

Developing ML driven NLP pipeline in AWS for Scalable Solutions
MA AI 136/2023

Dhruvkumar Ashokbhai Shihora

A Language Independent Natural Language Processing Model for an Intelligent Document Processing

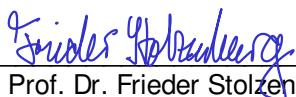
More than 80% of documents exist today are in unstructured format, most of them are not in machine readable format. Structured document understanding has got significant attention and made considerable progress in recent time. Therefore, the related work is focused on finding different models, methodologies, language/s and the datasets those models been trained on. Evaluating by comparing them to select the best possible model and methodology for the classification of the document content (Headers, Questions, Answers, Other) and fine-tune the model for better performance in one desired language for production use, evaluate the fine-tuned model to choose the best model for a deployment using cloud infrastructure that is scalable.

This thesis contains the following tasks:

- Reviewing existing NLP models and methods that can perform multi-lingual operations and comparing them to choose the model and methods for production use.
- Find the dataset and pre-trained models.
- Fine-tune the model for one specific language for better performance in one specific language (German and English) for production use.
- Developing the cloud-infrastructure using AWS (amazon web services) for deployment of the model for SaaS (software as a service).

The goal of this thesis is:

Developing SaaS using fine-tuned model and cloud-infrastructure to make multi-lingual documents machine-readable and classifying the content of multi-lingual documents.



Prof. Dr. Frieder Stolzenburg
1st Supervisor



Ms Warda Khan
2nd Supervisor