Institute of Mathematics and Physics (IMSP) , Bénin

06 -10 Novembre 2023

Automatic text classification of research results using Deep neural networks: an overview of classifyers

PAR :

Félicité G. DOMGUE

Université De Yaoundé I, Yaoundé - Cameroun

PLAN

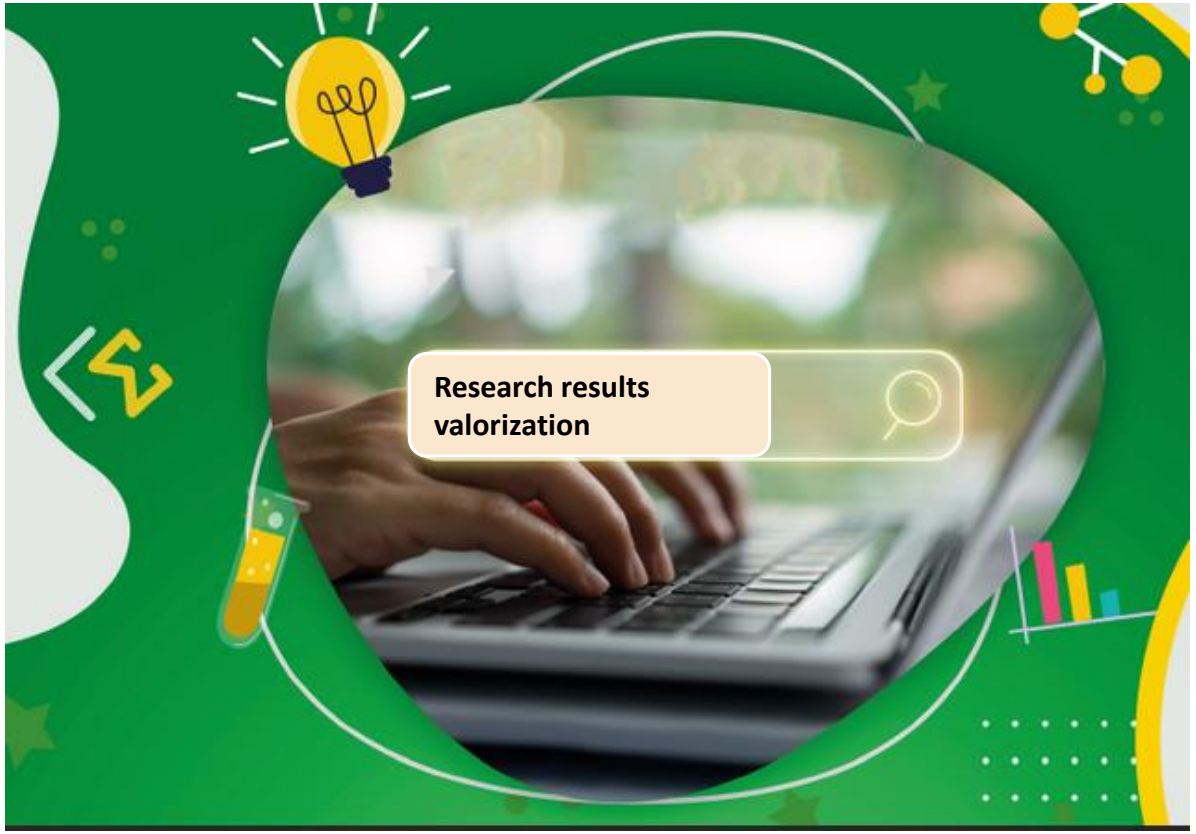
- Introduction
- Document modelling
- Classifier architectures
- Methodology of classifier comparison
- Experiments
- Conclusion and perspectives

Introduction
context

- 80 percent of text [Headmind Partners 22]
- With the exploding collection and storage of textual data, there is a growing need to analyze and extract relevant information from this huge volume.
- The rise of deep learning models for automatic natural language processing (NLP) has facilitated the use of textual data in operational problems:
  - Automatic text abstraction,
  - Question-answering,
  - Similarity analysis,
  - Document classification,
  - and more.

Introduction
context

- The valorization of results from research is the process of:
  - highlighting the knowledge and technologies resulting from research,
  - and making them available to researchers, markets, companies and government agencies
  - the economic growth and societal development of a nation.

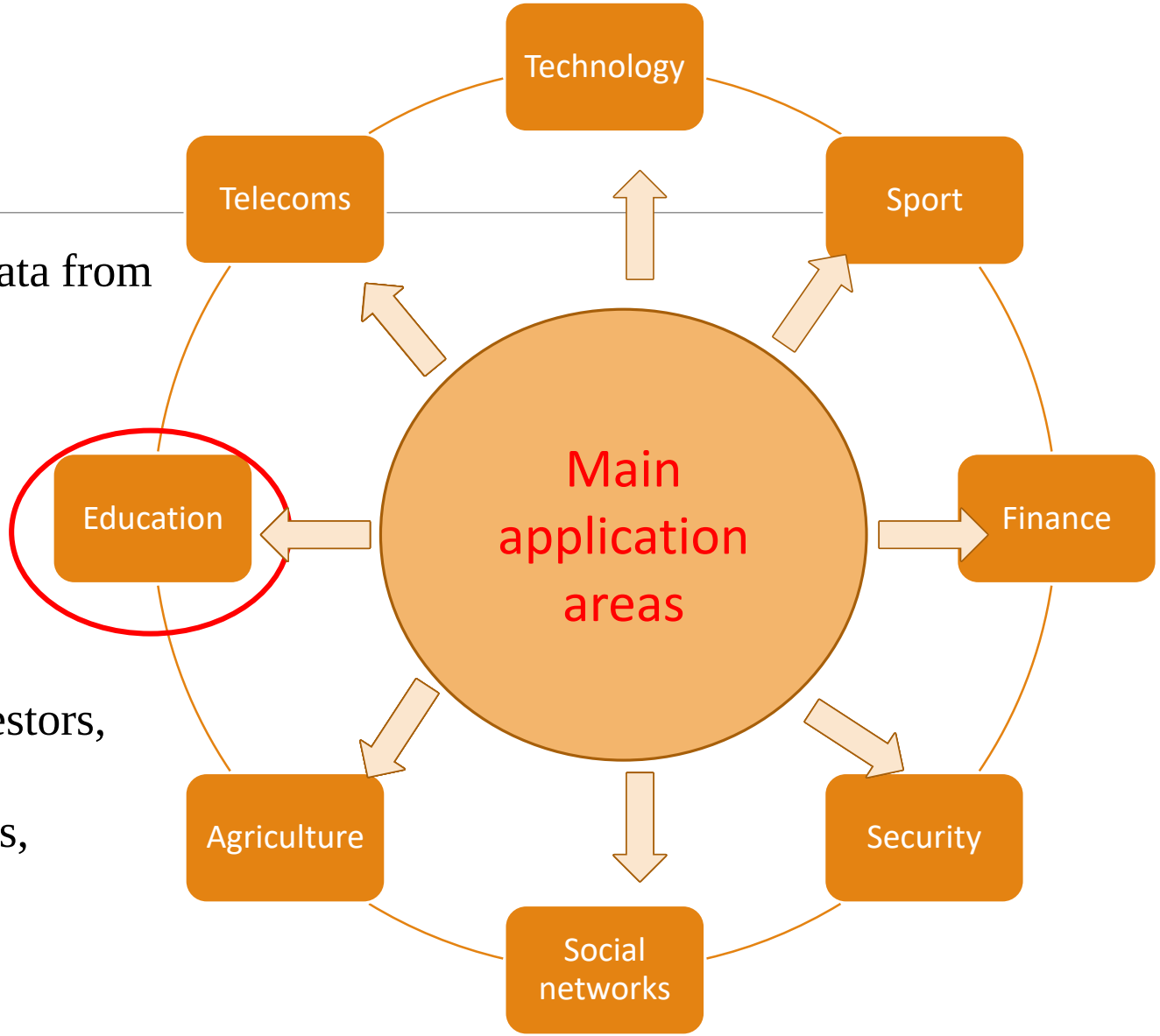


Introduction
context

- It results in a huge amount of hard-to-use data from a variety of fields.

- Researchers, the main stakeholders for whom these research results are intended, aspire to a deeper understanding of research results and better exploitation of them, with a view to improving the impact of research on businesses, governments and investors,

- indexation DB (DBLP, IEEE), blogs, books, Web pages, articles or tweets, etc…



Introduction
Motivation

- Natural language processing (NLP) is a branch of Artificial intelligence (AI) that helps computers understand, interpret and manipulate and respond to human in their natural language.

- The literature on AI is very extensive and still growing fast
  - It summarizes AI methods, tools, areas and applications from several scientific database repositories (DBLP, IEEE) and scientific search engines (ResearchGate).

- The literature search process consists in
  - seeking relevant information to find the answer to a specific question,
  - or to identify emerging or non-emerging application media of interest

Introduction
Motivation

- Tweets, smart fora, social networks : educating researchers or learners to perform acts that are appropriate or beneficial to the use of LLM (ChatGPT, Bard, concensus, Elicit.org, scite.ai, etc)

- ChatGPT redefines the future of academic research; but most academics don't know how to use it intelligently

- Mushtaq Bilal believes that incremental querying creates a better output. This involves providing indicators such as specific relevant concepts, authors, their ideas, source journals and so on...

Introduction
Problematic/objectives

Text classification of datasets for which few labelled examples are available, in a supervised learning

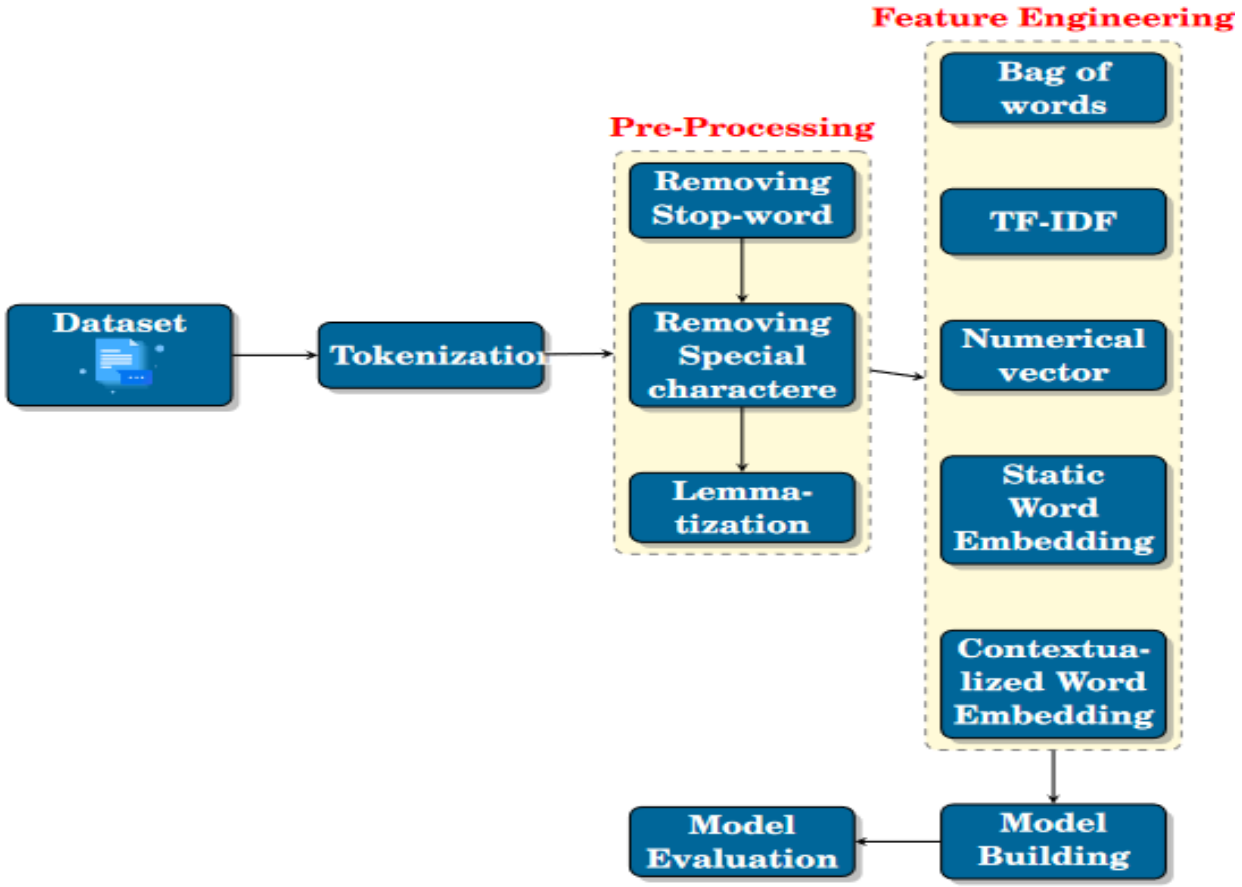
Introduction
Research question ?

How to represent input texts for classification algorithms,

avoiding domain-specific feature engineering steps ?

How to choose the most suitable learning algorithm ?

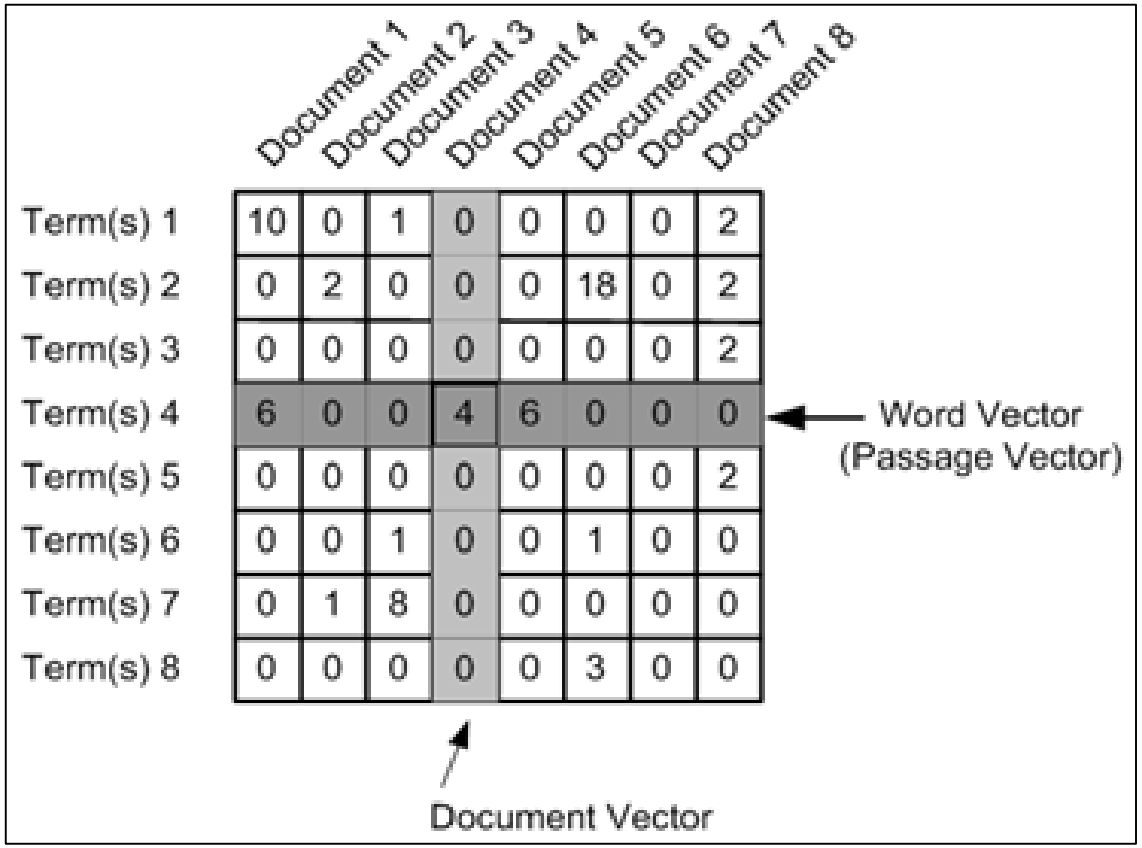
Document modelling
Automatic Text Analysis Process



Document modelling
Text structure

- Converting this mass of information into structured form is a major challenge, and the starting point for the development of data query and classification tools.

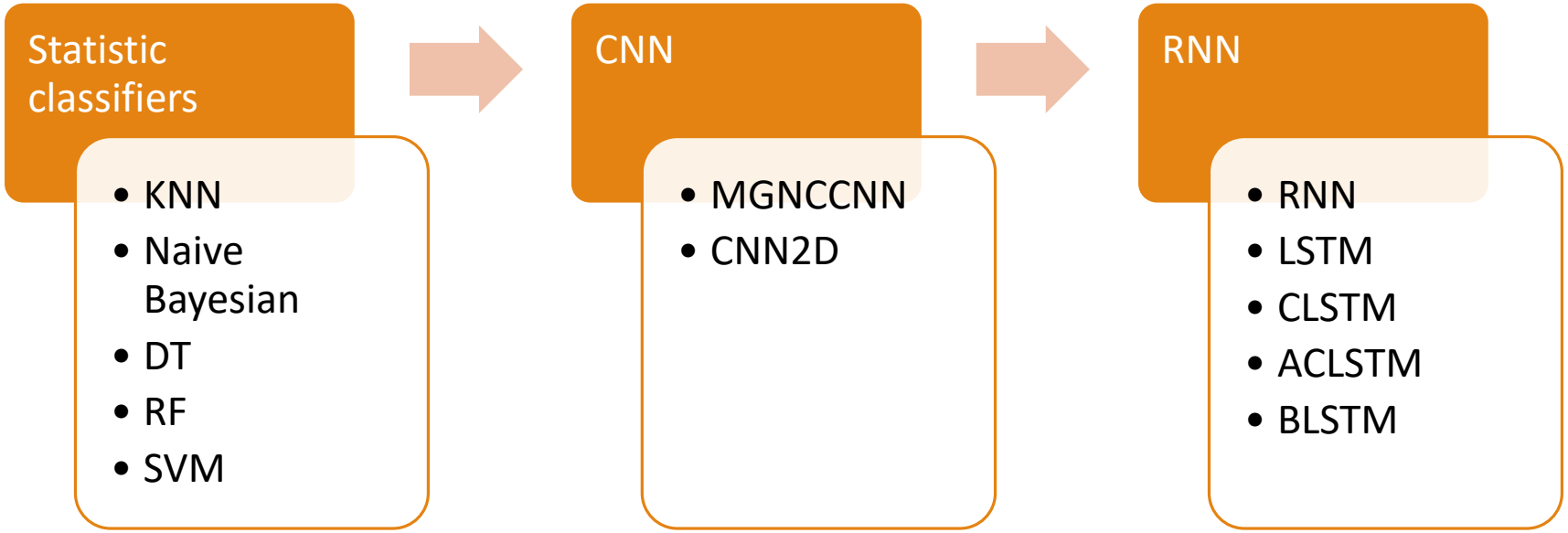
- This modelization has a strong impact on the accuracy and generalization of the learning system.



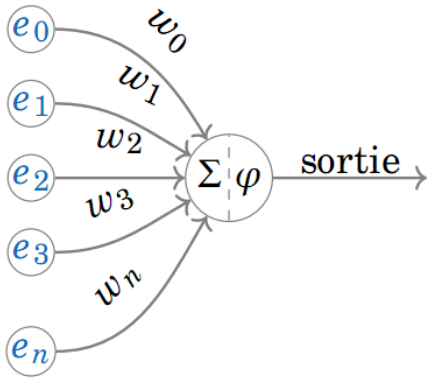
Document modeling
Text structure

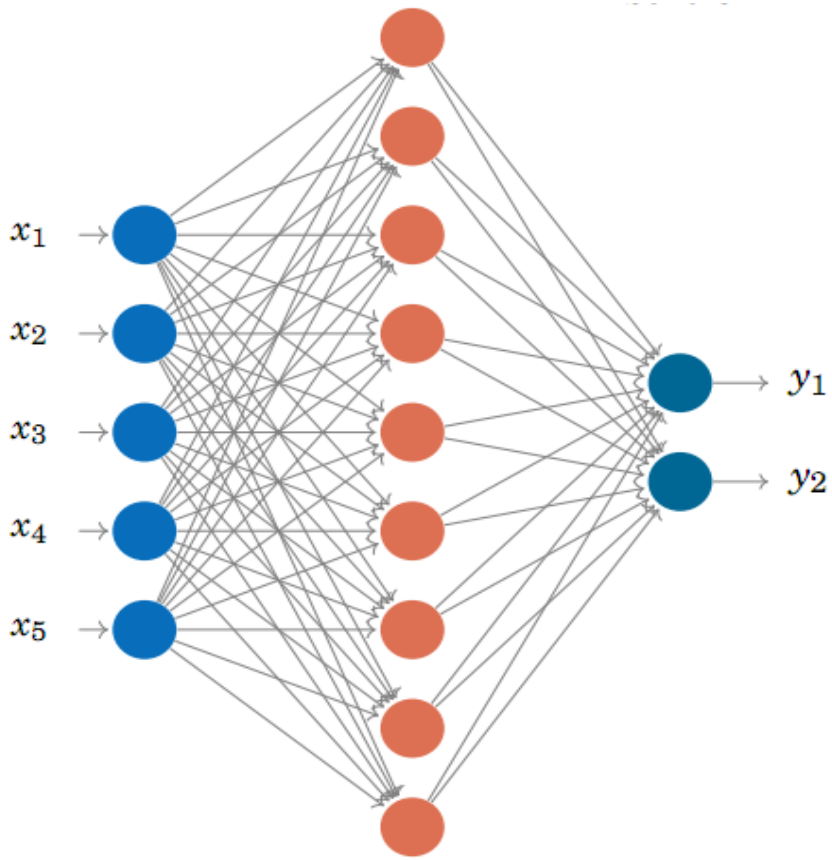
- « **Bag-of-words** »
  - Dictionary of corpura words
  - *d = {d1,..., dn}* being the number of occurrences of each word in the document *d*
- **TF-IDF**
  - measures the frequence of apparition of a word into a document
- **By ID**
  - Do not take into account the semantic and syntaxic relationship netween words
- **Word embedding**
  - Word2vec, glove, fastText

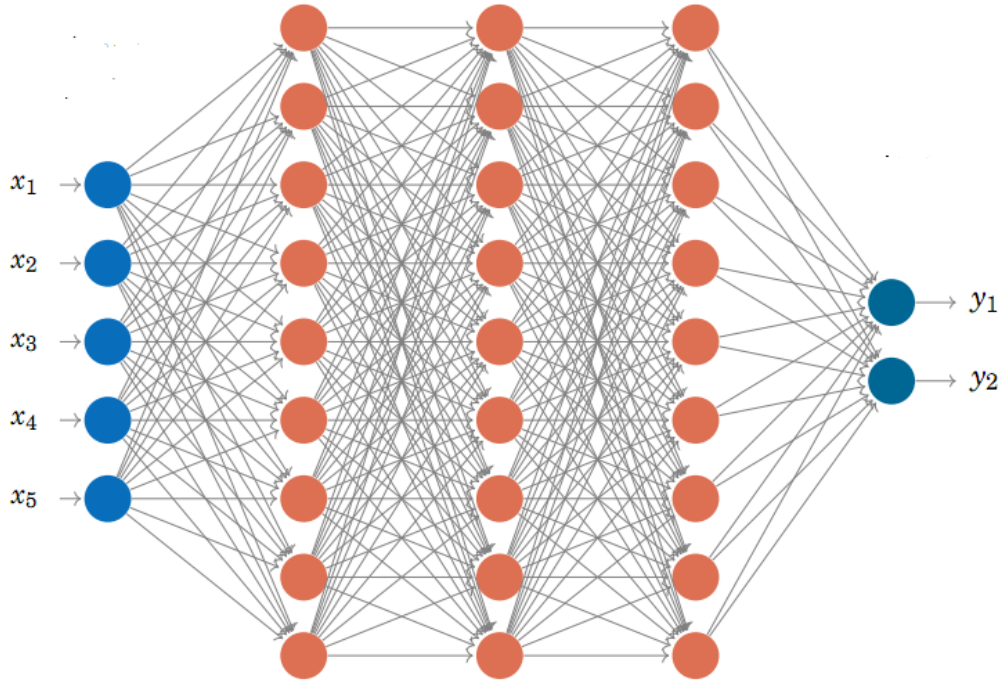
Classifyers architectures



Recurrent neural networks: LSTM
From RNN to LSTM

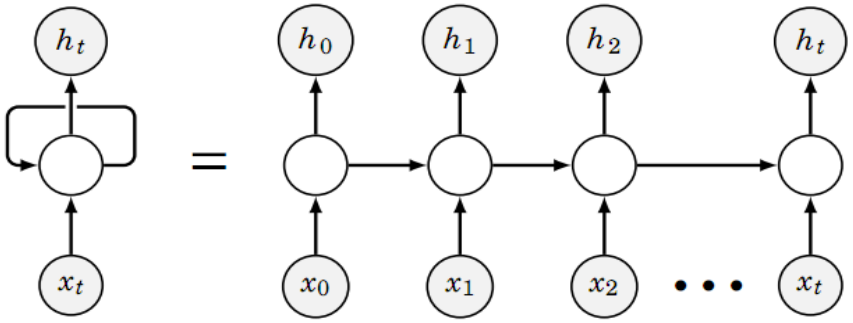




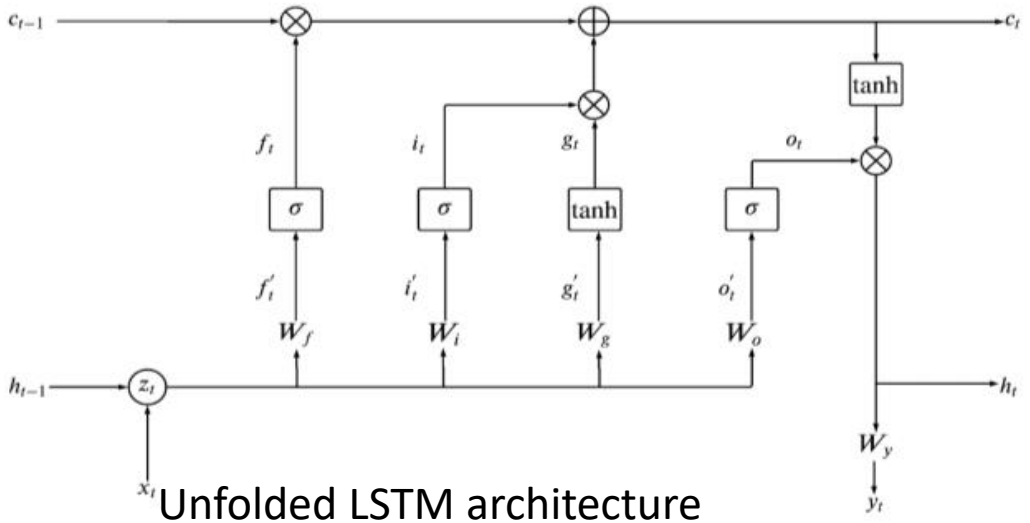


Recurrent neural networks: LSTM
From RNN to LSTM

- **RNN**:
- Information can be propagated backwards and forwards from the deepest to the most upstream layers,
- Dynamicity: weights depend not only on learned inputs, but also on previous outputs



Unfolded RNN



Unfolded LSTM architecture

Recurrent neural networks: LSTM
From RNN to LSTM

- The problem of LT dependencies of RNN
  - Difficult to capture LTD : Difficulty of diffusing the error gradient backpropagation through the feedback loop
  - No prediction of the word stored in long term memory
  - More accurate predictions from the recent informations
  - As the gap length increases, RNN does not give an efficient performance
- **LSTM**: RNN handling sequential data (times series, speech,text)
  - AI application Areas: Language translation, speech recognition, times series forecasting, anomaly detection, recommender systems, video analysis, etc,
  - Memory cell : 3 gates (input, forget, output gates)

Recurrent neural networks: LSTM
Advantages and disadvantages

- **Advantages of LSTM**
  - Long-term dependencies can be captured by LSTM; they have a memory cell that is capable of long term storage
  - In RNN, there is a problem of vanishing and exploding gradients when models are trained over long sequence; LSTM uses a gating mechanism that selectively recalls or gorgets information
- **Limitations of LSTM**
  - Computationnally more expensive: bad scalability for large-scale datasets
  - Cannot parallelize the work of processing the sentences, because of word by word process

Recurrent neural networks: LSTM
Different models

Long short-term memory (LSTM)	[Hochreiter and Schmidhuber, 1997]
Convolution Long short-term memory (CLSTM)	[Zhou et al., 2015]
Asymmetric Convolutional Bidirectional Long short-term memory (ACLSTM)	[Liang and Zhang, 2016]
Bidirectional Long short-term memory (BLSTM)	[Song et al., 2018]
Attention Bidirectional Long short-term memory (ABLSTM)	[Vaswani et al., 2017b]

Comparison methodology for different classifiers

- **1- Classifyer inputs / outputs**

  - For statistic based models : PreProcessing based on Word2Vec Vectorization
  - For LSTM Deep learning models: retain 15,000 most frequent(stop words excluded) words and we represent each document by a sequence of words to preserve their order
  - Embedding Input layer of dimension 300
  - Intermediare layers are described
  - The output layer contains as many neurons as there are classes in our dataset

Comparison methodology for different classifiers

- **1- Classifyer inputs / outputs**
- **2- Data partitioning and training**

  - The dataset is divided in four parts : 3 for training process and 1 for test step

  - Classic models: sklearn tool (https://scikit-learn/stable)

  - LSTM based models :
    - Batches: number of training instances to be considered simultaneously: 32
    - Each word is described by a vector of dimension 300
    - Epochs: number of iterations: 25

Comparison methodology for different classifiers

- **1- Classifyer inputs / outputs**
- **2- Data partitioning and training**
- **3- Performance metric : Accuracy, Recall, Precision, F-mesure**

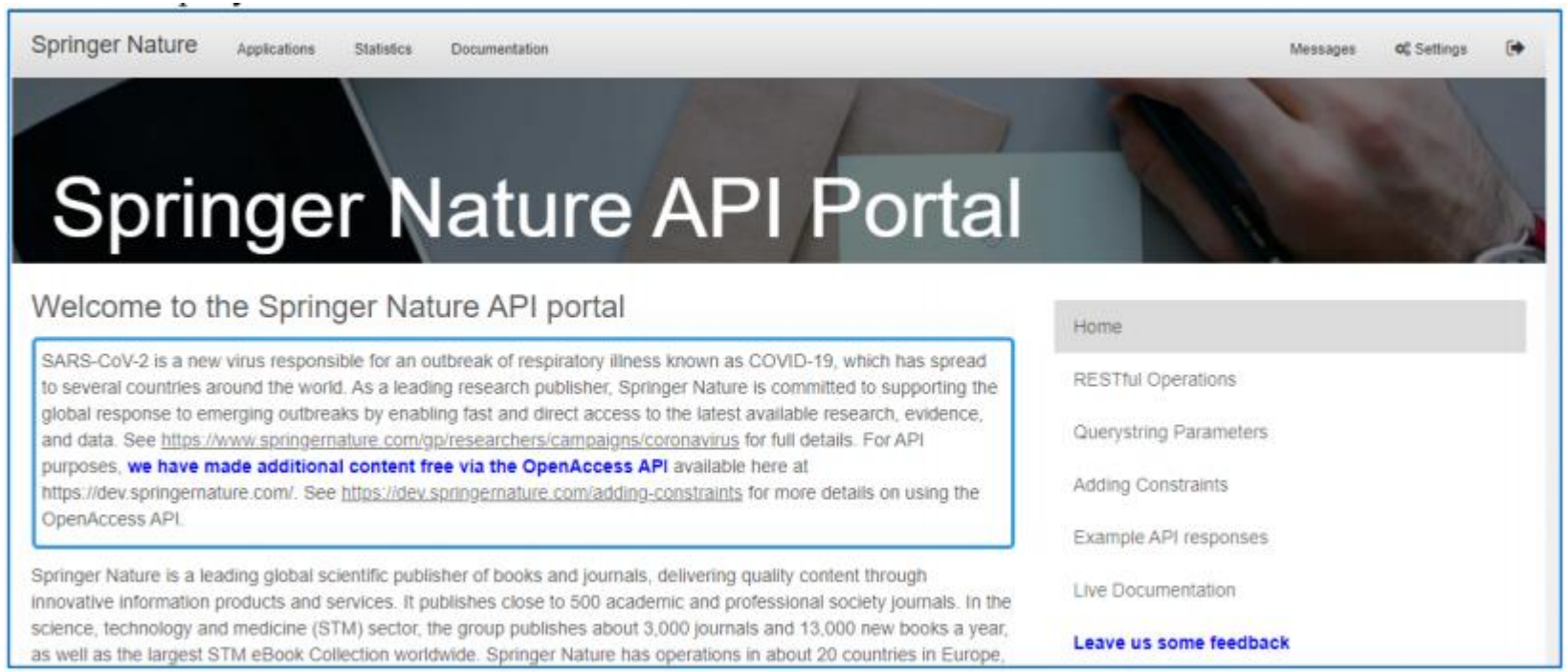
- Each output is a *n-dimension* vector corresponding to *n* classes to be predited

	In Class	Not in Class
Positive test	Positive True	Positive False
Negative Test	Negative False	Negative True

- *Accuracy*: $\frac{TP+TN}{TP+FP+FN+TN}$

Experiments

- **Springer Nature Metadata API**



Experiments: Accuracy

Caracteristics	Naive Bayesian	SVM	Rand-Forest	Decision Tree	KNN
Title	31.83	32.6	33.68	41.46	**28.05**
Full (Title+Abstract)	**58.25**	**67.18**	**70.34**	**75.05**	25.76
Keywords	52.70	56.43	67.68	70.01	16.42

Caracteristics	LSTM	CLSTM	ACLSTM	BLSTM	ABLSTM
Title	91.22	91.22	91.22	87.84	90.68
Full (Title+Abstract)	**93.42**	**93.42**	**95.29**	**92.65**	**92.33**
Abstract	92.51	92.51	92.51	92.45	91.89

Conclusion et perspectives

- Effectiveness of deep learning architectures for predicting categories on search results
- **Limits**: Only accuracy for performance classifyer validation
  - Black tools: poor interpretation of models
  - No explication, although good prediction [Shwartz-Ziv and Tishby, 2017]
- **Pespectives:**
  - Metrics: Recall, Precision, Specificity
  - How can we set up an active learning process combined with a deep learning model to select samples that will improve the model as speedily as possible?
  - Add a characteristic based on application support

Thank you for your kind attention