Title of the Thesis:

“Natural Language Processing For Clinical Documents: Addressing Label Scarcity To Support Healthcare Policies And Clinical Practice”

Abstract:

Clinical documents play a crucial role in modern healthcare, containing rich yet unstructured information about patient history, diagnoses, and treatments that are only partially captured by structured databases. In recent years, Natural Language Processing (NLP) has evolved rapidly, achieving remarkable results in various tasks. However, clinical NLP faces significant challenges, including the scarcity of labelled data, privacy constraints, the complexity of medical language, and the heterogeneity of medical documents. These limitations have historically slowed the adoption of NLP in healthcare, particularly in languages other than English.

This thesis addresses these challenges, particularly the scarcity of labelled data, by proposing a wide spectrum of techniques, including unsupervised, weakly supervised, and data augmentation approaches for fully supervised learning. The algorithms and models developed span the entire range of NLP techniques, from rule-based methods to the most recent large language models, including a dedicated study on model interpretability. The main contribution of this thesis is the adaptation of all these diverse techniques to the clinical NLP setting, demonstrating how they can be effectively used to face the challenges posed by this domain.

The studies presented in the thesis fall within the contexts of the Italian and Dutch healthcare systems, where research on clinical NLP has been particularly limited. Different types of textual data are analyzed, including referrals, discharge summaries, and progress notes, across various medical specialties such as cardiology, pediatrics, angiology, and gastroenterology. Each study addresses a real-world problem, working within computational and data constraints and proposing tailored methods to overcome them. The first two studies focus on public health applications, developed in collaboration with health authorities in the Lombardy Region, while the remaining studies tackle challenges in epidemiology and clinical practice, supporting medical researchers and clinicians across different institutions.

By demonstrating how a broad range of NLP techniques can be applied effectively in clinical settings, overcoming the previously mentioned barriers, this thesis lays the groundwork for future advancements in healthcare text analysis. The findings highlight the potential of NLP to optimize healthcare administration, support clinicians, and facilitate epidemiological research. Continued progress in this field will require further interdisciplinary collaboration aimed at the development of multilingual, privacy-conscious models with limited computational requirements, that can be seamlessly integrated into real-world healthcare systems.