

Spark NLP: A Versatile Solution for Structuring Data from Endoscopy Reports

Andrei Constantin IOANOVICI, Ștefan Marius MĂRUȘTERI*, Andrei Marian FEIER*, and Alina Dia TRAMBITAS-MIRON

”George Emil Palade” University of Medicine, Pharmacy, Sciences and Technology, Faculty of Medicine, 540142 Târgu Mureș, Romania

E-mails: andrei.ioanovici@umfst.ro; marius.marusteri@umfst.ro; andreifeier@gmx.com; dia.trambitas-miron@umfst.ro

* Author to whom correspondence should be addressed

Abstract

Artificial intelligence (AI) can be applied in the practice of gastroenterology to acquire and analyze information. Besides speed and duplicability, AI has the potential of also offering insight with results that surpass medical specialists. Natural language processing (NLP) is being used to extract information from text, organize and categorize documents. Processing unstructured data with NLP will result in structured data and medical codes can be extracted more easily (ICD10, medical procedure codes, etc) for reimbursement purposes among others. Recent research is studying the use of AI for automated interpretation of text from endoscopy and medical documents for better quality and patient phenotyping as well as enhanced detection and descriptions of endoscopic lesions such as colon polyps. In this paper, we present a method of extracting medical data using Spark NLP (John Snow Labs, DE, USA), by annotating endoscopy reports and training a model to automatically extract labels in order to obtain structured medical data. This can be used in combination with other forms of structured data for an optimal and novel patient profiling.

Keywords: Health record digitization; Spark NLP; Gastroenterology; Structured data