Post-doc position: Multi-view Topic-Modeling for Medical Report Analysis

Job Location: Marseille, Aix-Marseille University (AMU)
Job quota: 100%
Date of recruitment: 01/09/2023
Nature of recruitment: Position open to holders of a doctorate, 1-year fixed-term contract.
Keywords: Clustering, Topic-Modeling, Multi-views, Textual data, Analysis of medical records, Thoracic oncology, Lung cancer, Diagnostic wandering

Working context:
Aix-Marseille University (AMU) is a multidisciplinary university structured around five major disciplinary sectors (Arts, Literature, Languages and Human Sciences; Law and Political Science; Economics and Management; Health; Science and Technology and a multidisciplinary sector). AMU trains 80,000 students in 18 components (faculties, schools, institutes) spread over four departments (04, 05, 13, and 84) and 10 cities (find out more: www.univ-amu.fr).

Context. With the ever-increasing amount of textual data in the biomedical field, unsupervised classification (Text Clustering), which aims to group a set of texts into classes, has become more critical than ever. However, to group similar texts, it is essential to grasp their meaning, and for this, many textual representations exist, from the simplest, such as Bag-Of-Words (BOW), to the most complex, such as recent transformers (Bert, Roberta, etc.).
Purpose. Our objective will be to develop multi-view probabilistic models (which can consider an unlimited number of textual representations), to compare their performance to several recent algorithms that we have developed, and beyond that, to develop a metric of homogeneity of handling—load on the one hand and a diagnostic wandering metric on the other hand.
Methodology. We will first conduct a retrospective study. The textual reports of the pneumology consultations of the hospital center n°16 (cardiovascular and thoracic) of the Assistance Publique - Hôpitaux de Marseille since 2010, as well as the PMSI, acts, and the CIM-10 pathologies (international classification of diseases), will be extracted from the medico-administrative database of the medical informatics department (DIM) of the AP-HM. All data will, of course, be anonymized at the source as part of the processing by the DIM. We will identify through the ICD-10 coding the patients who, at the end of their observed medical history, actually fell under thoracic oncology. This will include patients with a proven diagnosis of cancer of the pleura, lung parenchyma, oesophagus, thymus, etc. A probabilistic multi-view theme detection model will be developed. Indeed, several representations (multi-views) of the consultation reports will be extracted using NLP tools (BioBERT, Biomedical-Entity-Linking, Sytora, etc.) intended for biomedical data. These representations reflect, on the one hand, the semantic meaning contained in the reports by considering the context, and on the other hand, representations of the pathologies, symptoms, treatments, etc., appearing in the text of the reports. Finally, a probabilistic topic-modeling model will receive these multi-view representations of medical reports as input to unsupervised discover relevant classes of reports. A clustering of PMSI procedures and pathologies according to ICD-10 selected for each patient will make it possible to propose a joint proximity metric of the semantics of the reports, the procedures performed, and the diagnoses selected, opening the way to identification and characterization of the diagnostic error and uniform treatment. Finally, a metric reflecting the medical prognosis of the patients will be proposed (the death and the date of death of the patients are also accessible and available).

Responsible

Under the hierarchical authority of Project Manager Dr. Rafika Boutalbi, she will guarantee the proper integration of the agent. Also, she will supervise the actions carried out and the good progress of the program defined in the "Main activities and tasks" section.

Main activities and tasks

Tasks:

- Literature review of work carried out previously for the analysis of medical reports.
- Data processing: Cleaning and extracting the different useful representations for the model.
- Modeling a multi-view approach for clustering medical reports.
- Implementation of the proposed approach.

Main activities:

- Development of the theoretical part of the proposed approach modeling.
- Implementation of the proposed approach.
- Evaluation of the implemented approach using existing evaluation measures and new ones to be defined.

Work conditions:

The selected candidate will be assigned to the QARMA team of the LIS, located on the Châteaux Gambert campus.
The candidate will be required to regularly travel to the Aix-Marseille site or carry out missions as part of his/her activity.
Possibility of working remotely (2 days/week) after 6 months of the contract.

REQUIRED SKILLS:

1. Main skills:

- In-depth knowledge of word processing and familiarity with different text representations, such as Transformers (BERT, RoBERTa, etc).
- In-depth knowledge of Machine Learning, in particular, Topic Modeling and clustering.
- High-level publications in international conferences and journals (SIGIR, NIPS, ACL, KDD, etc.)
- Knowledge of named entity detection and disambiguation.
- Have been confronted with or have worked on health data.

2. Technical skills:

- Have an excellent command of the Python programming language.
- Familiar with health issues.

3. Soft skills

- Scientific curiosity and involvement in the project
- Sociable and open to collaboration with researchers from other teams, particularly in health and NLP.
- High-level publications in international conferences and journals.

This post-doc can be extended if the results obtained are conclusive and make it possible to achieve the objectives set.

RESEARCHER PROFILE

Education: PhD level in Data Science, NLP, Machine Learning
Desired experience: 2 years of experience will be a plus.

The application file (CV, cover letter, and, where possible, your last professional interview), should be sent electronically to: rafika.boutalbi@univ-amu.fr, stephane.delliaux@univ-amu .Fr