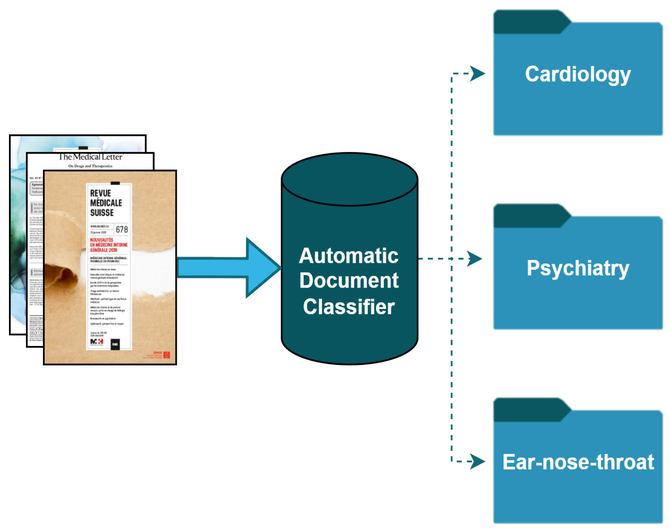
Performance of Machine Learning Methods to Classify French Medical Publications

Jamil Zaghir, Jean-Philippe Goldman, Mina Bjelogrlic, Daniel Keszthelyi, Christophe Gaudet-Blavignac, Hugues Turbé, Belinda Lokaj, Christian Lovis

INTRODUCTION

- Information overload remains a challenge many healthcare organizations face.
- Seeking precise information in a motley collection of data is resource-demanding.
- A way of helping information retrieval is to separate the dataset into multiple topics.
- **Automatic Document Classification (ADC)** systems can be used to assign labels.
- **Natural Language Processing (NLP)** methods coupled with Machine Learning techniques.
- A three-phase methodology to classify French medical free-text to their closest subject:
  i. Extracting data to get document – topic pairs
  ii. Running NLP pipeline to preprocess documents
  iii. Classifying using **Traditional Machine Learning (TML)** or **Deep Learning (DL)** models.
- First paper involving document classification for French medical articles.

Goal: automatic document classification of medical documents into clinical specialties



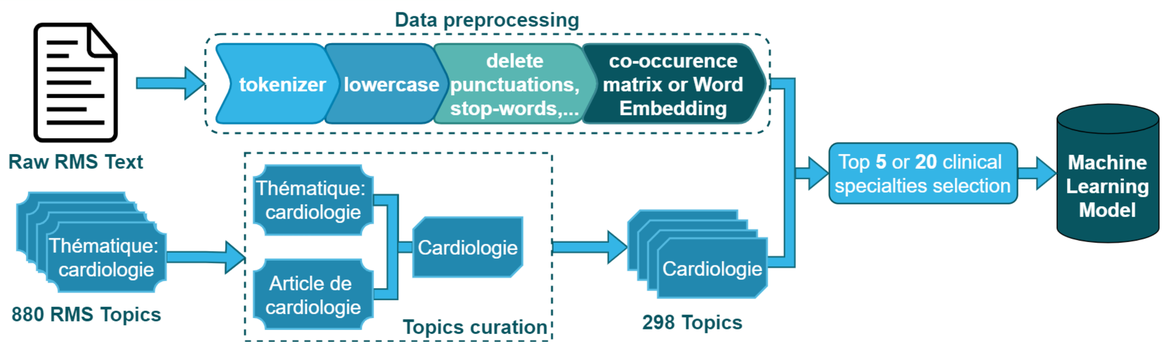
METHODS AND RESULTS

- Dataset: 13’000+ articles from **Revue Médicale Suisse**, a peer-reviewed French medical magazine (15 years of publication).

- Labeled dataset:
  i. Collect texts and topics from the RMS;
  ii. Curation of the topics (880 to 298);
  iii. Top-5 and top-20 of clinical specialties are kept.

- The pipeline starts with text preprocessing.
- DL classifiers use word embedding.
- CNN (4 layers, filters size 128).
- Training / Test split: 80% / 20%.

- **Results:** 5 and 20 multiclass classification with TML and DL models.
- FNN has the best results with an accuracy of 81% in top-20 classification.



Model	5-topic classification				20-topic classification			
	Precision	Recall	F1-Score	Accuracy	Precision	Recall	F1-Score	Accuracy
Support Vector Machine (RBF)	0.92	0.91	0.91	0.91	0.73	0.71	0.71	0.71
Support Vector Machine (Lin.)	0.93	0.93	0.93	0.93	0.76	0.76	0.76	0.76
Naïve Bayes	0.92	0.92	0.92	0.92	0.75	0.75	0.74	0.74
Logistic Regression	0.94	0.93	0.94	0.94	0.78	0.78	0.78	0.78
Mean Embeddings + FNN	**0.96**	**0.95**	**0.96**	**0.96**	**0.83**	**0.81**	**0.81**	**0.81**
Embeddings + 2D-CNN	0.92	0.91	0.91	0.92	0.74	0.71	0.70	0.70

DISCUSSION

- FNN performed the best, but Logistic Regression could be an interesting candidate: up to 10 times faster training time.

- CNN performed the worst: only 151 documents per class for top-5 classification and 111 documents for top-20.

- Higher amount of data might improve CNN performance as a similar work [1] had CNN outperforming Logistic Regression with 4000 samples per class.

CONCLUSIONS

- Comparison between TML and DL methods applied to medical document classification in French.
- TML methods are suitable: simple to build and low computational cost.

Next steps

- More state-of-the-art Deep Learning models (RNN-based, or more recently, Transformer-based).

- Transfer-learning with French BERT models (e.g. with CamemBERT [2]).

SIMED@HUG SIMED@UNIGE

1. M. Hughes, I. Li, S. Kotoulas, and T. Suzumura, Medical text classification using convolutional neural networks, in: Informatics for Health: Connected Citizen-Led Wellness and Population Health, IOS Press, 2017: pp. 246–250.
2. L. Martin, B. Muller, P.J.O. Suárez, Y. Dupont, L. Romary, É.V. de la Clergerie, D. Seddah, and B. Sagot, CamemBERT: a Tasty French Language Model, Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. (2020) 7203–7219. doi:10.18653/v1/2020.acl-main.645.

nccr evolving language NATIONAL CENTRE OF COMPETENCE IN RESEARCH
FONDATION LEENAARDS
UNIVERSITÉ DE GENÈVE FACULTÉ DE MÉDECINE
HUG Hôpitaux Universitaires Genève
SIMED