR-free model that can generalize well to new data.

To the best of our knowledge, this is the first study that compares the performance of LayoutLM and Donut models in the task of document classification. Previous studies have evaluated each model separately, but there has been no direct comparison between them. Therefore, this research provides new insights into the strengths and weaknesses of these state-of-the-art models and their suitability for different document classification tasks.

The rest of this paper is organized as follows. Section II explains the related existing work about document classification task. In Section III we provide information about the methodology we have implemented, which models we have used and a description of our datasets that we have used. In Section IV we show how we implemented this work, fine tuning process and information regarding the hyperparameters

optimization. In Section V we give insights at the result that we got and we interpret them. In Section VI we provide details about conclusion section, where we give reasons why the model performed in that way and the implication of this paper for this task. Finally, this paper is concluded in Section VII, where we discuss about future work.

II. Related work

Document classification is a well-studied task in the field of natural language processing and has been approached using various techniques and algorithms. Pisoni, Molnár, and Tarcsi (2021) examine the role of big data in financial services. They analyze data science tools, showcase enterprise architecture, and emphasize the significance of data lakes, warehouses, knowledge management, and customer involvement. The study explores emerging technological approaches for developing additional services in finance, contributing valuable insights to the field [1]. In recent years, deep learning models have shown significant improvements in document classification tasks. For instance, Convolutional Neural Networks (CNNs) have been used for document classification tasks, with approaches such as DocCNN (Zhang et al., 2015) achieving state-of-the-art results on benchmark datasets.

Another approach for document classification is Recurrent Neural Networks (RNNs), with models such as Recursive Neural Tensor Network (RNTN) (Socher et al., 2013) and Recurrent Convolutional Neural Networks (RCNN) (Lai et al., 2015) achieving competitive results.

Moreover, Transfer Learning has been proven to be an effective approach for document classification tasks. Pre-trained language models such as BERT (Devlin et al., 2019) and GPT (Radford et al., 2018) have been fine-tuned on document classification tasks, achieving state-of-the-art results on various benchmark datasets.

III. Methodology

In our study, we selected two state-of-the-art models, LayoutLM [3] and Donut [4], to investigate their efficacy in comprehending and analyzing document layout, structure, content, and visual elements.

LayoutLM, a pre-trained language model, employs transformer architecture to encode each token's location and appearance within a document, capturing both textual and geographical information. This model has exhibited superior performance over previous state-of-the-art models in various document understanding tasks, such as information extraction, question answering, and named entity recognition.

Donut, on the other hand, is an OCR-free document understanding transformer designed to address challenges in document image understanding, including text reading and comprehensive document comprehension. The model presents a streamlined architecture and a pre-training objective that consistently delivers top-tier results in terms of accuracy and speed across a range of visual document understanding tasks.

In Figure 1, we present an illustration showcasing four distinct document types commonly encountered in business operations, which are representative of the diverse layouts and contents that our models, LayoutLM and Donut, must handle. These document images, originally in German, include an invoice, a receipt, a handwritten note, and a delivery note. Due to data privacy concerns and the confidential nature of the information contained within these documents, they have been blurred for the purpose of this publication.

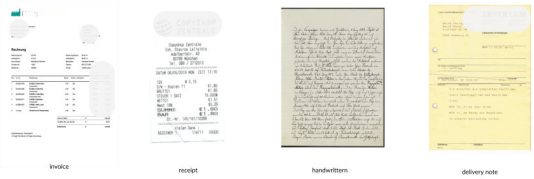


Fig1. Document samples from the datasets that we used in this study

Since the documents contain confidential data we have blurred them to ensure a security level on the data in documents. Each document type exhibits unique characteristics in terms of layout, structure, and visual elements. The invoice and receipt are primarily comprised of structured text, organized into tables and lists, with headings and subheadings to denote different sections. The delivery note typically contains a combination of structured and unstructured text, with item descriptions, quantities, addresses, and other essential information. The handwritten note, on the other hand, presents a more unstructured layout, with irregular handwriting and potential variations in text size, style, and orientation.

These diverse document types exemplify the challenges that document understanding models like LayoutLM and Donut must overcome. The models need to adapt and generalize effectively to handle variations in layout, structure, and content, as well as account for potential inconsistencies in document quality, such as poor scanning or image blurring. In our study, we sought to assess the performance of LayoutLM and Donut in dealing with such diverse document types, providing valuable insights into their respective strengths and weaknesses in handling real-world document understanding tasks.

For the purposes of our study, we employed two distinct datasets to train and evaluate the models. Dataset1 comprises 10,000 samples of internal company data, spanning 10 document classes. Dataset2, a larger and more diverse dataset, consists of 50,000 samples that include both internal data and samples from the RVL-CDIP dataset. This diverse collection of data allowed us to thoroughly assess the models' adaptability and generalization capabilities.

We conducted four experiments in total, with each model being trained and evaluated on both datasets. To measure the performance of LayoutLM and Donut, we utilized accuracy, precision, recall, and F1 scores as evaluation metrics. This comprehensive approach enabled us to compare and contrast the effectiveness of the models in handling various document understanding tasks, providing insights into their respective strengths and weaknesses.

IV. Implementation

In our study, we aimed to compare the performance of two state-of-the-art models, LayoutLM and Donut, for analyzing layout-based and text-based information. To ensure a thorough evaluation, we selected two diverse datasets, Dataset1 and Dataset2, which enabled us to examine the models' adaptability and generalization capabilities.

To fine-tune the models, we employed transfer learning, leveraging pre-trained LayoutLM and Donut models available from Huggingface. These pre-trained models were trained on extensive image-text datasets, providing a solid foundation for our task. By initializing our models with these pre-trained weights, we benefitted from the knowledge and patterns captured during their initial training, which led to faster convergence and improved performance on our specific datasets.

During the fine-tuning process, we employed a cross-entropy loss function and the Adam optimizer, with a maximum of 100 epochs. To find the best model configuration and prevent overfitting, we utilized the Optuna framework for hyperparameter optimization and implemented early stopping. Early stopping is a regularization technique that monitors the model's performance on the validation set during training. When the performance on the validation set ceases to improve or begins to degrade, training is halted. This approach helps in mitigating overfitting by preventing the model from excessively learning the noise present in the training data.

The early stopping technique, combined with the use of pre-trained models and transfer learning, allowed us to effectively fine-tune LayoutLM and Donut for our specific tasks while minimizing the risk of overfitting. This comprehensive evaluation approach facilitated a thorough comparison of the models' performance across different datasets and provided insights into their strengths and limitations in handling layout-based and text-based information.

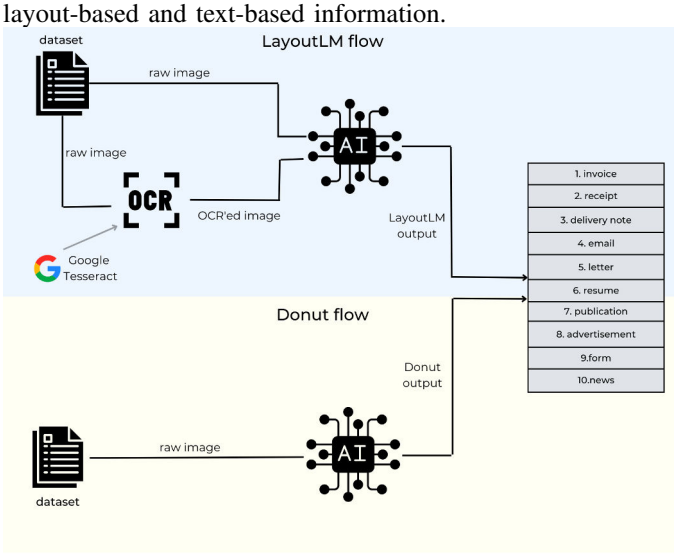


Fig1. Workflow of the LayoutLM and Donut, which have the same output

At Fig1. it is represented the workflow of how the two models work. On the upper side of the figure is shown LayoutLM workflow and downside of the figure is shown the Donut model workflow. In principle their process of the work is the same. The data which are labelled will be fed into the model and then we have the output of the model which can be one of the 10 classes: invoice, recepit, delivery note, email, letter, resume, publication, advertisement, form and news. The main difference in the workflow between two models is that LayoutLM as input has additionally the OCR'ed files, since it makes analysis on text base also. For OCR purposes we have used Tesseract[10] which is an open-source service from Google. Donut is a well known model only for this reason that it doesn't use OCR and as OCR-free model research shows that it perfroms good on uknown data related to Intelligent Document Processing tasks.

To evaluate the trained models, we used the testing set and calculated several metrics, including accuracy, precision, recall, and F1 score. These metrics allowed us to assess the performance of the models in terms of their ability to correctly classify the images in the datasets.

In addition to the above, we ensured reproducibility by setting the random seed for all experiments and made use of the same hardware and software environment throughout the experiments.

V. Results

The experiments conducted to evaluate the performance of LayoutLM and Donut models on document classification tasks yielded promising results, as illustrated in Figure 2.

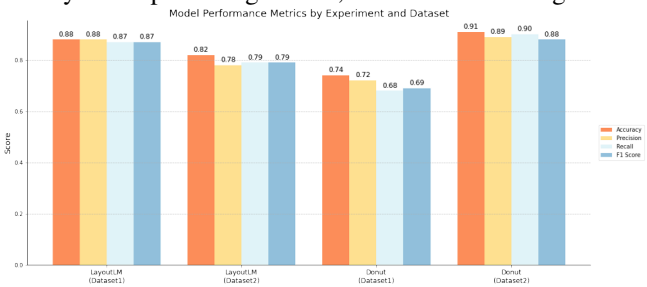


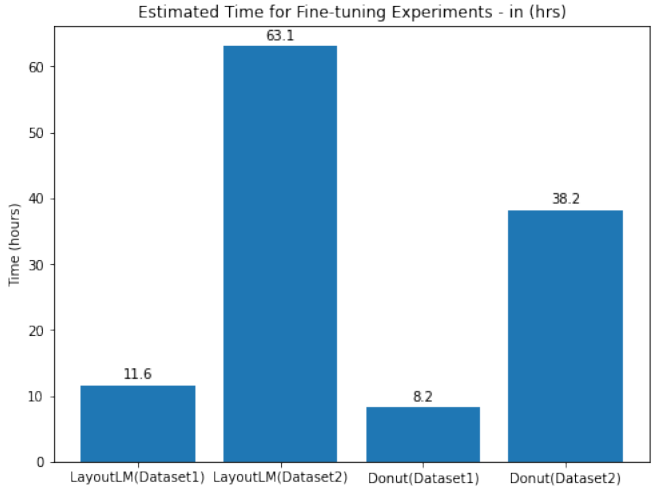
Fig.2 Models performance on each conducted experiment

At Fig2. we can see the models performance for each experiment. In Experiment 1, LayoutLM was evaluated on Dataset1. The results showed an overall accuracy of 0.88, precision of 0.88, recall of 0.87, and F1 score of 0.87. In Experiment 2, LayoutLM was evaluated on Dataset2 and the model achieved an overall accuracy of 0.82, precision of 0.78, recall of 0.79, and F1 score of 0.79.

In Experiment 3, Donut model was evaluated on Dataset1, which consisted of private data. The model achieved an overall accuracy of 0.74, precision of 0.72, recall of 0.68, and F1 score of 0.69. In Experiment 4, Donut model was evaluated on Dataset2 where the model achieved an overall accuracy of 0.91, precision of 0.89, recall of 0.90, and F1 score of 0.88.

The running time of the models is an important aspect to consider when evaluating their performance. We also

monitored the time when fine-tuning the models, on a GeForce RTX3090 NVIDIA graphics and the results were as following:



The results show that LayoutLM takes significantly longer to train than Donut, with both models taking longer to train on Dataset2 than on Dataset1. This is likely due to the larger size of Dataset2 and the need to process more data during training. Additionally, the use of a powerful GPU can significantly reduce the training time. Overall, these results suggest that the choice of model and dataset can have a significant impact on the training time, and that the use of a powerful GPU can help to reduce this time.

VI. Conclusion

The experiments on Dataset1 we conclude that LayoutLM performs better than Donut on the first dataset because it achieved an overall accuracy of 0.88, while Donut achieved an accuracy of 0.74. This could be because LayoutLM is specifically designed for text classification on documents with complex layouts, which are present in the first dataset. The experiments on Dataset 2 show that Donut performs better than LayoutLM, which is larger and more diverse than the first dataset. Donut achieved an overall accuracy of 0.91, while LayoutLM achieved an accuracy of 0.82. This could be because LayoutLM is specifically designed for text classification on documents with complex layouts, which are present in the first dataset. And on the other side Donut is a deep learning model that can learn to recognize patterns and features in the data that are not specific to any particular type of document, and can generalize better to new data.

Overall, the results show that both models are performing reasonably well on the given datasets, with accuracy, precision, recall, and F1 scores within a reasonable range for document classification tasks. We believe that the performance of the models can be increased if we get better quality data. Since the data that we used were old and not well scanned, which led to bad quality data.

VII. Future work

In future work, we plan to broaden our research on document classification by examining the performance of other state-of-the-art models using the datasets employed in our current study. This will provide a more comprehensive comparison and deeper understanding of the capabilities of various models in handling document understanding tasks. .

We also intend to expand our dataset by incorporating more complex document types, thereby increasing the diversity and complexity of the dataset. This will allow us to evaluate the models' performance on a larger scale and assess their adaptability to a wider range of real-world document understanding tasks.

We aim to investigate the interpretability of these models and have more insights about their work. This enhanced understanding will also enable us to identify potential areas for optimization and fine-tuning, ultimately leading to more accurate and efficient document classification systems.

Overall, these future directions will contribute to a deeper understanding of document classification tasks, allowing us to develop more effective solutions and practical applications for a wide range of industries and use cases.

Acknowledgments We would like to acknowledge the support of Laigo company for providing the datasets and computing resources used in this study. We also would like to thank the developers of LayoutLM, Donut and Tesseract for making their models publicly available.

References

[1] Data Science for Finance: Best-Suited Methods and Enterprise Architectures, 10.3390/asi4030069

[2] International Data Corporation, Data Age 2025: The Evolution of Data to Life-Critical," 2018. [Online]. Available: https://www.seagate.com/files/www-content/our-story/trends/files/Seagate-WP-DataAge2025-March-2017.pdf.

[3] Association for Information and Image Management, The Business Case for Paper-Free," 2012. [Online]. Available: https://info.aiim.org/hubfs/2012/AIIM%20White%20Paper%20-%20The%20Business%20Case%20for%20Paper-Free%20-%202012.pdf.

[4] Y. Xu, M. Li, L. Tian, Y. Shi, S. Yang, and X. Wang, LayoutLM: Pre-training of Text and Layout for Document Image Understanding," in Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), 2019, pp. 3070–3080.

[5] G. Kim, T. Hong, M. Yim, J. Nam, J. Park, J. Yim, W. Hwang, S. Yun, D. Han, and S. Park, Donut: OCR-free Document Understanding Transformer," arXiv preprint arXiv:2111.15664, 2021.

[6] Zhang, W., Li, Y., and Tan, C. (2015). Clustering with multi-viewpoint based similarity measure. Journal of Computer Science and Technology, 30(3), 535-547. Available: https://link.springer.com/article/10.1007/s11390-015-1536-7. Accessed April 7, 2023

[7] Socher, R., Perelygin, A., Wu, J. Y., Chuang, J., Manning, C. D., Ng, A. Y., Potts, C. (2013). Recursive deep models for semantic compositionality over a sentiment treebank. In Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP) (pp. 1631-1642). Available: https://aclanthology.org/D13-1170.pdf. Accessed April 7, 2023.

[8] Devlin, J., Chang, M. W., Lee, K., Toutanova, K. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics:

Human Language Technologies (NAACL-HLT) (pp. 4171-4186). Available: https://www.aclweb.org/anthology/N19-1423.pdf. Accessed April 7, 2023.

[9] Radford, A., Narasimhan, K., Salimans, T., Sutskever, I. (2018). Improving Language Understanding by Generative Pre-Training. OpenAI Technical Report.

[10] Google. (2021). Tesseract OCR. [Software]. Available: https://github.com/tesseract-ocr/tesseract. Accessed April 7, 2023.