Contribution ID: **92** Type: **Oral Presentation**

Text Classification on COVID-19: a Transformer-based Approach

Wednesday, 27 March 2024 17:00 (30 minutes)

During the COVID-19 pandemic, there has been a rapid growth of literature and the need to access useful information quickly to understand its disease mechanisms, define prevention techniques, and propose treatments, therefore text classification has become an essential and challenging activity. LitCovid represents an example of the extent to which a COVID-19 literature database can grow: it has accumulated over 390,000 articles with millions of accesses since 2020. Approximately 10,000 new articles have been added to LitCovid every month since May 2020.

Text classification is the process of assigning predefined labels to text by its content. Label selection is a demanding task due to various factors. Firstly, a deep knowledge of the topic domain is compulsory. Secondly, a label structure has to be established since a label may be connected or imply other labels, i.e. diagnosis implies tomographic, medical imaging, and radiation. Finally, the depth of this label structure has to be defined and be consistent for every topic, i.e. treatment and prevention must be described using the same level of detail. For text classification, we have considered transformer models that have achieved unprecedented breakthroughs in the field of Natural Language Processing. The core function that drives the success is the attention mechanism, which provides the ability to dynamically focus on different parts of the input sequence when producing the predictions and capturing the relationships between words in the sentence.

In this study, we have proposed labels that can satisfy the need to assist title and abstract screening for supporting COVID-19 research on detection, diagnosis, treatment, and prediction. Our labels extend what LitCovid and CORONA Central corpora provide for COVID-19 literature. Furthermore, we have classified literature by using different pre-trained transformer models, mainly based on BERT and ELECTRA models, such as BioBERT, PubMedELECTRA, and BioFormer. The selected papers have been identified through the usage of PRISMA, the Preferred Reporting Items for Systematic Reviews and Meta-Analysis, an incident approach originally developed in the chemical process industry and afterward well-distinguished in the medical field.

All the models have been compared by considering micro average, weighted average, and sample average methods for performance metrics. During this study, we have tackled the problem of the computational requirements, e.g. 10 hours per 5 epochs (2 hours per epoch) with a GPU P100 for PubMedBERT-large; and of the domain specificity of the model performances.

Primary authors: Ms TODESCHINI, Sofia Camilla (INFN CNAF); Dr CANAPARO, Marco (INFN CNAF); Prof. RONCHIERI, Elisabetta (INFN CNAF)

Presenters: Dr CANAPARO, Marco (INFN CNAF); Prof. RONCHIERI, Elisabetta (INFN CNAF)

Session Classification: Health & Life Science Applications

Track Classification: Track 2: Health & Life Sciences (including Pandemic Preparedness Applications)