Digital Health and Informatics Innovations for Sustainable Health Care Systems
J. Mantas et al. (Eds.)

doi:10.3233/SHTI240559

A Comprehensive Natural Language Processing Pipeline for the Chronic Lupus Disease

Livia LILLI[a,b], Silvia Laura BOSELLO[a], Laura ANTENUCCI [a,b,1], Stefano PATARNELLO[a], Augusta ORTOLAN[a], Jacopo LENKOWICZ[a], Marco GORINI[c], Gabriella CASTELLINO[c], Alfredo CESARIO[a], Maria Antonietta D'AGOSTINO[a] and Carlotta MASCIOCCHI[a]
[a]*Fondazione Policlinico Universitario Agostino Gemelli IRCCS, Rome, Italy*
[b]*Catholic University of the Sacred Heart, Rome, Italy*
[c]*AstraZeneca Italy, MIND, Milan, Italy*
ORCiD ID: Livia Lilli https://orcid.org/my-orcid?orcid=0009-0005-3319-7211 Silvia Laura Bosello https://orcid.org/0000-0002-4837-447X, Stefano Patarnello https://orcid.org/0009-0008-2765-5935, Jacopo Lenkowicz https://orcid.org/0000-0002-8366-1474, Marco Gorini https://orcid.org/0009-0008-1455-3884, Alfredo Cesario https://orcid.org/0000-0003-4687-0709, Maria Antonietta D'Agostino https://orcid.org/0000-0002-5347-0060, Carlotta Masciocchi https://orcid.org/0000-0001-6415-7267

Abstract. Electronic Health Records (EHRs) contain a wealth of unstructured patient data, making it challenging for physicians to do informed decisions. In this paper, we introduce a Natural Language Processing (NLP) approach for the extraction of therapies, diagnosis, and symptoms from ambulatory EHRs of patients with chronic Lupus disease. We aim to demonstrate the effort of a comprehensive pipeline where a rule-based system is combined with text segmentation, transformer-based topic analysis and clinical ontology, in order to enhance text preprocessing and automate rules' identification. Our approach is applied on a sub-cohort of 56 patients, with a total of 750 EHRs written in Italian language, achieving an Accuracy and an F-score over 97% and 90% respectively, in the three extracted domains. This work has the potential to be integrated with EHR systems to automate information extraction, minimizing the human intervention, and providing personalized digital solutions in the chronic Lupus disease domain.

Keywords. Natural Language Processing (NLP), Information Extraction (IE), Artificial Intelligence (AI), Electronic Health Record (EHR), Systemic Lupus Erythematosus (SLE).

1. Introduction

The high concentration of unstructured data coming from Electronic Health Records (EHRs) led to the development of Natural Language Processing (NLP) systems to extract information from free text [1]. NLP is a subfield of Artificial Intelligence (AI) used for

[1] Corresponding Author: Laura Antenucci; E-mail: laura.antenucci@policlinicogemelli.it.

the automatic elaboration of human language, making it a valuable tool in healthcare, especially for tasks such as Information Extraction (IE) and text classification [2-4]. For these purposes, pattern-matching is a widely implemented technique in IE applications, and it identifies desired sequences of words within large texts, through the development of a rule-based system. This approach is largely used in the clinical field, for extracting specific entities from clinical reports [5,6]. However, rule-based systems have limited adaptability in terms of data sources (same rules can be not valid on diverse report types) and these approaches usually require a significant degree of human involvement in rules' definition and formulation. This paper aims to provide an NLP pipeline for the extraction of unstructured information from Italian-language EHRs. We will demonstrate the benefits of our approach, where a pattern-matching system is integrated with transformer-based approaches and clinical support, in order to detect information like therapies, diagnosis, and symptoms from EHRs of patients affected by Systemic Lupus Erythematosus (SLE). The overall framework is shown in Figure 1.

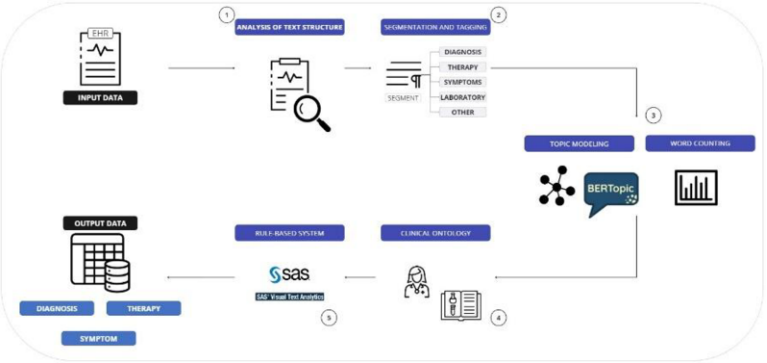

Figure 1. Natural Language Processing Pipeline for the extraction of Lupus-related features from Italian EHRs.

2. Methods

Analysis of Text Structure. We start with a first explorative analysis for a better understanding of EHRs structure. Then, our clinical reports typically cover five principal conceptual areas: diagnosis, treatments, symptoms, laboratory analysis and clinical follow-up.

Segmentation and Tagging. According to the previous analysis, segmentation and tagging constitute the second step of our NLP pipeline. In this stage, input text is divided in shorter segments, each one tagged with one or more of the targeted concepts, which are therapy, diagnosis and symptom.So, the extraction of specific information is limited to segments assigned to the relevant conceptual area, with a consequent noise reduction. The tagging was performed through personalized regex rules, indicated by physicians.

Topic Modeling and Word Counting. We used a transformer-based topic model to find out the most appropriate patterns usable for the development of rules in the pattern matching algorithm. To this purpose, we chose BerTopic [7], a topic model which uses BERT embeddings to represent document semantics, applying clustering techniques to group similar documents. We also implemented a word counting algorithm as further explorative analysis in order to get the most frequent words overall the reports, by excluding an input list of stop words.

Table 1. NLP performance for the overall diagnosis, therapy and symptom domains.

Overall Category	Accuracy	F1	AUC	Precision	Recall
Diagnosis					
Articular	1.0	1.0	1.0	1.0	1.0
Cutaneous	0.99	0.99	0.99	1.0	0.98
Hematologic	1.0	1.0	1.0	1.0	1.0
Neurologic	1.0	0.99	0.99	1.0	0.98
Renal	1.0	1.0	1.0	1.0	1.0
Serositis	1.0	1.0	1.0	0.99	1.0
Systemic	0.99	0.95	0.99	0.93	0.98
Vascular	1.0	0.99	0.99	0.99	0.98
Therapy					
Antimalarials	0.93	0.95	0.92	0.94	0.95
Glucocorticoids	0.98	0.98	0.98	0.99	0.97
Biological Immunosuppressant	1.0	0.99	1.0	0.98	1.0
Conventional Immunosuppressant	0.97	0.97	0.97	0.96	0.99
Symptom					
Oral aphthae	0.99	0.87	0.91	0.91	0.83
Alopecia	0.99	0.87	0.91	0.92	0.83
Anasarca	1.0	N/A	N/A	N/A	N/A
Hemolytic Anemia	1.0	N/A	N/A	N/A	N/A
Angioedema	1.0	N/A	N/A	N/A	N/A
Arthritis	0.99	0.92	0.94	0.95	0.88
Pulmonary Embolism	1.0	0.0	N/A	0.0	N/A
Epilepsy	1.0	N/A	N/A	N/A	N/A
Erythema	0.97	0.91	0.96	0.88	0.94
Raynaud's Phenomenon	0.99	0.95	0.97	0.95	0.95
Fever	0.99	0.9	0.96	0.88	0.92
Gangrene	1.0	N/A	N/A	N/A	N/A
Stroke or transient ischemic attack	1.0	N/A	N/A	N/A	N/A
Papula	1.0	0.97	0.97	1.0	0.94
Pericarditis	1.0	N/A	N/A	N/A	N/A
Pleuritis	1.0	0.0	0.5	N/A	0.0
Psychosis	1.0	1.0	1.0	1.0	1.0
Joint swelling	0.97	0.88	0.97	0.79	0.98
Deep vein thrombosis	1.0	1.0	1.0	1.0	1.0
Ulceration	1.0	0.96	1.0	0.92	1.0
Ascites	1.0	N/A	N/A	N/A	N/A

Clinical Ontology. Starting from the word counting results, we provided a team of expert physicians with a set of the most frequently occurring terms. Based on these findings, they built up a specialized clinical ontology, meaning a set of relations among relevant concepts and corresponding terms in the SLE clinical practice.

Rule-Based System. As the last stage of our pipeline, we created an advanced pattern-matching system for the extraction of all the terms included in the clinical ontology. The developed system matches patterns and rules defined according to the earlier topic analysis. The extraction process was performed using the SAS® Visual Text Analytics 8.3 software [8].

3. Results

The NLP pipeline was applied on a cohort of 262 Lupus patients, each one having a median of 15 ambulatory contacts, in a temporal window of 7 years on average, for a total number of 4567 ambulatory EHRs to analyze. Our system aimed to extract several concepts related to three areas of interest from unstructured data: diagnosis, therapy, and symptom. For this purpose, 70% of patients were chosen as training set and was passed to the entire pipeline for the textual analysis and the rule-based system development. A portion of approximately 10% of patients was allocated for validation, to refine the rules, based on a first quality assessment conducted by the data science team. Finally, about the remaining 20% of patients (i.e. 56 patients with a total of 750 EHRs) was annotated by physicians and used as the gold standard for performance evaluation of the rule-based system. Specifically, Table 1 shows Accuracy, F-score, Auc, Precision and Recall at singular category levels, for each domain. We leave metric values as not applicable (N/A) in cases of extremely unbalanced categories where we did not receive adequate frequency for the computation of metrics. Additionally, we also compute NLP performance at the overall domain level for a better comparison of the three areas of interest (results are shown in Table 2).

4. Discussion

The study describes the development of an NLP pipeline for the automated extraction of Lupus information from ambulatory EHRs. Looking at NLP performances, Table 1 shows results of our system, with respect to the gold standard. For highly unbalanced symptoms (like Anasarca, Hemolytic Anemia, Angioedema, Epilepsy, Gangrene, Ischemic attack, Pericarditis, Ascites, Pulmonary Embolism and Pleuritis), where only positive or negative samples are present, it wasn't possible the computation of certain metrics: thus we prompted a focus just on the accuracy. In the other cases, we also opted for F-score together with Precision, Recall and AUC, which consider the different distributions among positives and negatives, during evaluation. Specifically, psychosis and deep vein thrombosis are the top performing balanced symptoms, showing an F-Score of 1. For the other domains, articular, hematologic and renal diagnosis with a F-Score and Accuracy of 1, and biological immunosuppressant with a F-Score of 0.99, stand out as the top performing categories. In general, diagnosis and therapy present relevant results, with an overall F-Score equal to 0.99 and 0.97, respectively, as shown in Table 2. On the other hand, symptoms exhibit a slightly lower overall F-score, reflecting an intricate nature of the category identification for that domain.

Table 2. NLP performance for the overall diagnosis, therapy and symptom domains.

Overall Category	Accuracy	F1	AUC	Precision	Recall
Diagnosis	1.0	0.99	0.99	1.0	0.99
Therapy	0.97	0.97	0.97	0.97	0.97
Symptom	0.99	0.90	0.96	0.89	0.92

5. Conclusions

This paper aims to implement a NLP approach to extract useful information from Italian EHRs of patients affected by SLE. Our findings show that NLP is a valuable solution for streamlining the process of IEfrom EHRs. By automating multiple processes, such as text segmentation, pattern recognition, andIE, our system reduces the need for human intervention.. This study intends to empower healthcare professionals with valuable insights that will facilitate better-informed decisions, optimize treatments, and ultimately enhance the care provided to patients affected by SLE.

Acknowledgements

The project has been developed with the financial contribution of AstraZeneca and it has been approved by the Institutional Review Board of the Fondazione Policlinico Universitario Agostino Gemelli IRCCS (protocol number 0034012/22). This study received partial funding from Italian Ministry for University and Research (MUR) under the Program PON "Research and Innovation" supporting the development of artificial intelligence platform Gemelli Generator at Policlinico Universitario A. Gemelli IRCCS".

References

[1] Sheikhalishahi S, Miotto R, Dudley JT, Lavelli A, Rinaldi F, Osmani V, et al. Natural language processing of clinical notes on chronic diseases: systematic review. JMIR medical informatics. 2019;7(2):e12239.
[2] Adamson BJ, Waskom M, Blarre A, Kelly J, Krismer K, Nemeth S, et al. Approach to Machine Learning for Extraction of Real-World Data Variables from Electronic Health Records. medRxiv. 2023:2023-03.
[3] BIR: Biomedical Information Retrieval System for Cancer Treatment in Electronic Health Record Using Transformers. Sensors. 2023;23(23):9355.
[4] Romanowski B, Ben Abacha A, Fan Y. Extracting social determinants of health from clinical note text with classification and sequence-to-sequence approaches. Journal of the American Medical Informatics Association. 2023:ocad071.
[5] Deshmukh PR, Phalnikar R. Anatomic stage extraction from medical reports of breast cancer patients using natural language processing. Health and Technology. 2020;10(6):1555-70.
[6] Topaz M, Murga L, Gaddis KM, McDonald MV, Bar-Bachar O, Goldberg Y, et al. Mining fall-related information in clinical notes: Comparison of rule-based and novel word embedding-based machine learning approaches. Journal of biomedical informatics. 2019;90:103103.
[7] Grootendorst M. BERTopic: Neural topic modeling with a class-based TF-IDF procedure. arXiv preprint arXiv:220305794. 2022.
[8] Anandarajan M, Hill C, Nolan T, Anandarajan M, Hill C, Nolan T. SAS Visual Text Analytics. Practical Text Analytics: Maximizing the Value of Text Data. 2019:263-82.