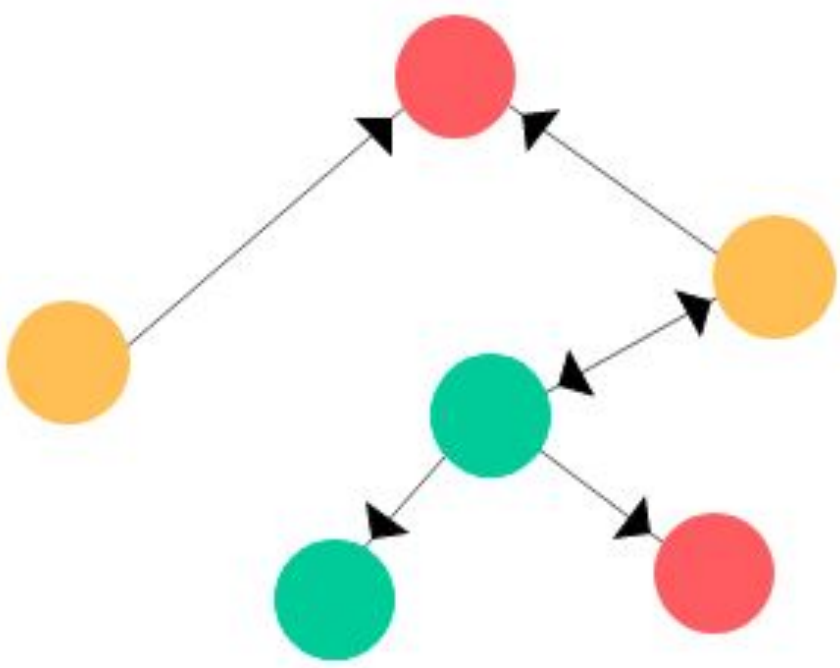
Citation prediction



Agenda

01 Context and objective

02 Methodology overview

03 Data exploration and processing

04 Feature engineering

05 Modeling

06 Tuning

07 Perspectives

01

Context and objective

Context and objective

Project Aim

This project aims to predict if two papers have a **citation link** by using machine learning or deep learning techniques. It is a binary classification task with two outcomes :

0 – no link

1 – existence of a link

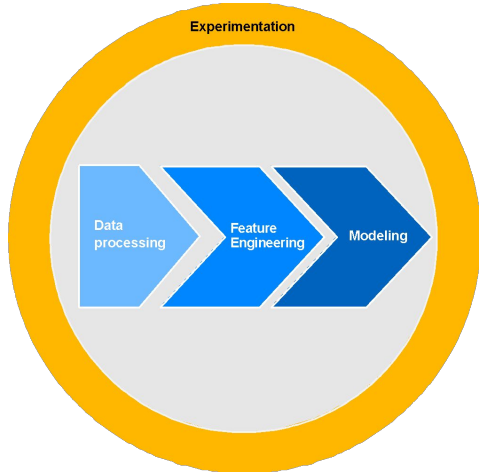
Challenge

3 Différents datasets to combine into a coherent piece. Egde list, abstracts and authors.

11 Languages detected in the abstracts (eng, fr, ca, etc)

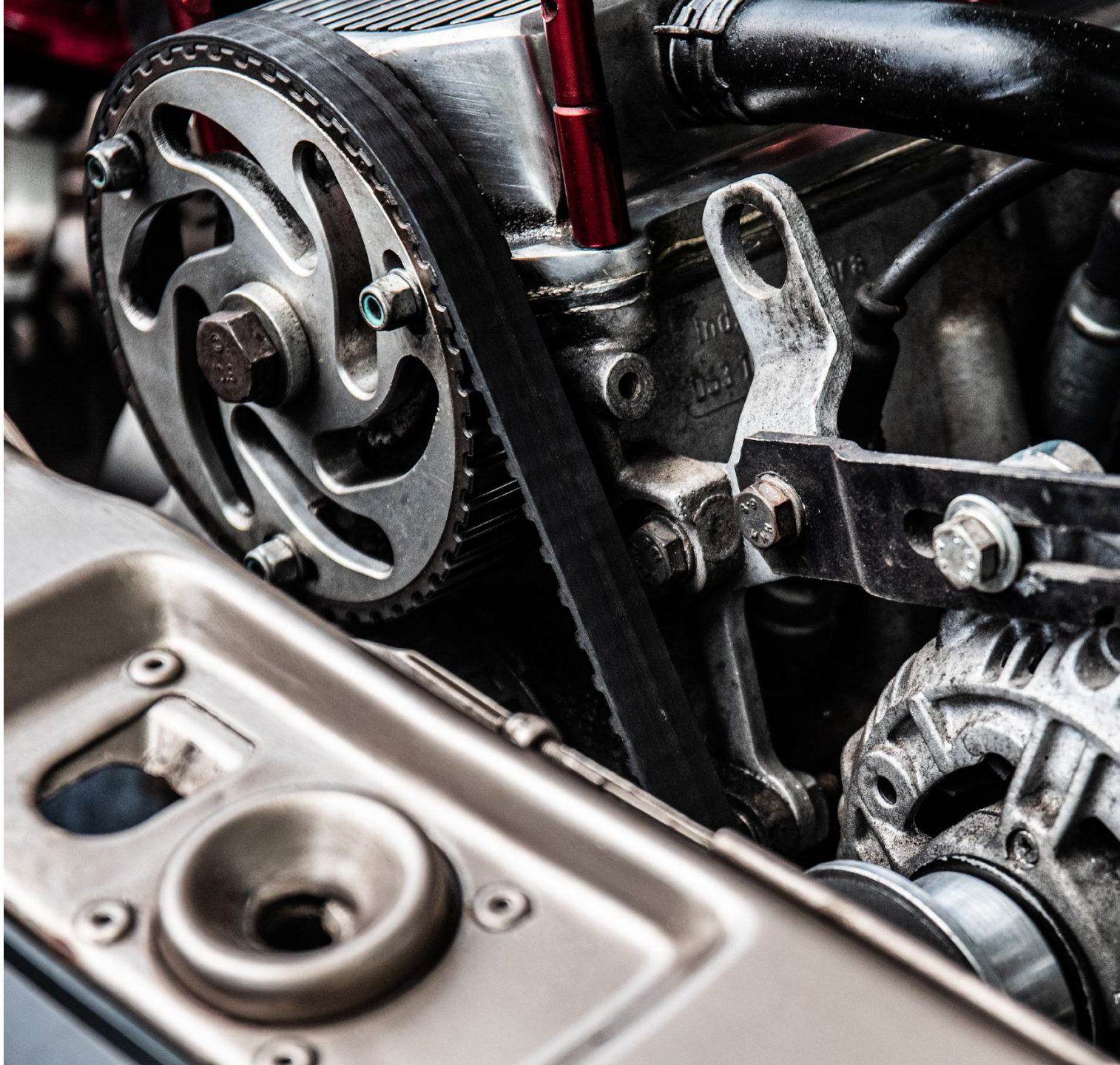
153 Authors for a single paper

Methodology

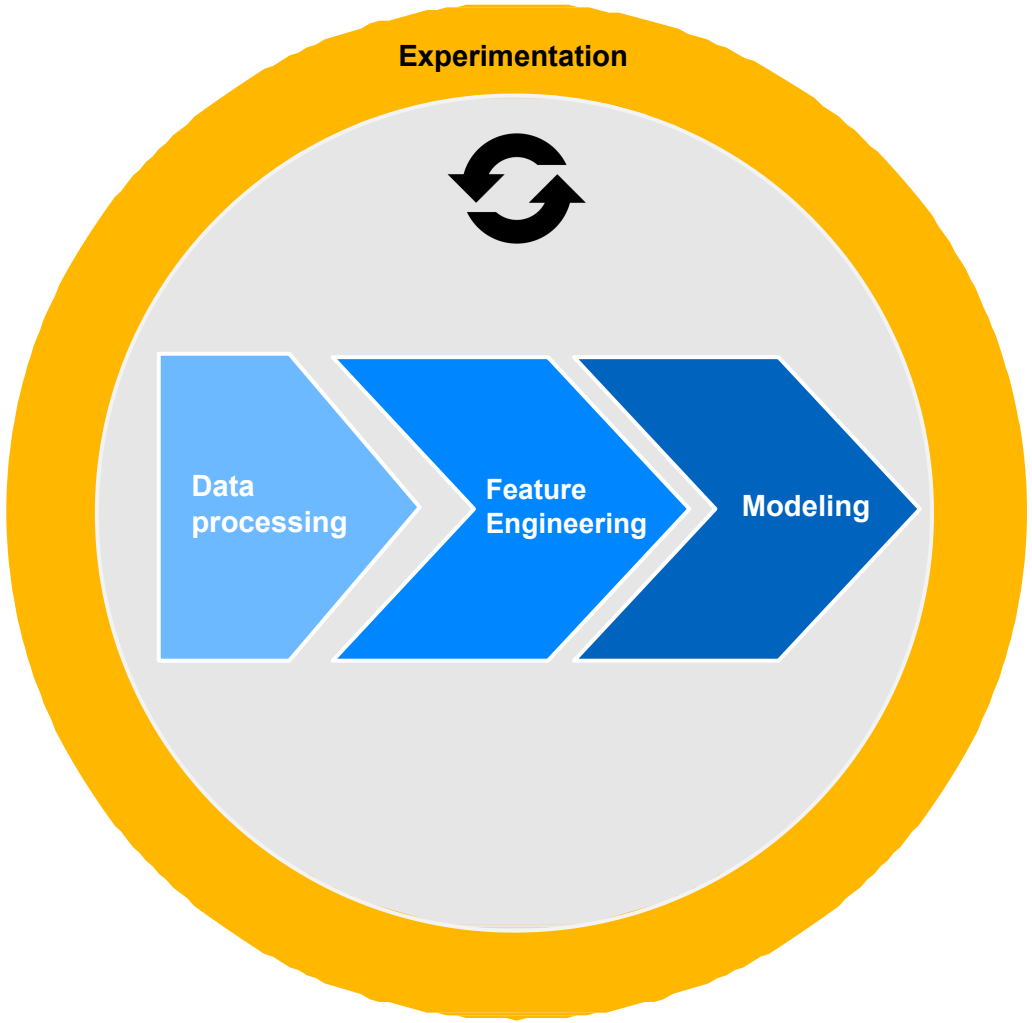


02

Methodology overview



Methodology overview



Data processing:
- Generate a validation set
- Removing stop words
- Preserving order of words for embeddings

Feature engineering:
- Extracting the most relevant features
- Eliminating unnecessary features

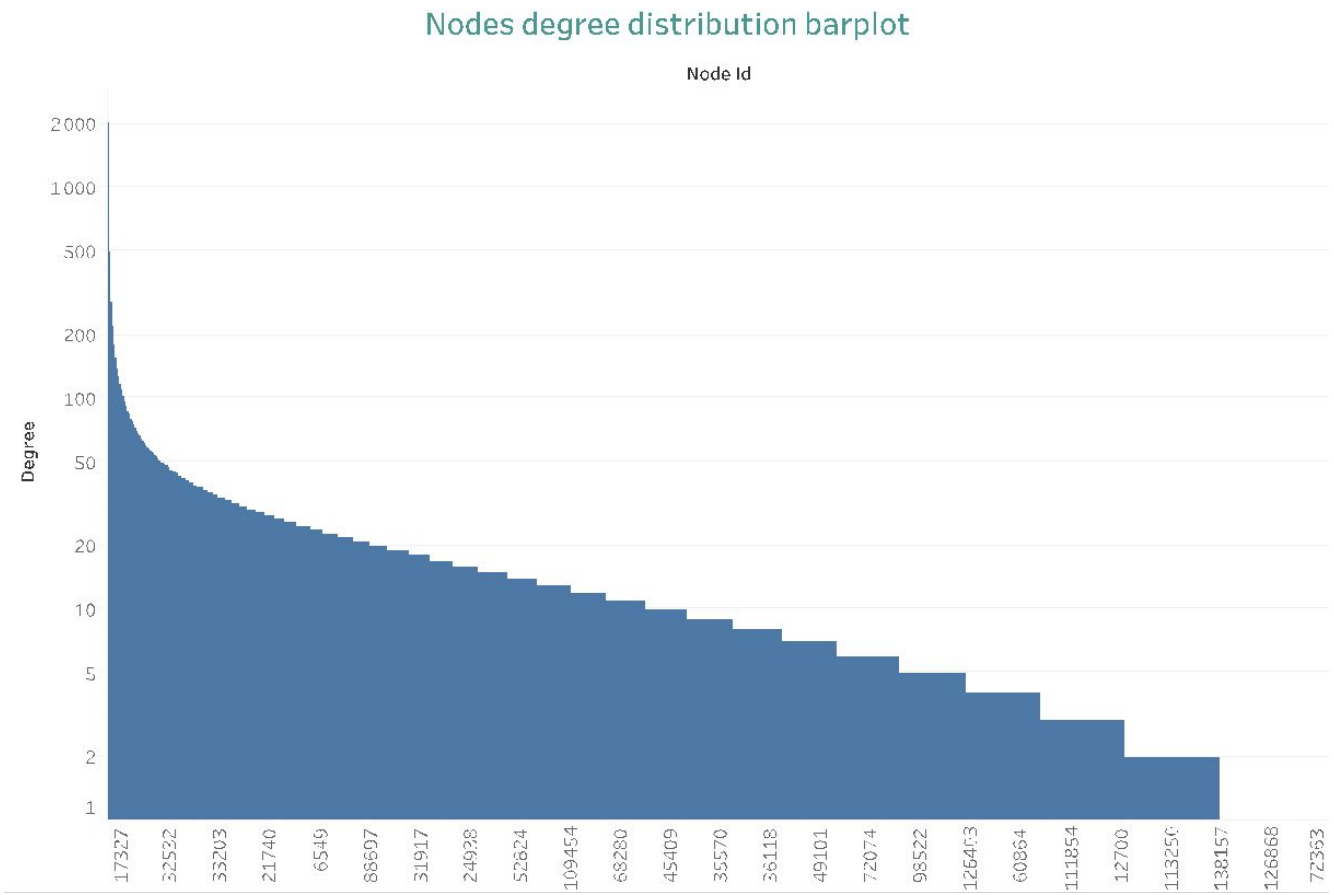
Modeling:
- Testing several algorithms
- Tuning

Experimentation loops:
-Experiment differents approaches and learn

03

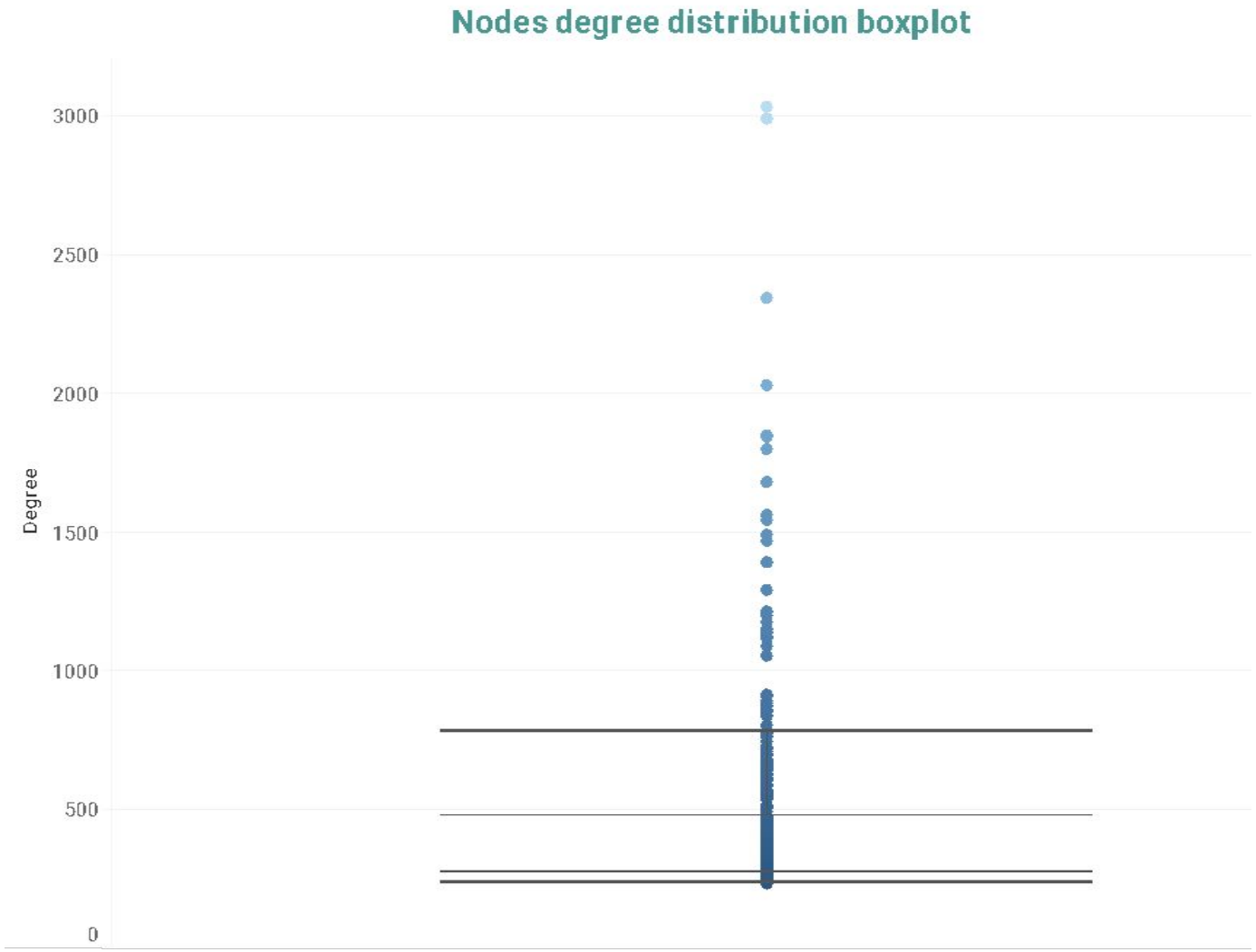
Data exploration and processing

Data exploration and processing



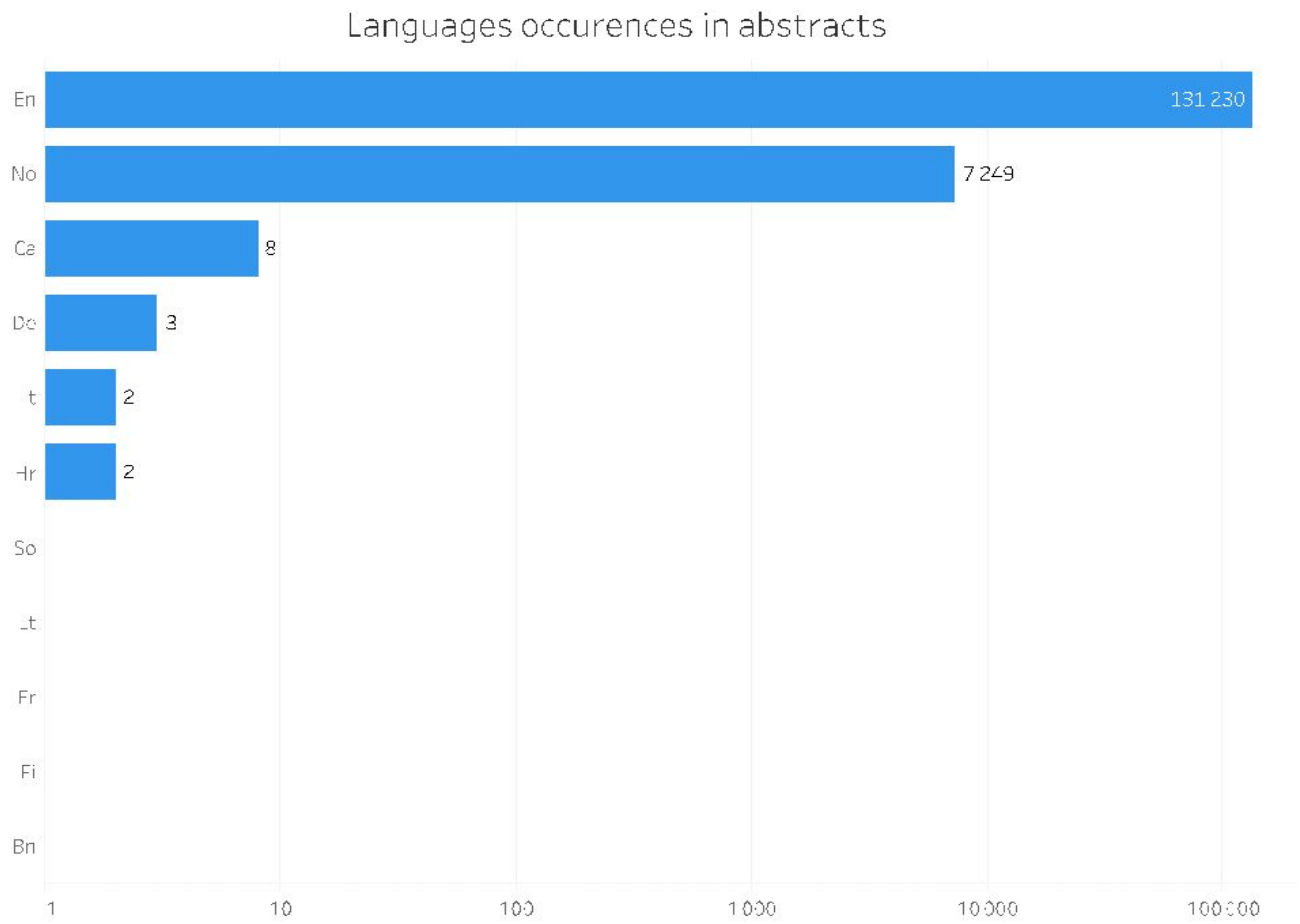
- Most papers have few degrees. 80% of papers have a degree inferior to 20 with a minimum degree of 1.
- Only a **small fraction** papers have a very high degree. 1% have a degree superior to 115. With a maximum degree of 3037
- Not a surprising observation

Data exploration and processing



- 75th percentile : 477
- Median : 337
- 25th percentile : 272

Data exploration and processing



- 11 detected languages in the abstracts. English is the most represented
- Language-specific processing technics

04

Features engineering

Feature engineering

Graph

- Sum of degrees
- Difference of degrees
- Jaccar coeffincient
- Adamic adar index
- Resource allocation
- Clustering coefficient
- Degree centrality
- Common neighbor centrality

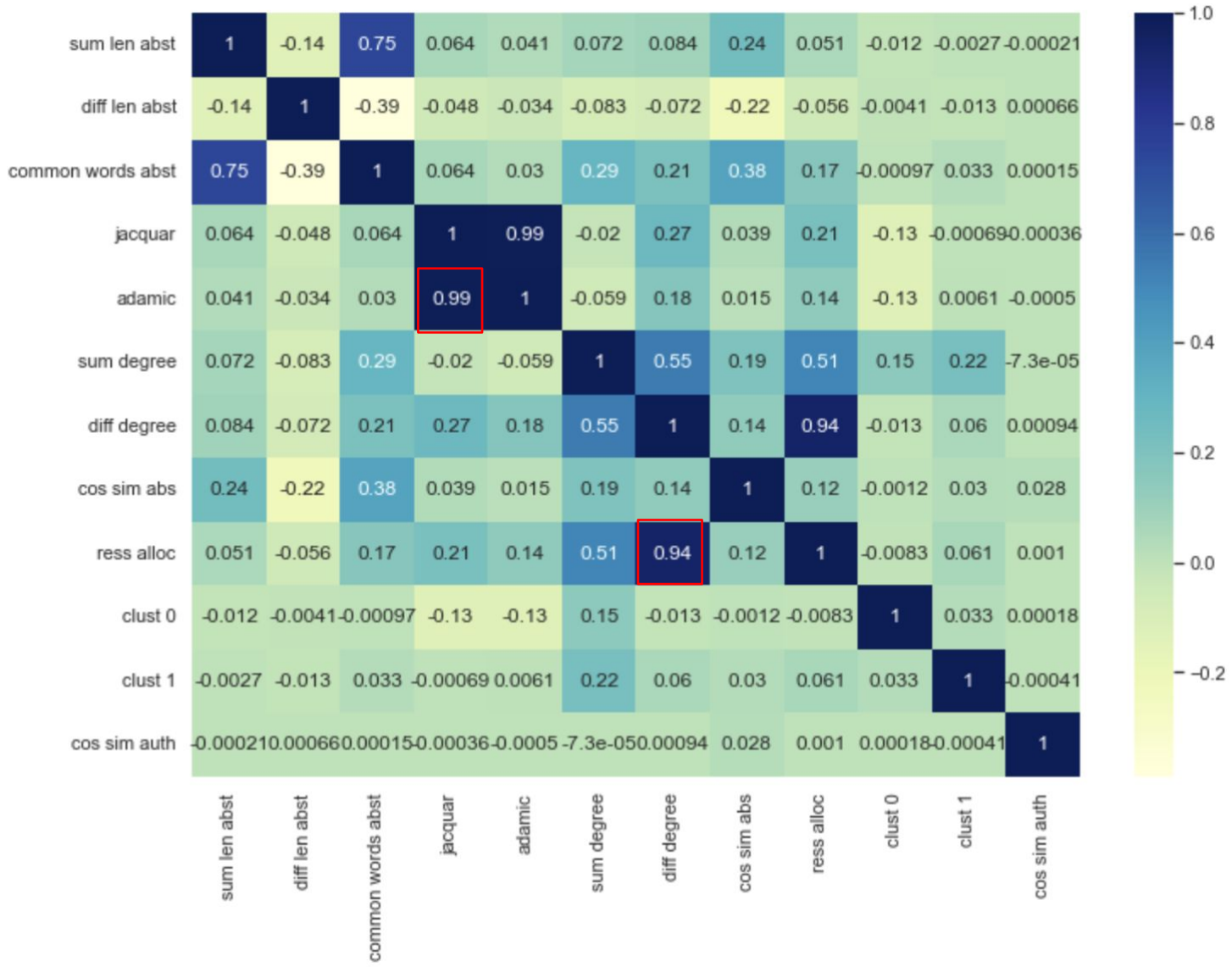
Abstracts

- Sum of unique words
- Difference of uniques words
- Common words
- Doc2vec cosine similarity

Authors

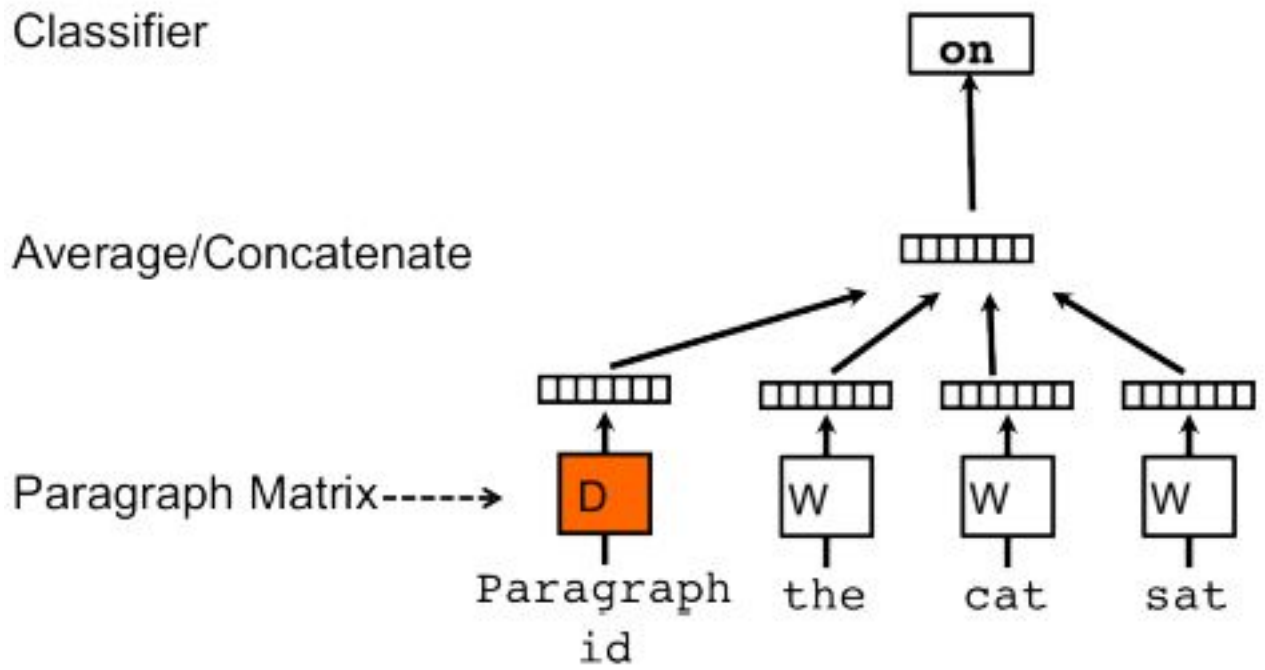
- Doc2vec cosine similarity

Feature engineering



- Correlation of **0.99** between '**adamic adar index**' and '**jaccard index**'
- Correlation of **0.94** between '**ressource allocation**' and '**difference of degrees**'
- Keeping the combination that produice the best results

Doc2vec

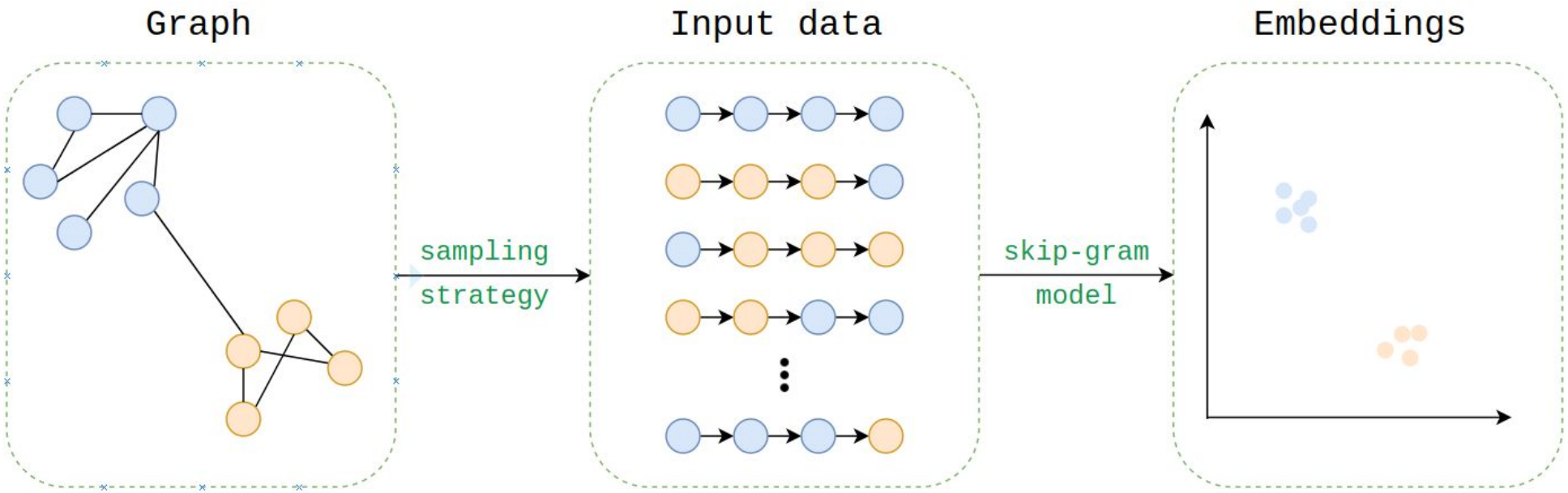


- Most digital transformation processes of digitally mature manufacturing firms are **management** or **innovation-driven.**
- Only a **small fraction** of digitally mature companies have used **IT** to transform.
- There are **no consistent approaches** visible in the process of digital transformation.

Doc2vec VS Tf-IDF

Doc2vec	Tf idf
Dimensionnality reduction Paragraph embedding Work well for large corpora. Many to many mapping between input and output. Cut-off : size of the vector for the new dimension Pr(x,y) of for a word to appear in the context of another word Most suitable for a wide ranging corpus content with no common vocabulary	Compute each tearm tf-idf Most suitable for small corpora One to one mapping between input and output Cut-off : tf-idf threshold . Pr(x) of a word to appear Most suitable for a focused corpus content with a « core » vocabulary

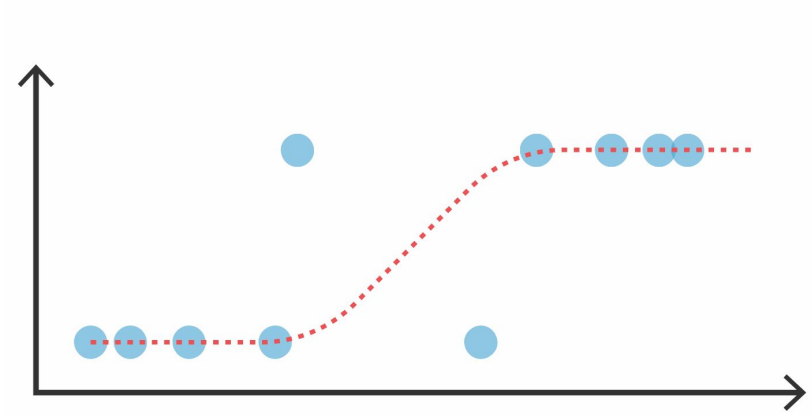
Node2vec


Graph
Input data
Embeddings
sampling
strategy
skip-gram
model

05

Modeling

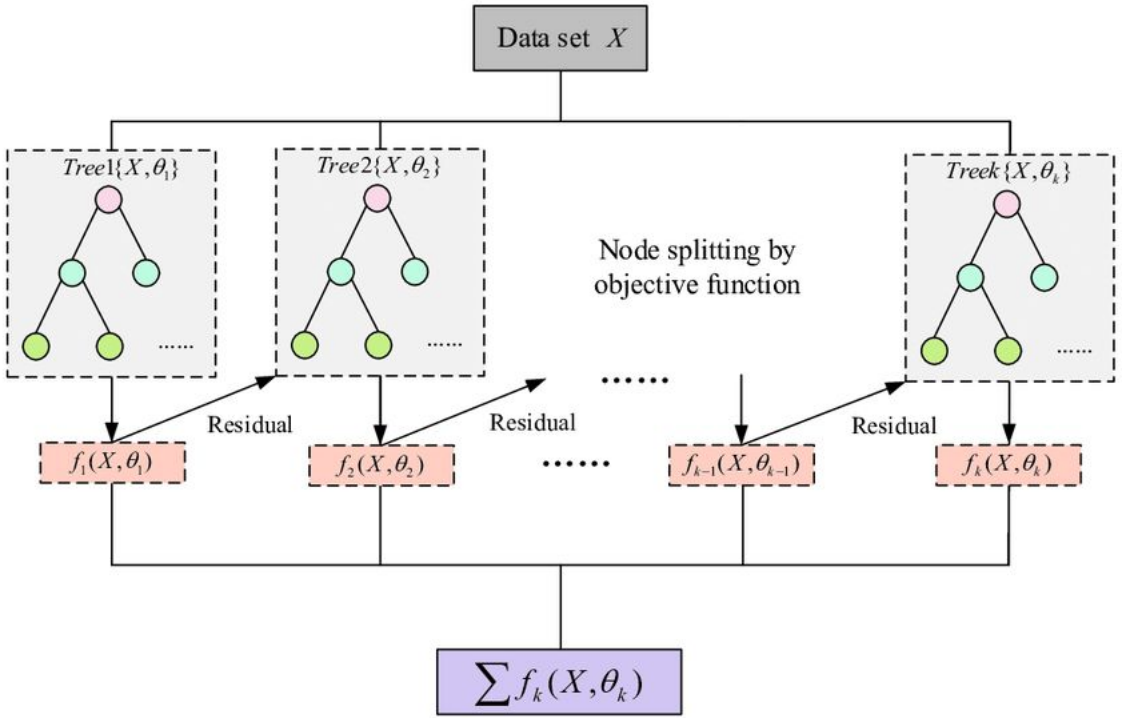
Logistic regression



	precision	recall	f1-score	support
0.0	0.85	0.91	0.88	174503
1.0	0.90	0.84	0.87	174503
accuracy			0.87	349006
macro avg	0.87	0.87	0.87	349006
weighted avg	0.87	0.87	0.87	349006

Kaggle : 0.22935

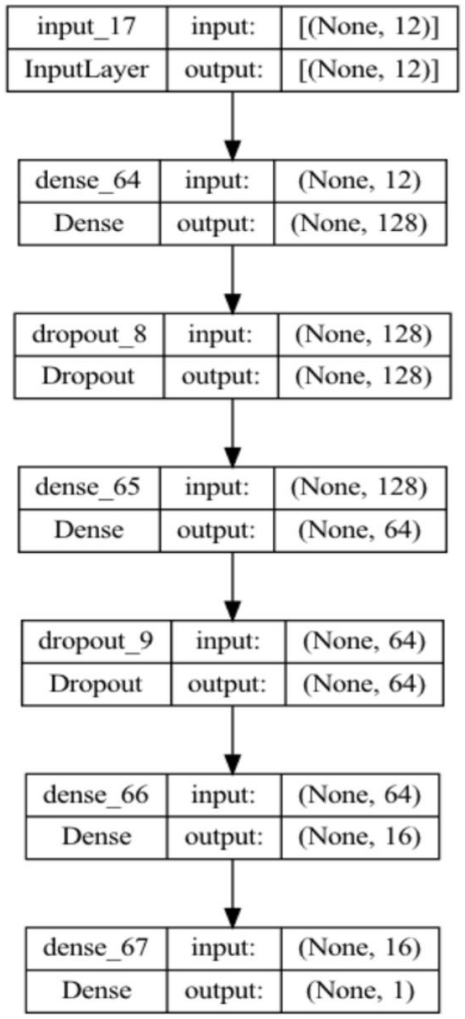
XgBoost classfier

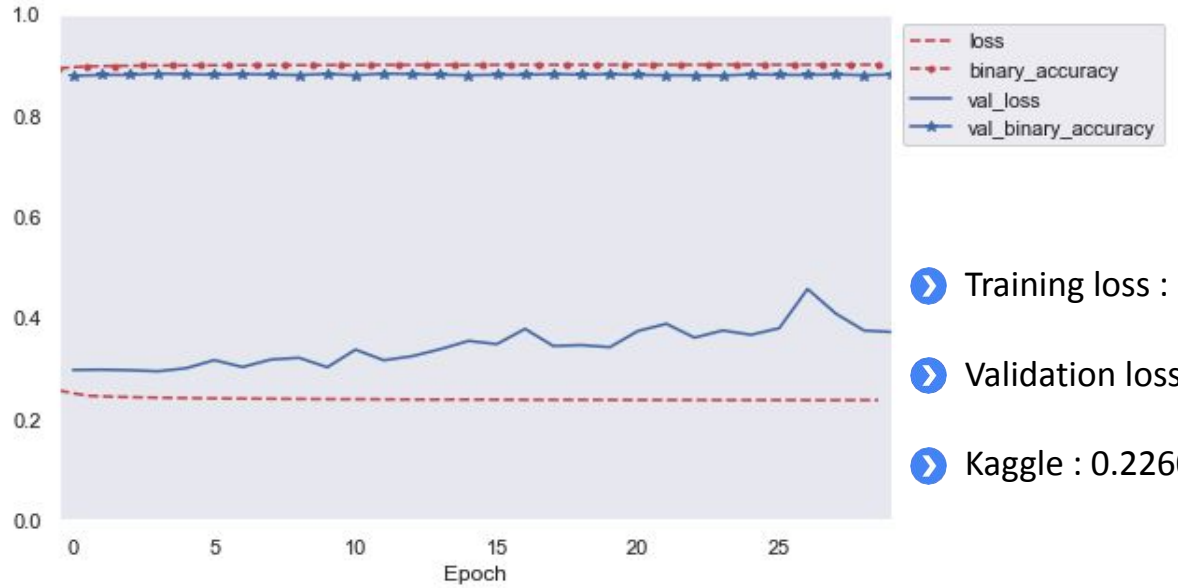


	precision	recall	f1-score	support
0.0	0.85	0.91	0.88	174503
1.0	0.90	0.84	0.87	174503
accuracy			0.87	349006
macro avg	0.87	0.87	0.87	349006
weighted avg	0.87	0.87	0.87	349006

Kaggle : 0.22601

Multi Layers Perceptron





- Training loss : 0.2552
- Validation loss : 0,2959
- Kaggle : 0.22601

	precision	recall	f1-score	support
0.0	0.86	0.91	0.88	174503
1.0	0.90	0.85	0.88	174503
accuracy			0.88	349006
macro avg	0.88	0.88	0.88	349006
weighted avg	0.88	0.88	0.88	349006

06

Tuning

Tuning

- Gridsearch for LR and XgBoost
- Manual tuning for MLP

07

Results and perspective

Perspective

- Node2vec
- Tf/idf
- Scibert
- Hyperparameter tuning for MLP