# CRP FINAL PRESENTATION:
# *ILLUIN TECHNOLOGY*

# AGENDA


ILLUIN TECHNOLOGY

# WHAT IS THE PROJECT ?

# CORPORATE RESEARCH PROJECT (CRP) OBJECTIVES

Our goal is to **deliver a pipeline** that facilitates the creation of a **database** from **web-scraped data** using **NLP** techniques. Illuin seeks to have a **technical proof-of-concept** but does not have yet a business use case.

**What objectives should we focus on to bring some added value?**

**Real use case**

The project should respond to potential existing needs of future clients

**Visual tool**

The project should be easily understandable and facilitate the use of a database and dashboard

**Reusable**

The project should be as automated as possible and be reusable for other similar needs

**Model benchmark**

The project should propose a benchmark of different NLP techniques envisioned and used in the process

# PROJECT SCOPE DEFINITION

**Initial project**
The project was initially focused on building a database of **eco-startup**

## OBSTACLES

1

**Difficulty of classification**
In theory we built a scale to determine the level of eco-friendliness, but in practice it was difficult to evaluate this level just based on a few lines of an article, and thus to annotate.

2

**Lack of sources**
Few sources report on eco-friendly start-ups. When webscraping BusinessWire a well-known journal for start-up news, out of 250 articles only 1 was eco-friendly.

**New project**
The project was initially focused on building a database of **biotech startups**

## WHY

**$41B** RAISED FUNDINGS OF BIOTECHS STARTUPS IN 2021

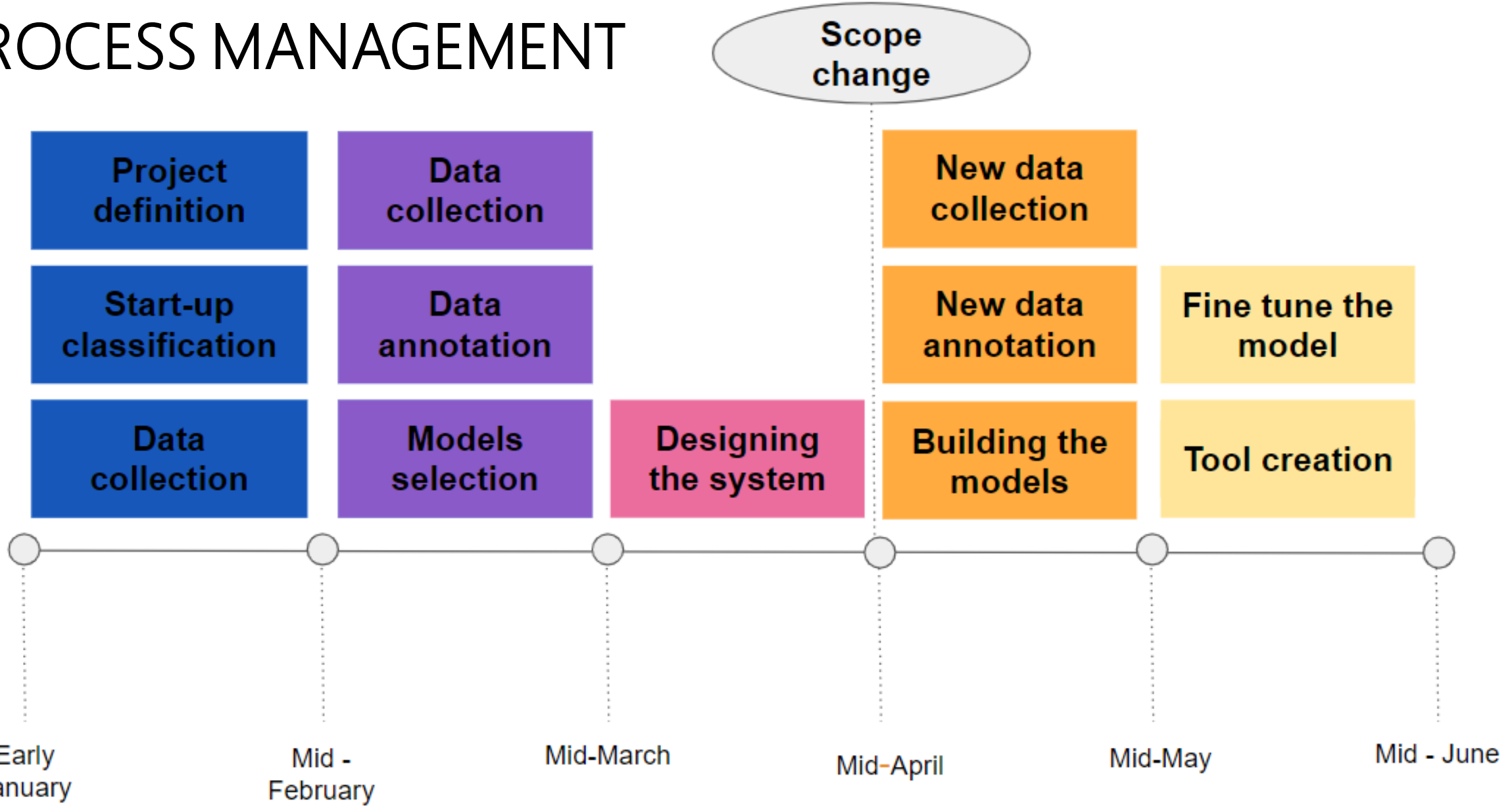
**4.1x** INCREASE IN BIOTECHS STARTUPS VALUATION SINCE 2017

**1st** SECTOR OF INVESTMENT FOR V.C. FIRMS IN U.S. FOR 2 YEARS

Focusing on biotech startups allows to :
- Respond to the **needs of V.C. firms**
- Have access to **numerous sources**
- **Classify** easily startups

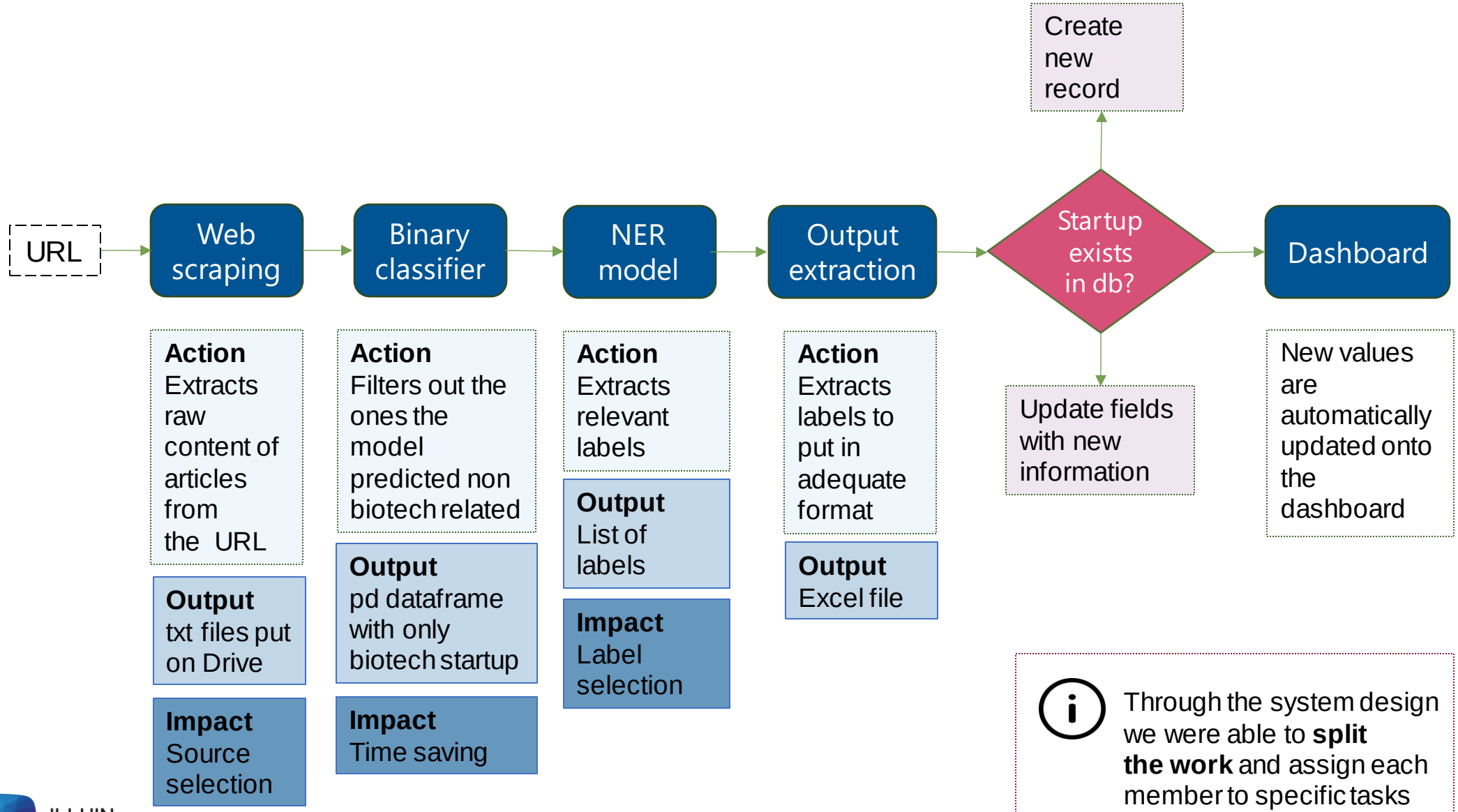
ESSEC BUSINESS SCHOOL
ILLUIN TECHNOLOGY

# PROCESS MANAGEMENT




ILLUIN TECHNOLOGY

# PROCESS MANAGEMENT THROUGH SYSTEM DESIGN




ILLUIN TECHNOLOGY

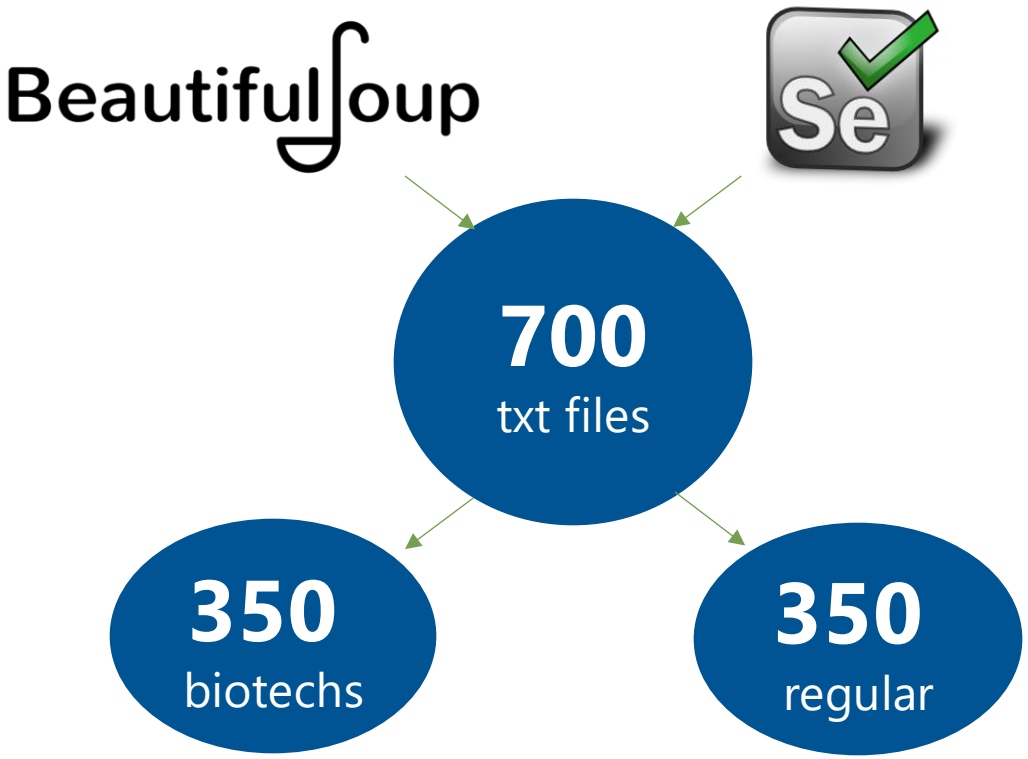
# WHAT ARE THE TECHNICAL ASPECTS ?

# SOURCES - WEBSCRAPING

**Sources**
Although web sources introduce some **bias**, we tried to limit this by using **diverse sources**, so the model learns from **different formats** of news and presents **heterogenous startups.**

**More than 10 sources...**

sifted / backed by FT FINANCIAL TIMES
BioPharma -Reporter.com
.SEEDTABLE
businesswire A BERKSHIRE HATHAWAY COMPANY
exploding topics
Startups.Bio
TechCrunch
LABIOTECH.eu
tech eu




**Webscraping**
We built webscraping tools to extract 700 text files using both the BeautifulSoup and Selenium libraries

ESSEC BUSINESS SCHOOL
ILLUIN TECHNOLOGY

# BUILDING A BINARY CLASSIFIER

## Model choice

We chose to use a **BERT Binary classifier** for the following reasons

1

**Better understanding of the context**

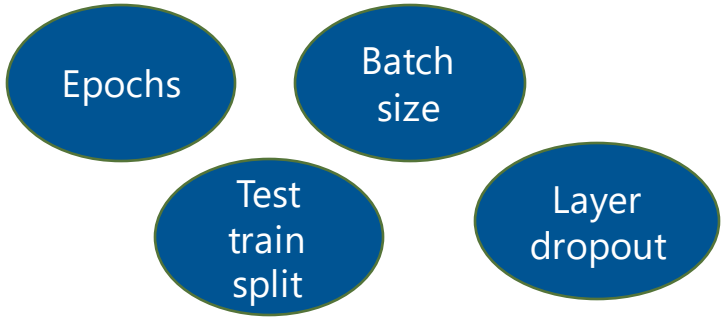
Through the **bidirectional** characteristics of BERT, the model **takes into account the surroundings of the words**, thus the context of the text is learned. It is a strong advantage as articles can contain multiple words, which would confuse classical models and context would be lost.

2

**Better performance**

Through the use of a **transformer and its two sub-layers**, the BERT Binary classifier usually has better performance than state of the art models.

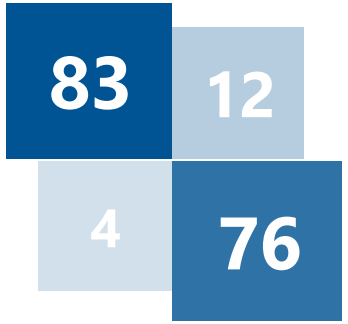
## PARAMETERS TO TUNE




## PERFORMANCE RESULTS

**91%** of accuracy

## CONFUSION MATRIX




## TYPE I & TYPE II ERRORS

**Variation according to chosen cut of predicted %**

Depending on the chosen cut of the prediction probability (output of the model), type 1 and type 2 errors will vary. The higher, the higher false negative we get and the opposite when it gets lower. We advise in terms of business to **limit as much as possible false positive errors**, and thus to **keep a high cut rate.**

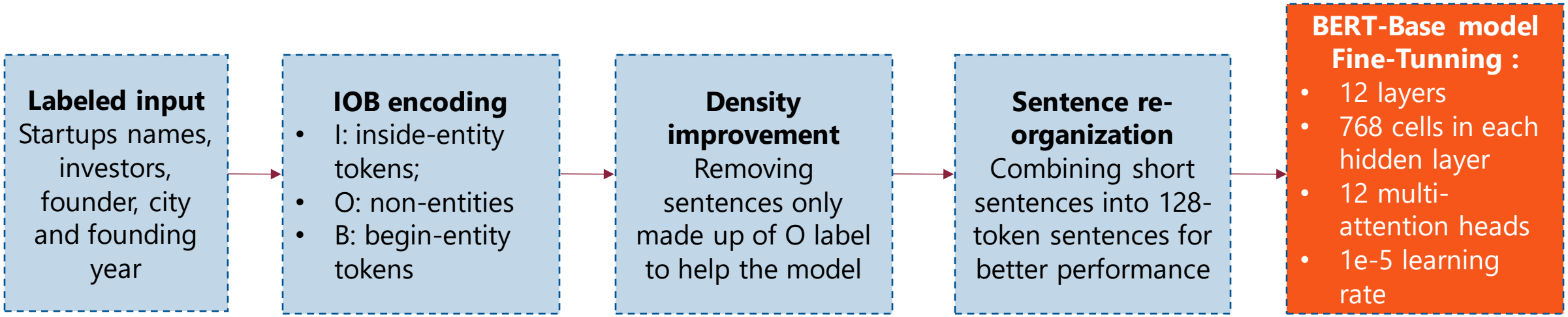
**Type I errors**

Most of the type I errors have some "health" related words. Examples: lifespan (recycling start-up), healthy (food tech)

**Type II errors**

There seem to be no explicit reasons for type II errors. Some of them have biotech words some don't, some are short and some are long and complex. They usually have a prediction probability between 0.4 and 0.5.

# NER




## PERFORMANCE RESULTS

**91%** f1 score average for predicting the labels

## BEST & WORST PERFORMANCE

**90%** f1 score for predicting the year at the beginning of a chunk

**27%** f1 score for predicting the place inside a chunk

## LIMITATION

**0%** f1 score for predicting the year at the inside of a chunk, this is due to some irregularities in annotation as some of us took both the month and the year and not only the year resulting in an inside chunk

ESSEC BUSINESS SCHOOL
ILLUIN TECHNOLOGY

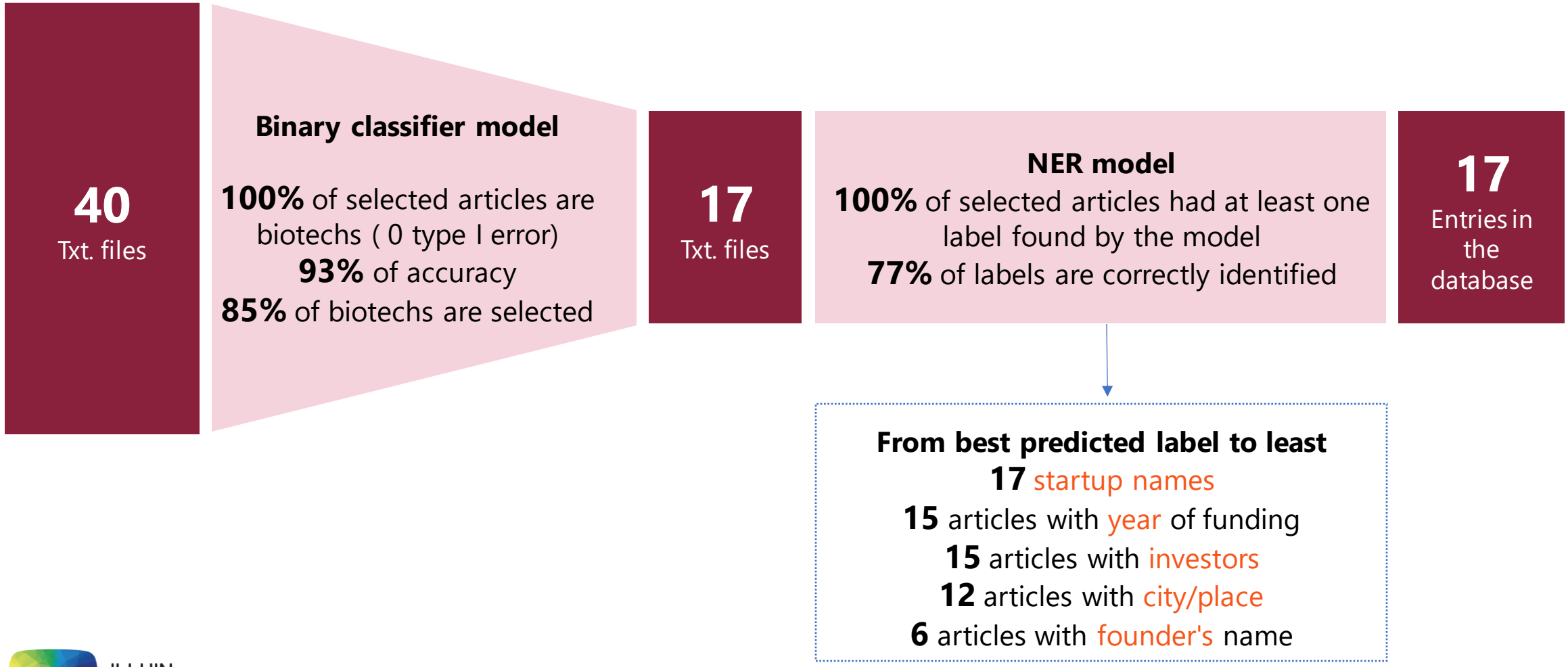
# LIMITATIONS

| | SOURCES | DATA ANNOTATION | AUTOMATION | MODEL UNDERSTANDING |
|---|---|---|---|---|
| PROBLEM | Sources have a s**trong impact on the model performance, source of bias.** | No strict rule of annotation resulting in **difference in labels** for each annotater | The proposed system is not entirely automated. | The **deep learning** aspect of the model is an obstacle to **understand how it works** (cf. Type 1 & 2 errors) |
| IMPACT | **Hazard in performance, additional work** in source selection. | **Hazard in performance** depending on the label. | **Need to regularly run the pipeline,** time consuming task. | Lack of means **to avoid false positive errors.** |
| SOLUTION FEASABILITY | Low | High | Medium | Low |

ILLUIN TECHNOLOGY

# OVERALL PERFORMANCE

## EVALUATION SET




ILLUIN TECHNOLOGY

# WHAT VALUE THROUGH THIS PROJECT ?

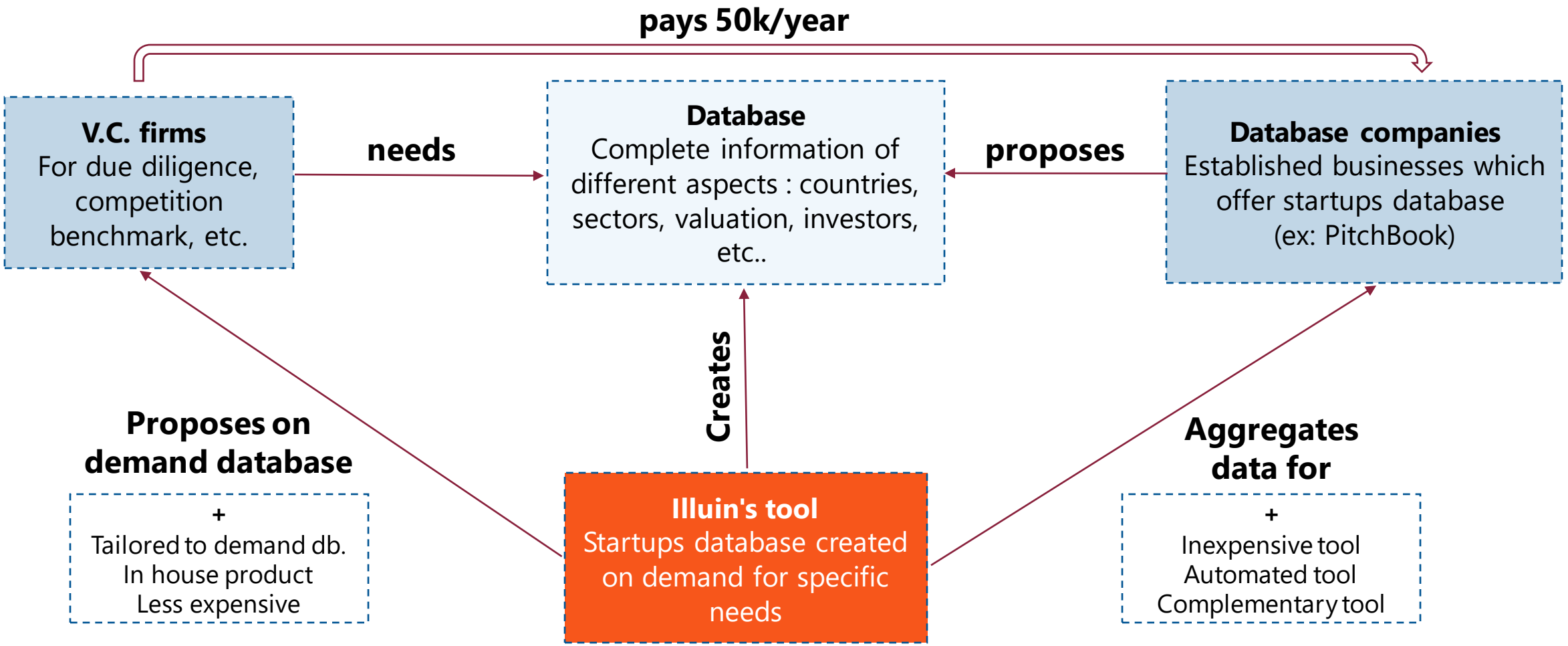
# FURTHER RESEARCH OPPORTUNITIES & BUSINESS RECOMMENDATIONS

Our model is up and running, but **only applies to one use case**. To bring some value and consistent revenues, we recommend to **develop certain axis of research** which would complete our work in order to propose a **tool which applies to multiple client requests.**

| RESEARCH AXIS | | ADAPTATION TO CLIENT SPECIFIC REQUESTS |
|---|---|---|
| **LABELS** | Develop other labels to predict : amount raised, funding round, etc. | **Example**<br><br>V.C. firm wants to invest in Series B funding, in the U.S. for start-ups evolving in biopharmacy sector or fintech sector. |
| **CLASSIFIER** | Develop a sector classifier maintaining the BERT structure | |
| **SOURCES** | Develop specific feature engineering and data preprocessing for complex news articles | |

# TANGIBLE VALUE OF THE PROJECT

Our project has a **concrete business value** behind it and can **generate some revenue streams** for Illuin. Through our model we can propose a tool which **responds to an existing need** of potential clients which are **both the V.C. firms and database companies**.



# THANK YOU!

# ADDITIONAL CONTENT

# WORKFLOW


URL

Web scraping

Web scraping of various websites using BeautifulSoup & Selenium libraries

Binary classifier training

BERT Binary classifier using the transformers library

Annotation

On *Etiquette*
→ Name
→ Investor name
→ Location
→ Year
→ Sector
Data corresponds to different news articles along with IOB tags for each words

Data loader

IOB encoding

- I: inside-entity tokens;
- O: non-entities
- B: begin-entity tokens

Density improvement

Removing sentences only made up of O label to help the model

Sentence re-organization

Model training

Using a pre-trained BERT on our data to predict the labels, and then compute the accuracy of our model

Fine tuning

Using AdamW HuggingFace as optimizer to fine tune our model

Model testing

Test model on test data and compute accuracy (only on active tags) to compare with training data

Results too low?

Fine tune the model on a larger dataset than ours, modify the architecture

Output extraction tool

Decide on a tool which automatically or not extract active labels

Database creation

Create a Excel file to store all the extracted elements

Database creation

Based on the database choose an end-tool (Tableau, PowerBI) and create a dashboard

ESSEC BUSINESS SCHOOL

ILLUIN TECHNOLOGY

# BINARY CLASSIFIER RESULTS

## FINAL PARAMETERS

**25%** Test-train split

**30** Epochs

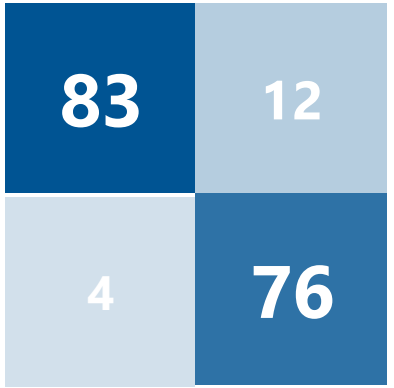
**24** Batch size

**0.1** Layer dropout

## FINAL SCORES

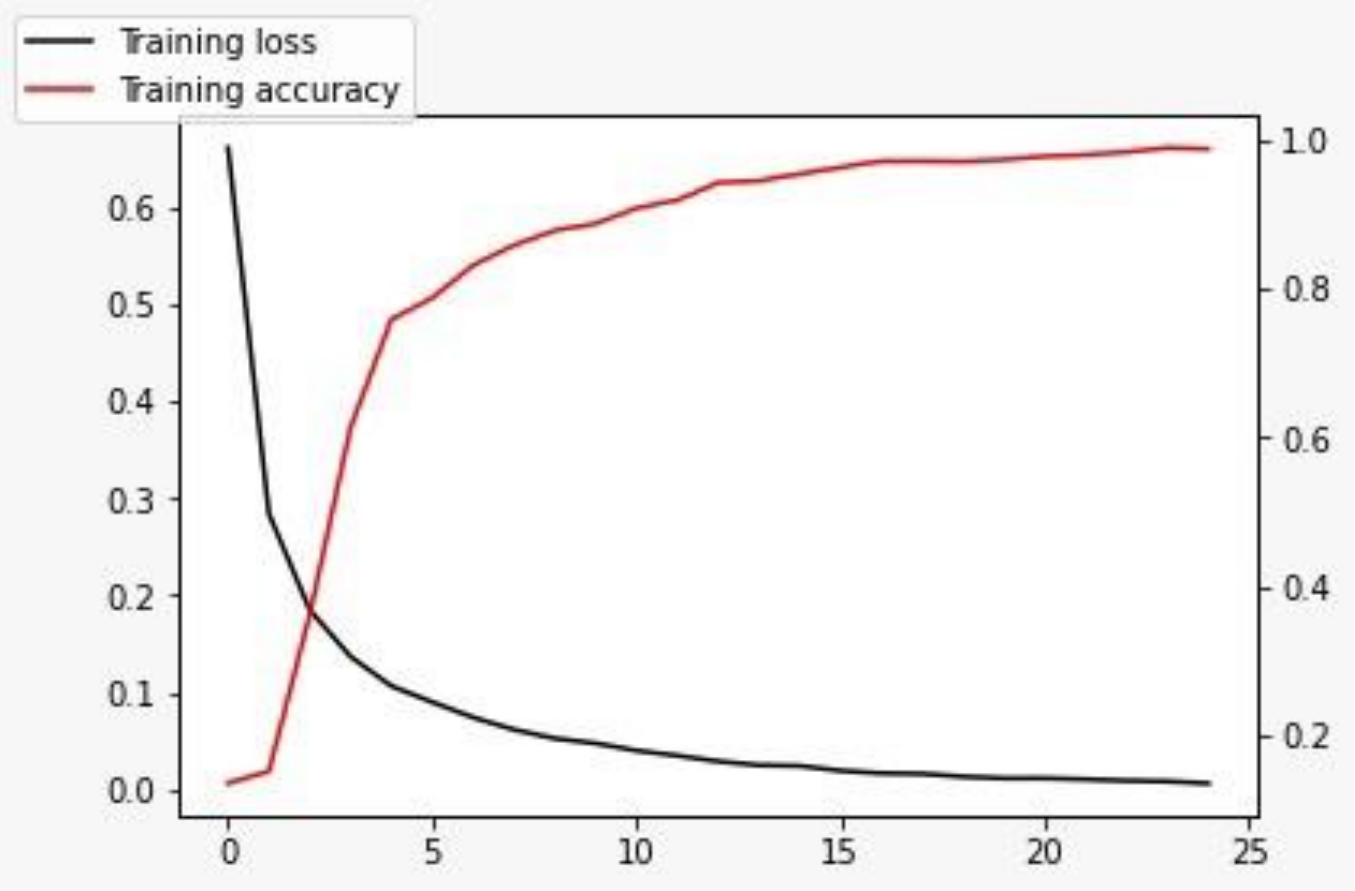
| | F1-Score |
|---|---|
| 0 | 91,2% |
| 1 | 90,5% |
| Accuracy | 90,9% |

## CONFUSION MATRIX




ILLUIN TECHNOLOGY

# NER TRAINING & LOSS ACCURACY




ILLUIN TECHNOLOGY

# FINAL DELIVERABLES

## CODE RELATED DELIVERABLES

- Web-scraping files
- Final file
- Binary classifier file
- Binary classifier model (.h5)
- NER model file
- NER model (.bin)
- 700 articles

## OUTPUT DELIVERABLES

- Final report
- Final presentation
- Database
- Dashboard file

ESSEC BUSINESS SCHOOL

ILLUIN TECHNOLOGY