**Invoice Recognition**

*Țăpuc Andrada-Ionela*

*Bodnar Ciprian-Ioan*

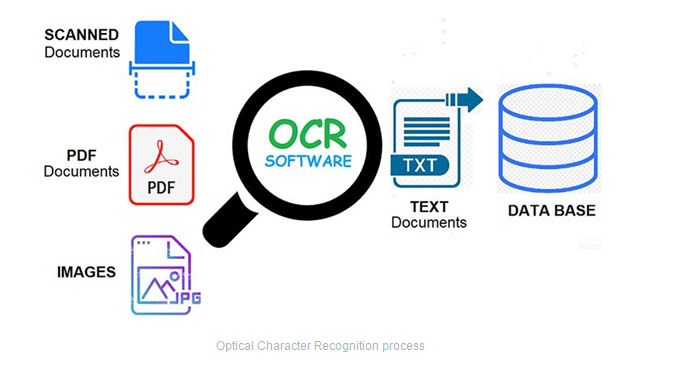
*Pintilie Cosmin*

**Project’s short description** **– project’s goal***:*

For an image with an invoice in any format, the aim is to extract important and relevant data from it, so that it can be easily observed in an excel file by a company's accounting.



**Architecture (schema):**

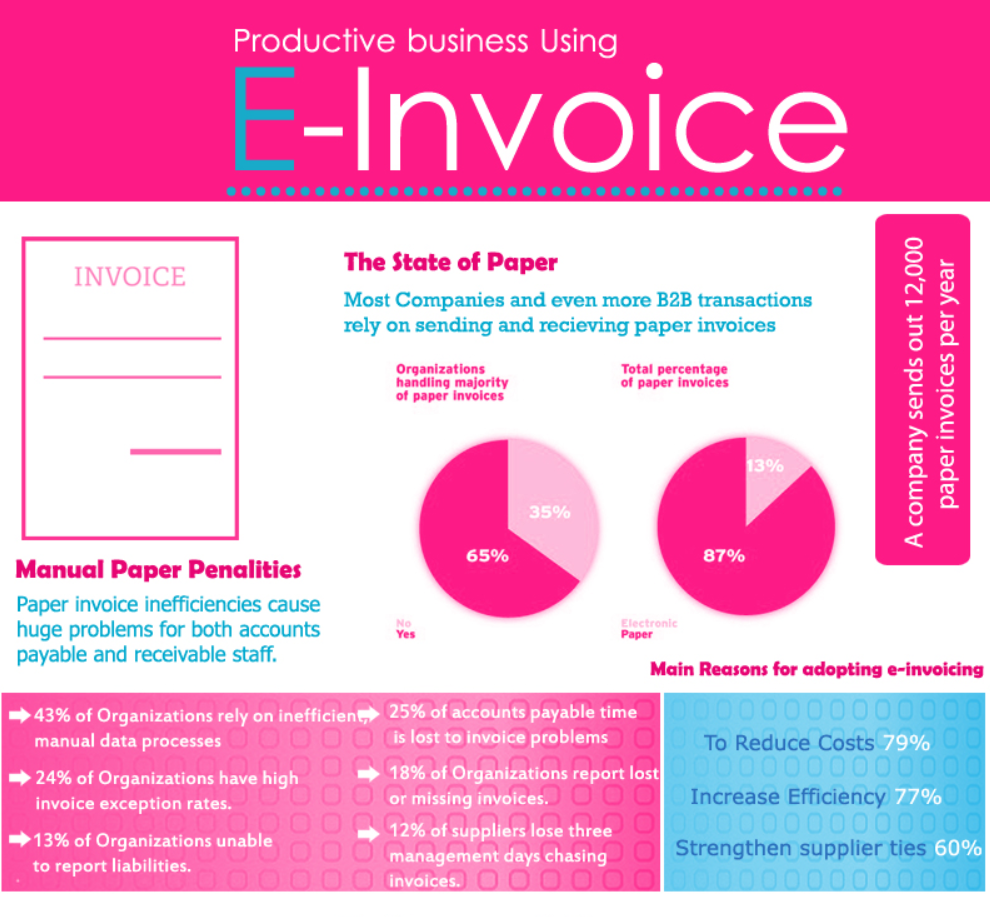


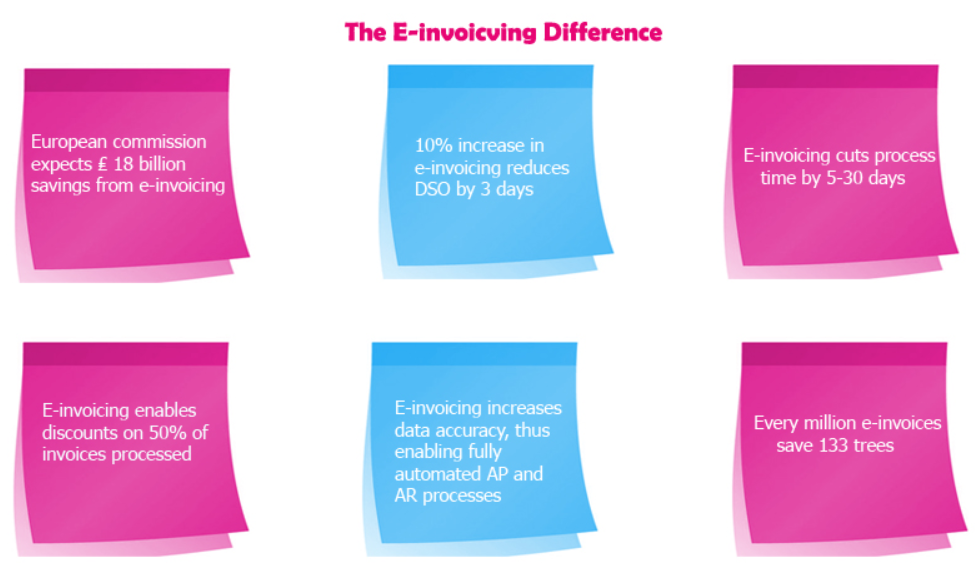


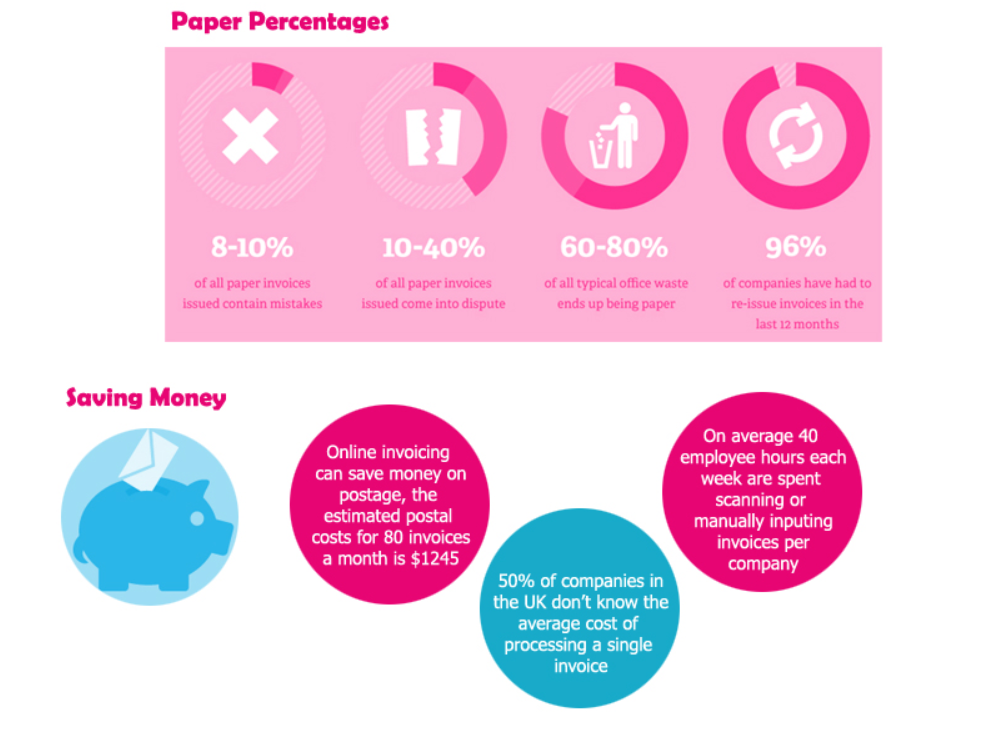
**Project’s details:**

* **Motivation:**

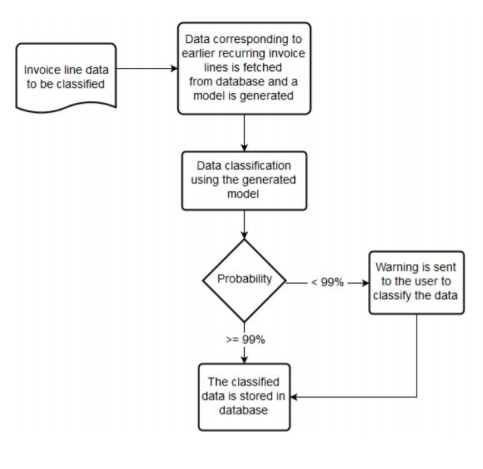
## Why digitize invoices? Digitizing information has several advantages a business can gain on several grounds. Businesses can track their processes better, can provide better customer service, improve the productivity of their employees and reduce costs.







* **Challenges:**
* **State of the art:**
  + “*Automated invoice handling with machine learning and OCR*”[[1]](#footnote-0) by Andreas Larsson and Tony Segras. By interpreting invoices with the OCR-engines, it results in the output text having few spelling errors. However, the invoice structure is lost, making it impossible to interpret the corresponding fields. If Naïve Bayes is chosen as the algorithm for machine learning, the prototype can correctly classify recurring invoice lines after a set of data has been processed. The conclusion is, neither of the two OCR-engines can interpret the invoices to plain text making it understandable. Machine learning with Naïve Bayes works on invoices if there is enough previously processed data. The findings in this thesis concludes that machine learning and OCR can be utilized to automatize manual labor.



* **Objectives:**
* **Short description of each module in the architecture:**

OCR = Optical Character Recognition. In other words, OCR systems transform a two-dimensional image of text, that could contain machine printed or handwritten text from its image representation into machine-readable text. OCR as a process generally consists of several sub-processes to perform as accurately as possible. The subprocesses are:

* Preprocessing of the Image (Tesseract and VisionApi)
* Text Localization (Header, Footer, Table)
* Character Segmentation
  + The invoice, buyer and supplier data can be found in the header part.
  + The product data can be found in the table.
  + The total invoice can be found in the footer part.
* Character Recognition ( keywords to make the search easier)
* Post Processing
* **Description of the integration of different modules**: problems foreseen, if any (incompatibility of programming languages, input/output format changes, etc.).

Knowing that we will use Tesseract, we chose the Python programming language, so as not to have incompatibility problems. An existing problem is the structure of the invoice, because not all invoices have the same pattern. Another problem is the very intense preprocessing of very degraded and difficult to understand images.

**Project management** (no imposed max. words no.):

* Assignment of team members for each different module (one team member per module)
* Will the project have a team leader? If yes, who is it? What are its main duties?
* **Detailed deadlines and milestones**
  + **4 March 2020**: Each member of the team must document himself and make a "State of art" in connection with the theme of the project, the optical recognition of invoices.
  + **12 March 2020:** Realization of the project architecture.
  + **19 March 2020**:
    - **Andrada**: Create the database with images of invoices.
    - **Ciprian:** Invoice extraction with Vision Google Api
    - **Cosmin**: Invoice extraction with Tesseract
  + **26 March 2020**: Creating a language model based on the CoRoLa corpus.
  + **9 April 2020**:
    - Dividing into zones (header, footer, table)
    - Parsing the table
  + **16 April 2020**:
    - Achieving the final excel
    - Correctly extract data from the table
  + **30 April 2020**:
    - **Andrada**: Extract data from header + Invoice type.
    - **Ciprian:** Extract data from table.
    - **Cosmin**: Extract data from footer.
  + **7 May 2020**: Versiune Beta - integrare totala
    - tabelul(coloanele separate cu un separator)
    - integrarea totala
  + **14 May 2020**:
    - Google Api - împărțirea pe linii și text blocks (identificare randuri tabel)
    - program de evaluare ( cate 3 facturi - se fac datele corecte și apoi se verifica)
  + **9 June 2020:** Forma finala
* Qualitative measures for the evaluation of the project (e.g. no. of annotations, no. of sentences processed, parsing accuracy/recall/precision, etc.). These measures **will** be used to assess the projects.

With this project we aim to achieve an accuracy of at least 70% in extracting information from invoices.

1. <http://www.diva-portal.org/smash/get/diva2:934351/FULLTEXT01.pdf> [↑](#footnote-ref-0)