Challenges of Trustable AI and Added-Value on Health
B. Séroussi et al. (Eds.)

doi:10.3233/SHTI220613

Performance of Machine Learning Methods to Classify French Medical Publications

Jamil ZAGHIR[a,b,1], Jean-Philippe GOLDMAN[a,b], Mina BJELOGRLIC[a,b], Daniel KESZTHELYI[a, b], Christophe GAUDET-BLAVIGNAC[a, b], Hugues TURBÉ[a, b], Belinda LOKAJ[a, b] and Christian LOVIS[a,b]
[a] *Division of Medical Information Sciences, University Hospitals of Geneva,*
[b] *Department of Radiology and Medical Informatics, University of Geneva, Switzerland*

Abstract. Many medical narratives are read by care professionals in their preferred language. These documents can be produced by organizations, authorities or national publishers. However, they are often hardly findable using the usual query engines based on English such as PubMed. This work explores the possibility to automatically categorize medical documents in French following an automatic Natural Language Processing pipeline. The pipeline is used to compare the performance of 6 different machine learning and deep neural network approaches on a large dataset of peer-reviewed weekly published Swiss medical journal in French covering major topics in medicine over the last 15 years. An accuracy of 96% was achieved for 5-topic classification and 81% for 20-topic classification.

Keywords. Document classification, unstructured medical data, machine learning, deep learning, natural language processing, Revue Médicale Suisse, French

1. Introduction

With the increasing amount of available data, information overload remains a challenge many organizations face. This is also true for healthcare research, where a large number of articles are published in medical journals. Seeking precise information in this motley collection of data is resource-demanding. As these articles cover the broad discipline of medicine, a way of helping information retrieval is to separate the dataset into multiple topics. If the collection size of data is too large to consider manual annotation, Automatic Document Classification (ADC) systems can be used to automatically assign labels. This research presents an approach for ADC, applying Natural Language Processing (NLP) methods coupled with Machine Learning pattern-identification abilities. A three-phase methodology to classify French free-text medical articles to their closest subject (i.e. multiclass classification) is proposed: (i) extract data to get documents and their corresponding topics; (ii) run NLP pipeline to preprocess documents, making them suitable for ADC methods; and (iii) classify documents using several approaches, from Traditional Machine Learning (TML) to Deep Learning (DL) models. To the best of our knowledge, no literature on document classification was published for French medical articles, the present paper aiming at filling this gap.

[1] Corresponding Author, Jamil ZAGHIR, Geneva, Switzerland; E-mail: jamil.zaghir@unige.ch.

2. Methods

The dataset contains more than 13 000 articles from the Revue Médicale Suisse [1], a weekly peer-reviewed medical magazine published, for the last 15 years, in French. The labeled set of data is built in two steps: (i) gathering textual content and the corresponding topics; and (ii) curation of the topics to reduce them from 880 to 298. Articles whose topics belong to the top-k of clinical specialties are kept, with k being 5 and 20 (6% and 17% of the dataset). For fair comparison, the text preprocessing step is identical for each learning method. The pipeline consists in converting text to lowercase, deleting punctuations, stop-words, and words occurring less than twice. Among models, DL classifiers have an embedding layer. One model computes the mean of word embeddings, and is based on a Feed-forward Network (FNN) while another one is a Convolutional Neural Networks (CNN) (4 layers, filters size 128). The train/test split is 80%/20%.

3. Results

The models' results are shown in Table 1 for 5 and 20 multiclass classification using TML and DL. FNN has the best results with an accuracy of 81% in top-20 classification.

Table 1. Classification performance with the following abbreviations P: Precision, R: Recall, Acc: Accuracy

	5-classes				20-classes			
Model	**P**	**R**	**F1-score**	**Acc**	**P**	**R**	**F1-score**	**Acc**
SVM (RBF)	0.92	0.91	0.91	0.91	0.73	0.71	0.71	0.71
SVM (Linear)	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-Emb. + FNN	**0.96**	**0.95**	**0.96**	**0.96**	**0.83**	**0.81**	**0.81**	**0.81**
Emb. + 2D-CNN	0.92	0.91	0.91	0.92	0.74	0.71	0.70	0.70

4. Discussion

Even though FNN performed the best, Logistic Regression could be an interesting candidate as training is up to 10 times faster on this dataset for a performance of only 2 to 4% smaller accuracy. Since there are on average 151 documents per class for top-5 classification and 111 documents for top-20, CNN performed the worst. However, a higher amount of data might be needed, as in a similar work [2] showed CNN outperforming LR with 4000 samples per class. This research has been co-funded by "NCCR Evolving Language, Swiss National Science Foundation Agreement #51NF40_180888" and by "Leenaards Foundation".

References

[1] Revue Médicale Suisse - Revue médicale francophone de référence, (n.d.). https://www.revmed.ch/
[2] Hughes M, Li I, Kotoulas S, Suzumura T. Medical Text Classification Using Convolutional Neural Networks. Stud Health Technol Inform. 2017;235:246-250. PMID: 28423791.