RESEARCH NEWS

Using deep learning to analyze texts

Analyzing documents faster using artificial intelligence from Fraunhofer

The flood of documents created every day in business and in society as a whole poses an enormous challenge. Information from countless different sources must be sorted, processed and evaluated. And this issue isn’t limited to companies: public authorities, research institutions and hospitals are affected, too. The Fraunhofer Institute for Intelligent Analysis and Information Systems IAIS has developed solutions that classify all kinds of documents and analyze their text content. The key to these solutions is AI-based language models trained with deep learning techniques.

E-mails, orders, delivery notes, quotes, contracts, reports – new data and documents are created every day in the course of doing business. Only when this flood of information can be structured in a meaningful way can companies make the right decisions and take action quickly. This applies equally to public institutions such as public authorities, libraries, research institutions and hospitals. To this end, Fraunhofer IAIS has developed a multi-part end-to-end solution: The “DocuLib” and the “NLU.Suite” are AI-based software solutions that make it possible to digitize, classify and analyze the content of any kind of documents almost automatically.

DocuLib – OCR software featuring deep learning technology

If paper documents are still available, they are scanned and captured by DocuLib, an OCR (optical character recognition) software. The text recognition works with deep learning models that were developed by the experts at Fraunhofer IAIS and that regularly rank at the top in international benchmarks. The software thus also recognizes poorly legible letters on paper that has become yellowed or torn. A patterned background, for example on transportation tickets, is no problem either.

The Fraunhofer IAIS solutions for document analysis also enable quick analysis of these digitized documents or those that are already available in digital format. For instance, they classify documents as invoices, travel receipts, flight tickets or e-mails. They also extract basic information such as names, dates and numbers. Also longer documents, such as letters or reports, can be analyzed and therefore linked to one another. This saves a lot of time anywhere numerous documents are created. For instance, the software can sort incoming mail and automatically forward it to the respective recipient.

Dr. Nicolas Flores-Herr, head of the Document Analytics business unit at Fraunhofer IAIS, says: “Our goal is to use artificial intelligence to largely automate information processing. This will speed up all document-based work processes for our customers.” The Fraunhofer IAIS software solutions are an ideal addition for companies that have installed document management or enterprise content management software. They are market-ready, in mainstream use at companies and being continuously refined by the Fraunhofer researchers.

Language models for analyzing text content

Natural Language Understanding (NLU) takes this technology one step further. NLU solutions are capable of analyzing the content of complex, unstructured documents. To enable semantic text analysis, the NLU team developed language models that are trained using deep learning techniques. The language models are first fed thousands of texts from a variety of fields. Newspaper articles, social media postings and e-mails are also among these texts. The AI module uses them to create a statistical model. Next, humans intervene and specify rules for targeted analysis. “Combining statistics and rules means that the software requires less data, and at the same time, it works faster and is more precise,” says Sven Giesselbach, who heads up the Natural Language Understanding team at Fraunhofer IAIS.

The NLU.Suite analyzes documents, extracts key data and, if necessary, even compiles a structured summary. These results, as well as the contents of the documents themselves, can be used to compare documents or find texts with similar information. If text A includes, for example, the terms “theft” and “chain”, and text B the terms “jewelry” and “stolen”, the language model recognizes the thematic relatedness. The software also understands that the sentence “The installment is due in advance at the start of the month.” in document A has a similar meaning to “The installment must be paid monthly in advance.” in document B.

“Our AI-based language models are clearly superior to conventional indexing. They not only find texts with predefined keywords, they also search intelligently for terms that occur in similar contexts or that are used as synonyms. Furthermore, the software also responds to morphological similarities,” explains Giesselbach. Court judgments are one example of a specific use case. For this, the IAIS experts collaborate with the faculty of law at the University of Cologne. The NLU.Suite, for example, independently searches for judgments that stand out based on such features as similar penalties or similarities in how crimes unfolded, making textual congruences between different documents visible.

In hospitals, the NLU solution can analyze medical diagnoses or physicians’ discharge reports. For new technical terms such as “Covid-19”, the AI software would recognize that the word “lung” occurs in the same context with above-average frequency and could thus find documents that deal with respiratory illnesses. The software also ensures that all data protection requirements are met. All personal data is anonymized, the servers are located in Germany, and the provisions of the General Data Protection Regulation (GDPR) are complied with.

Foreign-language documents

The NLU language models can also cope with foreign-language texts and can analyze both English and German documents in a single operation. Giesselbach and his team are continuously refining the deep learning language models. The system recognizes, for instance, positive or negative assessments in texts, and in some domains, such as the automotive industry, it is also able to recognize authors' emotions.

Users do not notice the NLU.Suite's complex structure. The application runs on normal desktop computers. Powerful hardware is only needed to create and train the language model. NLU applications that are already ready for use are those for analyzing court judgments, travel receipts and lease agreements. In healthcare, applications such as the analysis of specialist medical literature and clinical documents are ready for use.

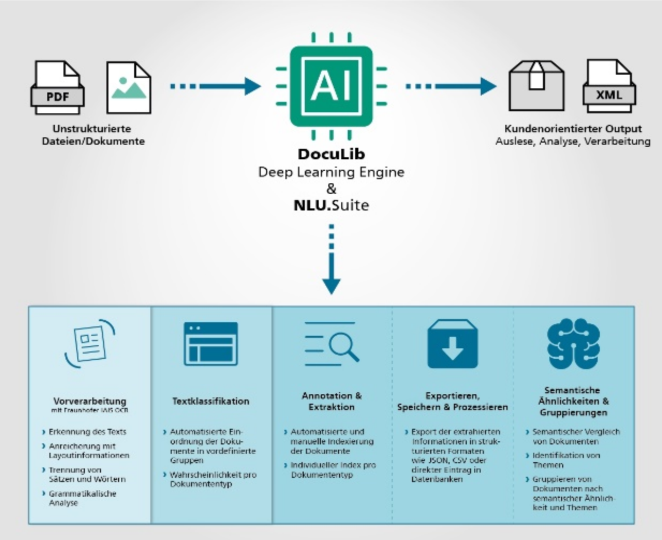
PDF
Unstrukturierte Dateien/Dokumente
AI
DocuLib
Deep Learning Engine
&
NLU.Suite
XML
Kundenorientierter Output
Auslese, Analyse, Verarbeitung
Vorverarbeitung
mit Fraunhofer IAIS OCR
› Erkennung des Texts
› Anreicherung mit Layoutinformationen
› Trennung von Sätzen und Wörtern
› Grammatikalische Analyse
Textklassifikation
› Automatisierte Einordnung der Dokumente in vordefinierte Gruppen
› Wahrscheinlichkeit pro Dokumententyp
Annotation & Extraktion
› Automatisierte und manuelle Indexierung der Dokumente
› Individueller Index pro Dokumententyp
Exportieren, Speichern & Prozessieren
› Export der extrahierten Informationen in strukturierten Formaten wie JSON, CSV oder direkter Eintrag in Datenbanken
Semantische Ähnlichkeiten & Gruppierungen
› Semantischer Vergleich von Dokumenten
› Identifikation von Themen
› Gruppieren von Dokumenten nach semantischer Ähnlichkeit und Themen

Fig. 1 **The Fraunhofer IAIS DocuLib solution and the NLU.Suite facilitate end-to-end text and document analysis – from OCR to understanding the text with the aid of artificial intelligence.**

© Fraunhofer IAIS

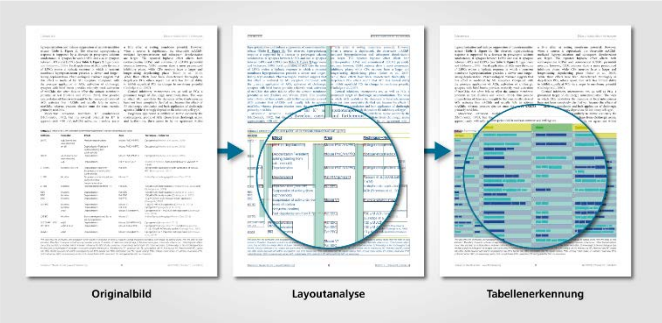

Fig. 2 An example of an AI-based layout analysis and table recognition of digitized documents.

© Fraunhofer IAIS