Title: Turning Dark Data into Informed Content. AI-driven Scalable Document Classification for Today's Enterprise Businesses

Name: Miguel Muñoz
Affiliation: Adlib Software

INTRODUCTION:
The use of Machine Learning and Natural Language Processing (NLP) techniques for document processing, classification, and information extraction is a very common scenario in the scientific literature. It is true that many old and new academic algorithms and methods tackle these problems and provide significant improvements—typically for very specific problems; however, to satisfy real-world business use cases, many other techniques and methods must be developed. Additionally, all of these methods and techniques have to be augmented and assembled into systems capable of addressing real-world business use cases, especially ones that demand broader applicability.

Numerous organizations across the globe need to process and analyze large volumes of very different unstructured content from documents in a variety of different formats. This content needs to be processed, standardized, and classified in an accurate way to maximize the benefit of the information it might contain. In addition, all the needed operations on this data have to be performed at a fast rate to meet the business requirements.

Join Adlib for a discussion on how the intersection of Machine Learning, Predictive Analytics, and Natural Language Processing techniques can be used for the purposes of document enrichment and document processing-classification automation. Classification and extraction of complex unstructured data is a common scenario in the scientific community; however, documented methods are often transactional in nature and rarely address real-world business applications, especially ones that demand broader applicability. Learn how you can turn dark data into informed content through techniques that leverage content standardization and AI-driven classification and information extraction.

AIM:
To design and implement a scalable and fault-tolerant production system that leverages existing and novel Machine Learning and Natural Language Processing techniques for document processing and classification for real-world business applications.

MATERIALS AND METHODS

A document processing system has been designed to implement and aggregate a number of Machine Learning and Natural Language Processing proprietary, open source, and licensed algorithms and subsystems to perform the needed document operations.
These operations are:

- Optical Character Recognition (OCR) on images and render to Portable Document Format (PDF)
- Image Processing
- Document fingerprint creation
- Document clustering
- Documents classification
- Information Extraction.

The above system was used to classify and extract data from multiple large document sets created across multiple lines of business. The resulting document classes and extracted data were validated using a semi-automated Quality Assurance process.

RESULTS:

The result is a system that ensures both reliability and scalability and incorporates both machine learning and software engineering metrics. This is achieved by developing a modular system with a collection of subsystems that can be deployed and coordinated from a central source. This scalable system is supported by a number of other engines—that can either be physical machines or cloud based virtual machines—which process large document volumes while allowing the flexibility to adapt to changing business requirements.

Regarding performance metrics, our system is capable of processing ~3000 documents/h in each deployed machine, which can lead to processing more than 10M documents per day in some of our largest deployed clusters.

Regarding our machine learning systems we were able to achieve the following classification and Information Extraction metrics in some of our proprietary solutions on real-world data:

	Document Classification	**Information Extraction – NER**
Precision	0.972	0.939
Recall	0.984	0.857
F1	0.978	0.892

CONCLUSIONS:

We have developed an enterprise document enrichment and process automation system that empowers customers to efficiently process, classify, and extract business critical data by leveraging some of the latest advancements in Machine Learning and Natural Language Processing. This gives organizations the ability to turn large volumes of unstructured data into standardized content and structured information. By revealing new insights in critical documents they can accelerate business decisions and improve customer experiences, as well as improve their adherence to compliance requirements and mitigate risks.

KEYWORDS:
Document Classification, Natural Language Processing, Machine Learning, Scalable Production Machine Learning Systems, OCR, PDF

BIOGRAPHY:
Miguel Muñoz is an MSc in Artificial Intelligence by the Polytechnic University of Madrid (Spain), where he also completed his BSc and MSc in Computer Science and Software Engineering. He has 7 years of experience in developing Artificial Intelligence, Machine Learning, and Natural Language Processing research, systems, and products in European and Canadian institutions and companies. Currently, he works as a Data Scientist at Adlib Software.