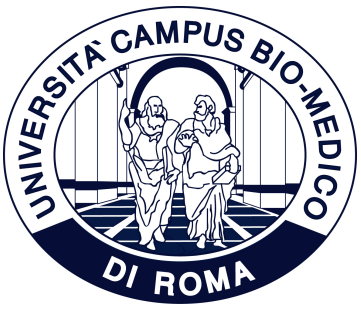

Università Campus Bio-Medico di Roma

Corso di Dottorato di Ricerca in
Scienze e Ingegneria per l'Uomo e l'Ambiente
XXXV ciclo a.a. 2019-2020

Exploring New Technologies in Healthcare: Advancing Natural Language Processing

Luca Bacco

Coordinatore
Prof. Giulio Iannello

Tutori
Dott. Felice Dell'Orletta
Dott. Ing. Mario Merone

13 Marzo 2023

Contents

List of Figures

List of Tables

Acronyms

AI Artificial Intelligence

AIn Averaged Infinity-norm

ATS Automatic Text Summarization

axSpA axial SpondyloArthritis

BART Bidirectional and Auto-Regressive Transformers

BERT Bidirectional Encoder Representations from Transformers

BoW Bag of Words

CAD Computer-Aided Diagnosis

CHV Consumer Health Vocabulary

CI Clinical Informatics

CS Computer Science

CSF CerebroSpinal Fluid

CSTS Clinical Semantic Textual Similarity

DH Distributional Hypothesis

DL Deep Learning

DS(M) Distributional Semantics (Model)

DSS Decision Support System

EHR Electronic Health Record

ExHiT Explainable Hierarchical Transformer

EST Expertise Style Transfer

FPT Further Pre-Training

FT Fine-Tuning

GPU Graphics Processing Unit

HIPAA Health Insurance Portability and Accountability Act

i2b2 Integrating Biology and the Bedside

ICD International Classification of Diseases

ID Incidental Durotomy

LASSO Least Absolute Shrinkage and Selection Operator

LBP Low Back Pain

LDA Latent Dirichlet Allocation

LM Language Modeling

LR Logistic Regression

MAE Mean Absolute Error

MAP Multimodal Automated Phenotyping

MLM Masked Language Modeling

MDS Multi-Document Summarization

mHealth mobile Health

ML Machine Learning

MNR Multiple Negatives Ranking

MQP Medical Question Pairs

MRI Magnetic Resonance Imaging

MT Machine Translation

MTS Manual Text Summarization

NLF Neural Foraminal Stenosis

NLG Natural Language Generation

NLP Natural Language Processing

NLU Natural Language Understanding

OhE One-hot Encodings

PHI Private Health Information

PICO Population/Problem, Intervention, Comparison, Outcome

PLM Pre-trained Language Model

PLS Plain Language Summarization

POS Part-Of-Speech

PRISMA Preferred Reporting Items for Systematic reviews and Meta-Analyses

QUADAS-2 Quality Assessment of Diagnostic Accuracy Studies

RNN Recurrent Neural Network

RSA Representational Similarity Analysis

SA Sentiment Analysis

SCC Sentence Classification Combiner

SCS Spinal Canal Stenosis

SDS Single Document Summarization

SimCSE Similarity Contrastive Sentence Embedding

SM Sentence Mask(ing)

SPE (sinusoidal) Sentence Positional Embeddings

STS Semantic Textual Similarity

SVM Support Vector Machine

TF-IDF Term-Frequency-Inverse Document Frequency

TST Text Style Transfer

UMLS Unified Medical Language System

VI Vascular Injury

VTE Venous ThromboEmbolism

WHO World Health Organization

WWW World Wide Web

XAI eXplainable Artificial Intelligence

XGBoost eXtreme Gradient Boosting

Abstract

In the last few years, Natural Language Processing (NLP) has gained impressive momentum in both academic and industrial research. Texts portray distinct characteristics from other kinds of data (such as images, audio, etc.), being inherently discrete, compositional, and hierarchical. NLP techniques allow the manipulation of such a peculiar source of information, providing researchers and practitioners with a way to automatize the analysis of textual data, enabling humans to communicate with machines (and vice versa) through natural language.

Understanding and generating natural language revealed to be a valuable ally in many fields, including healthcare. The process of digitalization taking place nowadays in healthcare, as well as in everyday life (e.g., social media), is pushing the need for tools to manage the high volumes of textual data available today. Leveraging unstructured, textual data from Electronic Health Records (EHRs) and Internet resources has disclosed plenty of applications, paving the way for improvements in the care of patients and their diseases.

The exploitation of NLP techniques in the healthcare domain is not only driven by the digitalization process we are living in but also by the advancements in the NLP field over the past years. In the last decade, in particular, we shifted the NLP paradigm from classical, machine learning-driven pipelines to end-to-end, deep learning ones. Especially in the very last few years, the NLP field was ruled by the Transformers architectures, which achieved state-of-the-art performance on numerous tasks. Besides the large improvements obtained with these kinds of architectures, concerns about their explainability have risen. The end-to-end paradigm, together with the complexity of deep learning models, makes it difficult to understand the motivations behind their decisions, which inhibits the interpretation from final users, linguists, or domain experts. Such an issue is particularly felt in a sensitive domain such as healthcare. Furthermore, being unable to understand the mechanisms behind their reasoning inhibits the researchers from getting rid of the current models and providing new solutions.

The present manuscript thus explores the landscape of NLP solutions in healthcare and provides significant contributions to the field. It demonstrates the worth of investigating such technology for improving healthcare, with particular focus on the explainability of the state-of-the-art models, i.e., Transformers, providing new solutions and analyses.

After providing an extensive background of NLP and its advancements, focusing on the solutions proposed in the healthcare literature, we investigated the use of Transformers in both Natural Language Understanding (NLU) and Generation (NLG). For the former, we collected the first dataset for sentiment analysis in Italian for healthcare. In our work we compared Transformer-based and Machine Learning (ML)-based NLP. Quite surprisingly, the classical model outperformed the other, for which we highlighted its sensitivity to data class imbalance.

For the latter, we faced the problem of reducing the expertise gap for patients reading medical texts by proposing a new system to simplify such documents. We employed Transformer-based bi-encoders (also known as Sentence Transformers) to collect new parallel datasets we analyzed in quality and then used to train an encoder-decoder model (again, based on the Transformers architecture). The analysis we conducted with human evaluators assesses without doubts our system to outperform models proposed in the past literature, while providing relevant insights on the automatic evaluation metrics usually employed in this kind of tasks.

Finally, we contributed to overcome the explainability issues, both from the end-user and researchers standpoints. First, we proposed two hierarchical architectures based on Transformers to perform document classification tasks while providing document summaries as an explanation of the decisions made. Using a well-known benchmark in sentiment analysis, we evaluated the two proposed models, highlighting their strengths and weaknesses. Both systems achieved good results, not so far from previous literature, while providing extractive summaries as an explanation of the sentences that were most relevant for the decision. Our proposed evaluation protocols ensured their ability to explain their reasoning.

Then, we conducted a study to investigate the robustness of Transformers when adapting to new domains through the further pre-training paradigm. By inducing minimal variations we disclosed surprising instabilities in fine-tuning.After testing a very large number of combinations, which we briefly summarize, our experiments focused on an intermediate phase consisting of a single-step and single-sentence masked language

modeling stage and its impact on a sentiment analysis task. We discuss a series of these unexpected findings which leave some open questions over the nature and stability of further pre-training and Transformers themselves.

1. Introduction

Natural Language Processing (NLP) is a discipline at the intersection between *Computer Science* (CS) and *Artificial Intelligence* (AI), and *Linguistics* that leverages unstructured human-interpretable (natural) language text. NLP aims to provide computational capabilities to either understand human language or naturally communicate with humans. In the former case, we refer to *Natural Language Understanding* (NLU), and in the latter to *Natural Language Generation* (NLG). In the last decades, NLP became increasingly popular: manipulating natural languages with machines has been shown to disclose plenty of applications, useful for both industrial and research scopes.

The explosion of NLP has affected several fields, including healthcare. Especially in recent years, NLP has been widely applied in health-related domains, from radiology [1] and oncology [2] to chronic diseases [3] and cardiology [4]. NLP can be used for health-specific tasks, for example,

- for mining medical records and clinical narratives, such as doctors' notes and discharge summaries, to interpret (NLU) and extract relevant information, helping to standardize and organize for easy access and analyses [5];
- for generating coherent, natural language text (NLG) based on structured data such as lab test results, which can ease the access to the implications of such results for patients or young medical professionals [6];

as well as more traditional ones, such as mining patients' opinions, which healthcare organizations can use to identify areas for improvement and track changes in patients' satisfaction over time [7]. Overall, NLP can potentially improve the efficiency and accuracy of healthcare information management and support better decision-making and care delivery. The potential applications of NLP in the healthcare scope are uncountable, leading to impressive benefits for all the actors involved in the healthcare process (patients, clinical personnel, etc.). The use of NLP in healthcare is an active area of research and development, and new applications are constantly being explored and developed.

However, the growth path of NLP in medicine and healthcare is far from reaching its end. As the healthcare industry continues to generate large amounts of unstructured text data, such as Electronic Health Records (EHRs) and clinical narratives, as well as other sources, e.g., social media, there will be increasing demand for NLP techniques. Such a path still shows numerous barriers to be faced. Researchers in the field are required to develop sophisticated and robust algorithms, leading to new and improved applications in a domain as particular as healthcare. In particular, they must deal with the lack of data, mainly due to privacy issues [8]. Healthcare-related free texts are usually filled with personal data, arising concerns about patient privacy. To be compliant with the regulations provided by the legislators[1] in the matter of processing and sharing personal data, healthcare organizations are often reluctant in providing access to the real-world data for researchers. As a result, it makes it difficult to collect and share the large amounts of data needed to train, evaluate, and compare NLP algorithms. Another issue to face is the demanding of explainability in automated decision-making processes [9], enforced by the concept of the *right to explanation* legislated by government organizations, as the right to provide *meaningful information about the logic involved in automated decisions* [10]. Especially in healthcare, where decisions made by an automatic system can have significant implications for patient care and treatment, healthcare providers need to understand the bases for such decisions. Delivering clear and understandable explanations for the decisions is directing the research on developing transparent, accountable, and trustworthy NLP algorithms.

The explainability concern, in particular, got emphasized by the adoption of end-to-end Deep Learning (DL) methods in the NLP literature of the last decade(s). Introducing technologies such as word embeddings, Recurrent Neural Networks (RNNs), and Transformers pushed NLP towards new horizons of improved performance with less human effort, at the cost of the interpretability of the designed systems. Prior models exploit features based on linguistic knowledge, for which one could observe the importance assigned by a statistical NLP model for a better understanding of the model and a justification of its output. Encoding the information end-to-end instead makes it harder to gain such insights.

Furthermore, these systems may be particularly affected by class imbalance [11], exacerbated by the lack of real-world data in healthcare and the particularities of such

[1]Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons about the processing of personal data and the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation, a.k.a. GDPR)

a domain. Overall, thanks to, or because of, these aspects, the growth path for NLP in healthcare is likely to continue for the foreseeable future. Hopefully, the present thesis will stand as one of the so many building blocks of it.

1.1. Objectives, Contributions, and Organization of the Manuscript

The main objectives of the present thesis fall into deepening the study of Natural Language Processing in Healthcare and demonstrating its potential for improving the efficiency and quality of care delivery while also tackling the emerging challenges of using state-of-the-art models, i.e., the Transformers neural networks. Alongside the manuscript, we present the contributions we made during the years of my Ph.D. in this scope.

Part I: Background To bring people unfamiliar with the topics here presented to appreciate the efforts and the results obtained in these years, in the next chapter, we introduce the reader to a technical background (Chapter 2) regarding the recent advances portrayed in NLP, and its role in the recent literature and real-world applications in the healthcare (Chapter 3). In particular, in collaboration with the physicians of the *Department of Orthopaedic Surgery at University Campus Bio-Medico of Rome*, we deepened the contribution of NLP in the care of Low Back Pain (LBP) and the related spine diseases, which we systematically present in Chapter 4.

Part II: Tackling Healthcare with NLP Then, we illustrate how we tackled some of the most anticipated tasks in the healthcare context under the perspectives of understanding (NLU) and generation (NLG) of natural language. For the former, we covered the demand for automatic tools for sentiment analysis for healthcare companies in Italy (Chapter 5). Being the earliest work facing such a task for Italian healthcare, we present the first dataset in the literature consisting of patients' reviews of care clinics and hospitals written in Italian. Our analysis highlights some flaws of the ground-breaking, Transformers models. In particular, the evaluated Transformer model resulted to be weakened by the class imbalance of the training data, and even the implementation of oversampling strategies did not lead to outperforming more traditional models such as

Support Vector Machines (SVMs). It led us to conclude that, under some conditions, more-powerful systems are not a priori the best choice, especially in lexical tasks such as sentiment analysis.

For the latter, instead, we aimed at developing a system for facilitating the communications between physicians and patients (Chapter 6). The development required two separate steps. First, we employed Semantic Textual Similarity (STS) techniques to collect new parallel datasets. Each sample consists of a pair of texts, i.e., a statement written for physicians (domain experts) and the version for patients (domain lay) associated with our trained model with a given similarity (content-preservation) score. By collecting such datasets, we overcome one of the main issues in Text Style Transfer (TST), allowing us to train, in a supervised way, models able to simplify experts' text for lay people. The analysis we conducted with both automatic metrics and human judgments shows that our system significantly outperforms state-of-the-art methods in simplifying medical texts, which would positively impact the communication between doctors and their patients. Furthermore, we collected a new parallel database, in which each sample consists of a pair of sentences and a score of content similarity expressed by physicians, which can be used for developing or evaluating new STS systems for healthcare.

Part III: Explaining Transformers In the third part of the manuscript, we illustrate our efforts in tackling the explainability issue of modern models, i.e., Transformers. In Chapter 7, we present the systems we designed as hierarchical Transformers for performing document classification while providing a summary of the document as an explanation of the decision made by the model by extracting the sentences that plausibly most influenced the model. In particular, such summaries enable the possibility to check for eventual errors in the made decision. Being interested in the validity of the proposed methodology, instead of focusing on a task in the healthcare domain, we assessed it on a well-known benchmark in the NLP community regarding sentiment analysis. Although, the proposed methodology is easily transferable to any document classification task, even in healthcare.

The success of Transformers is, in large part, given by their pre-training phase. In many cases, general-domain models were adapted to vertical domains, such as medicine and healthcare, with an intermediate pre-training phase. However, besides the importance covered by such phases in modern NLP, the mechanisms underlying their success

have been questioned. In Chapter 8, we illustrate the results of our experiments in investigating such mechanisms. In particular, we discovered a surprisingly non-robust behavior of this kind of models. While our work partly poses even newer questions, we believe our efforts may help to uncover new details on these models.

Part IV: Conclusions In the end, we leave space for the final considerations following the work portrayed so far.Furthermore, in the Appendix, we report an overview of the contributions in computer science and bioengineering made during this period but are unrelated to the topic of the present manuscript.

Part I.

Background

2. Natural Language Processing is What we Need

Natural Language Processing (NLP) is a remarkable technology that is already revolutionizing how we interact with computers, automatizing many language-based tasks. Since the advent of the first computers, we dreamt about giving machines the skills to understand and generate natural language. Such a dream lasts from the first rule-based machine translation systems in the early '50s [12] and domain-specific chatbots such as ELIZA [13] to the present day. However, the human language's vast variability and complexity make it difficult to process (and "understand") texts for these systems. The establishment of NLP as one of the most valuable technologies of our times occurred around the '90s. Then, we began to abandon deterministic, pattern-matching systems in favor of statistical methods [14], a.k.a., the conventional Machine Learning (ML), and later, in the early 2010s, with the adoption of *end-to-end* Deep Learning (DL) techniques.

2.1. What is Natural Language?

Textual data is intrinsically different from other sources of information (e.g., images, videos, audio), presenting unique characteristics that make them particularly challenging to handle [15]. First of all, text may be produced by different authors and languages (e.g., English and Italian), at times (e.g., modern English and Shakespearean English are different), and for functions (e.g., poetry and technical reports). Text is also inherently discrete, being combinations of symbolic units (e.g., characters, words); for example, I can use the units *{I, s, h, o, t, p, a, j, m, ..., shot, pajamas, elephant}* to build the sentence

> *"I shot an elephant in my pajamas."*

as well as hierarchical: starting from the morphemes (morphology), we create and com-

bine words following the given language rules (syntax) to convey some meaning (semantics). Furthermore, language can be ambiguous; in the sentence above, for example, the prepositional phrase *in my pajamas* may either be an adverbial modifier, telling under which conditions I shot an elephant, or an adjectival modifier indicating that it was the elephant wearing my pajamas. Language is also compositional: units such as words can combine to create phrases and again to create larger phrases. Combining different words/phrases can easily lead to change in the interpretation of the rest of the text; for example, the composition from Groucho Marx in the movie *Animal Crackers*

> *"One morning, I shot an elephant in my pajamas. How he got in my pajamas, I don't know."*

resolves the ambiguity in the previous example. Furthermore, any human language follows two laws:

- *Zipf's law*: there will be a few very frequent words, and a long tail of rare terms; as a consequence, NLP algorithms should be particularly robust to observations that do not occur in the training data.
- *Heaps law*: given a **corpus** (collection of documents) of N words (or *tokens*), the number of unique words (i.e., the vocabulary size $|V|$) is proportional with (a root of) N, which leads to sparse representations of documents in terms of the occurring words.

All these characteristics make it particularly difficult to work with textual data. However, natural language, in both spoken and written forms, is the primary mean of human communication. As is, these unstructured media are a precious source of information that cannot be ignored in today's world.

2.2. From Rule-based Systems to Transformers

The processing of natural language presents two perspectives: linguistic knowledge-based NLP works by *transforming text into a stack of general-purpose linguistic structures* to build applications on top of these linguistic structures [15], while the "natural language processing from scratch" [16] train end-to-end systems without linguistic annotation, currently dominated by Deep Learning (DL) methods. Figure 2.1 shows the evolution of NLP through time, summarizing the main pros and cons of the two paradigms.

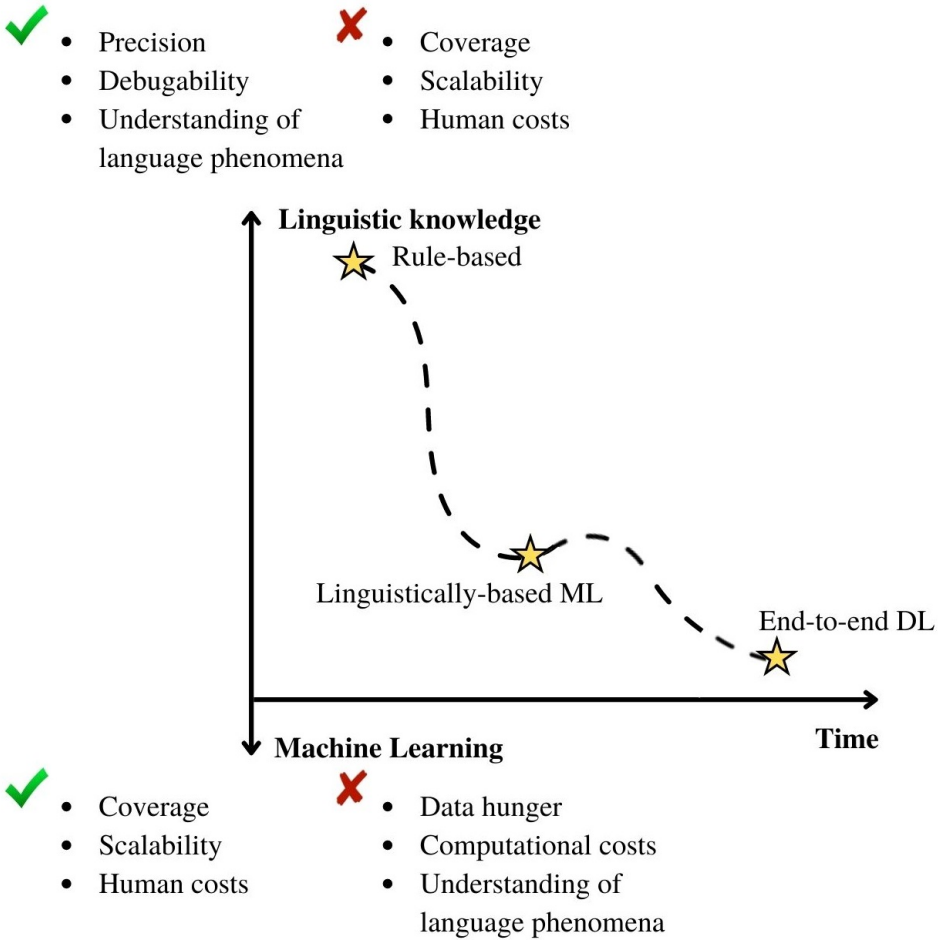

Figure 2.1.: NLP evolution: from linguistic knowledge to end-to-end machine learning.

Although their deterministic nature assures high precision and their being easy to debug and interpret, rule-based methods usually present only moderate recall and are hardly scalable to new data. Plus, they exploit plenty of linguistic knowledge, regular expressions, and heuristics, thus requiring relevant efforts from expert developers and linguists to design such systems.

With classical Machine Learning, we were able to mitigate the negative aspects of past systems. By extracting meaningful features from the text, we train ML models to obtain higher coverage in a relatively fast and scalable way. However, implementing this kind of systems still requires human efforts in designing relevant features for the task at hand. To manipulate text with ML models, we need to extract a mathematical representation of words and documents, i.e., numerical vectors passed in input to the models. To build such representations, we pull out features from the bare text. This aspect is, perhaps, the most crucial one in classical NLP.

We can roughly divide textual features into two categories, *statistical* and *hand-crafted* features. In the former, considering a vocabulary $V = \{w_1, \cdots, w_{|V|}\}$, where w_i represents the i-th word or character (or, generally speaking, token), we may represent documents as vectors of dimension $|V|$ as either

- One-hot Encodings (OhE), in which the i-th entry can assume the value of either 1 or 0, representing the presence or the absence of the token w_i in the document;
- Bag of Words (BoW), in which the i-th entry is the number of occurrences of the token w_i in the document;
- Term-Frequency-Inverse Document Frequency (TF-IDF), which improves BoW by integrating information about the frequency of the token w_i in a given (training) corpus as a weighting term [17, 18, 19].

As hand-crafted features, instead, we can extract from the raw text the linguistic information about syntactic and semantic structures, such as Part-of-Speech (POS) tags, orthographic and dependency labels, and named entities. The practice to extract this kind of features resides in using other ML and rule-based methods [20, 21]. Nevertheless, these methods are usually limited to a monolingual setting, with their performance tending to fade when used in domains different from the one(s) they were trained at first [22].

This kind of representation does not take into account the word orders. Recalling the principle of compositionality of language in the previous section, a common technique is to use *n-grams* of features. Consisting of n sequential features, e.g., words, they capture more complex structures, i.e., context. For example, while *good* conveys a positive meaning, *not good* conveys the opposite sentiment; similarly, while *Paris* indicates a location entity, *Paris Hilton* refers to a person. By the way, recalling the Heaps law, considering large values for n (≥ 3 for words) exacerbates the *sparsity issue* these kinds of representations bring.

Besides these complexities, the ML approach is still being successfully employed: in Chapter 5, we show how a Support Vector Machine can overcome most modern models, i.e., BERT [23], on a sentiment analysis task, using raw and hand-crafted features.

However, with such representations is not easy to deal with single words, e.g., in a task of token classification or in the scope of analyzing a sequence of tokens as is. Here, the *Distributional Hypothesis* (DH) [24] came to help. The hypothesis states

that *each language can be described in terms of the occurrence of parts relative to other parts* [25]. Such a hypothesis led to the traditional *Distributional Semantics* (DS), and its *Models* (DSMs), based on *counting* the co-occurrence of target words in the environment (context) [26, 27, 28, 29]. In other words, *difference of meaning correlates with difference of distribution*, thus similar items, e.g., words, lie as close co-occurrence representations in the n-dimensional vector space, as shown by a simplified example in Figure 2.2 [29].

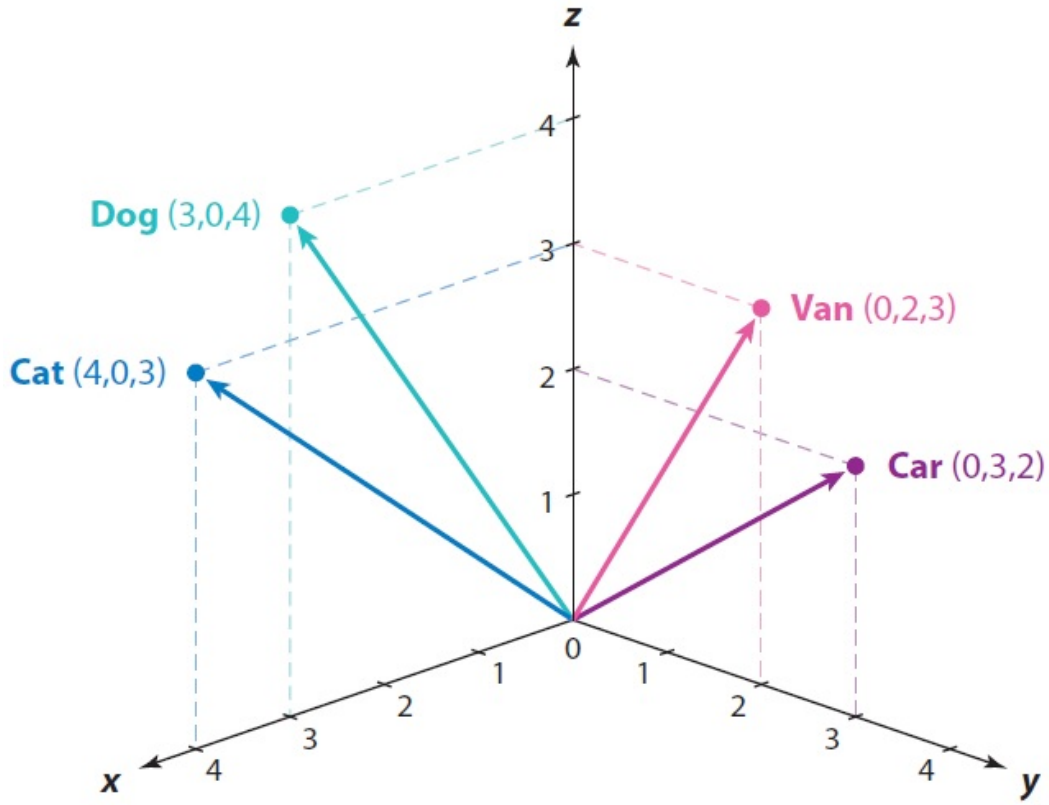

Figure 2.2.: Example of distributional vectors of the lexemes *car*, *cat*, *dog*, and *van*, in a simplified, three-dimensional visualization.

However, in real scenarios, n represents the number of items in the entire corpus at hand, which means that the vector space still tends to be very high dimensional.

It was in the last decade that researchers developed Deep Learning methods that successfully embedded dense representations for words. In 2013 *Word2Vec*, in both its *Skip-Gram* and *CBOW* variants [30, 31, 32] overcame the counting paradigm, advancing the representation learning towards a predictive paradigm [33]. Later, other good tools came out, e.g., *GloVe* [34] and *FastText* [35].

Deep Learning has widely improved the capabilities of NLP systems, allowing them to handle a range of languages and language styles, achieving better performance with the drawbacks of the need for high-volume datasets for training and computational costs. However, DL allowed the development of *end-to-end* systems, abandoning the *feature-engineered paradigm*, thus minimizing the need and efforts of linguists and experts in the application domain. Figure 2.3[1] shows the differences between classical and DL-based NLP.

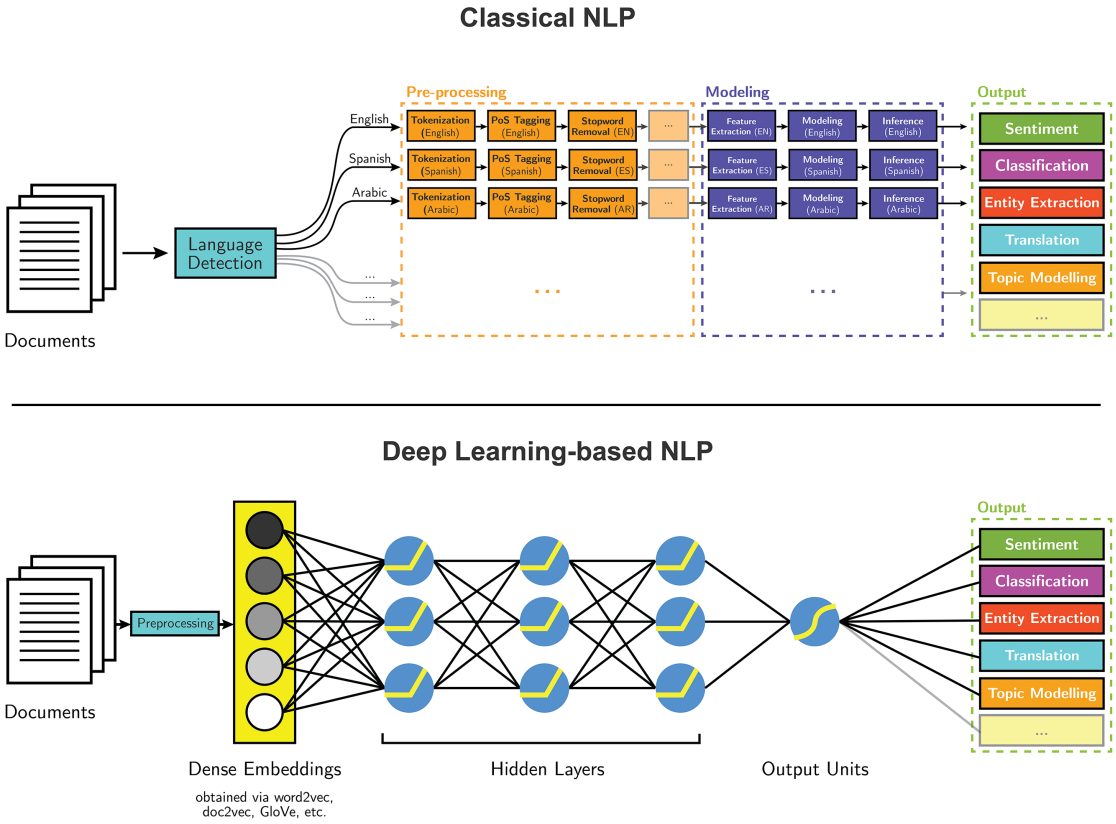

Figure 2.3.: Differences in the pipelines of classical NLP and Deep Learning-based NLP.

The success of Deep Learning in NLP is given by several factors [36], such as the increase in computational power, the availability of high-volume corpora (e.g., from the web), and architectures able to elaborate sequential inputs. Unlike traditional feed-forward ones, Recurrent Neural Networks (RNNs) [37] exploit feedback connections to let the information from previous timestamps flow through the network. Such a strategy

[1] https://s3.amazonaws.com/aylien-main/misc/blog/images/nlp-language-dependence-small.png

allows RNNs effectively process data sequences, capturing the structure and dependencies from the context, especially with the bidirectional variants [38] (Bi-RNNs). Such networks were largely implemented by the NLP researchers, even for computing contextualized word embeddings by computing other NLP tasks such as Machine Translation (MT) and Language Modeling (LM), as for *CoVe* [39], *ELMo* [40], and *ULMFiT* [41]. Unlike static word embedding mentioned above, in which each word is associated with a unique vector, contextualized representations change dynamically with respect to the context. For example, in the sentence

*"I had my booster **shot** today."*

the term *shot* has a completely different meaning from the example reported in the previous section. However, it would be represented with the same static word embeddings in both contexts.

Despite their common usage in past years in complex NLP tasks, such as LM [42, 43], such networks are affected by the exploding and the vanishing of the gradient during training [44], which let the capture of dependency to fade for long sequences. Furthermore, their sequential nature makes them computationally expensive to train, limiting their application to restricted corpora. While more powerful variations using Long-Short Term Memory [45] and Gated Recurrent Unit [46] cells (LSTMs and GRUs) helped to partially overcome these issues, their appeal faded in favor of Transformers, a very recently proposed feed-forward architecture.

Introduced in 2017 by researchers at Google [47], Transformers have since become one of the most successful architectures for NLP, becoming almost ubiquitous in the recent literature, especially after the *Bidirectional Encoder Representations from Transformers* model, a.k.a., BERT [23], came out. Their success resides mainly in the *self-attention* and *positional encoding* strategies. Based on the idea of the attention mechanism first introduced in computer vision [48] and neural machine translation [49], the former allows to simultaneously focus on different elements in the input sequence. The latter, instead, comes in help for recovering the information about the sequence order. Rather than employing recurrence to sequentially process the input sequences, these strategies allow Transformers to capture long-range dependencies in the input data and parallelize the processing. As a consequence, researchers were allowed to self-supervised (pre-)train deeper and deeper architectures on high-volume, unlabeled corpora. Examples of such models are the so-called large Language Models, such as BERT [23], in its *Base* (110 million parameters) and *Large* (340 million parameters) versions, its optimized variants

RoBERTa [50] and DistilBERT [51] (the latter, counting "only" 66 million parameters), and the OpenAI models family, i.e., GPT [52] and its 1.5 and 17 billion parameters successors GPT-2/3 [53, 54]. These task-agnostic pre-trained LLMs have shown extensive improvements when fine-tuned on a multitude of downstream tasks, either at token-, sentence-, and document-level, reaching state-of-the-art performance even in a benchmark like the GLUE one [55].

The transfer learning paradigm [56] has, indeed, shown great benefits instead of training these models from scratch on the target task. To further increase the performance, an intermediate step of pre-training to adapt a general-domain model, such as BERT and ALBERT, to a specific domain, such as the healthcare one, i.e., *BioBERT* [57] and *ClinicalBERT* [58], and *BioALBERT* [59], also for languages different from English, e.g., *KM-BERT* [60]. Although the success of this strategy is commonly attributed to the pre-training phases which would allow learning linguistic knowledge from large corpora [61, 62, 63, 64, 65] that is then exploited during fine-tuning, recent studies showed the benefits of pre-training with either small, noisy or even non-human language data [66, 67, 68, 69, 67, 70]. In Chapter 8, we present our contribution to the analysis of further pre-trained models, discovering how tiny, numerical differences induced by one only sentence lead to astonishing differences in fine-tuning. These recent results suggest that such benefits are induced by pre-training mechanisms not yet fully elucidated.

Another issue in these models is the lack of interpretability. Moving along from the more linguistically-based methods, understanding the mechanisms involved fades. In particular, while in classical NLP we may still be able to observe the importance given to each feature by the statistical models to figure out the reasoning behind a model decision, in the end-to-end paradigm, the information flows through repeated transformation processes, turning it into latent representations. The abstraction processes make it increasingly difficult to understand what kind of linguistic knowledge is encoded at each step [71] and explain the model decisions. Using these techniques for increased performance is countering the trend of *right to explanation* legislated by government organizations [10]. Shifting the paradigm from black-box models to models that provide understandable explanations to final users, or to domain experts [72], without impacting the performance is a well-known problem in the XAI (eXplainable AI) research [73, 74, 75], as well as in NLP [76].

Several methodologies have been proposed in the literature so far, such as *LIME* [77, 78] and *integrated gradients* [79]. Attention has been discussed [80, 81] as a less burden-

some alternative, even for Transformers: the *BertViz* tool [82], for example, provides an interactive interface to visualize attention weights between tokens for every attention head in every layer. In Chapter 7, we present our approaches to exploit hierarchical architectures involving Transformers, also in combination with the *attention-as-explanation* paradigm, to extract collateral summaries from the model during a document classification task.

We can see that more advances are portrayed in the processing of natural language, and newer problems need to be faced. As a reminder of this chapter, the research in NLP is far from reaching its end, increasingly evolving to handle problems of more and more complexity, making things that were inconceivable just a few years ago. A striking example is the *ChatGPT* dialogue-specialized chatbot trained by OpenAI[2]: just released the past November 30, in a few days, it has conquered the attention of the scientific and NLP community, as well as renowned newspapers, becoming a trending topic on social media and gathering 1 million users in less than a week.

[2] https://openai.com/blog/chatgpt/

3. The role of NLP in Healthcare

The healthcare industry is rapidly moving through a digital transformation. One need only thinks of the increasing adoption of Electronic Health Records (EHRs) in hospitals and clinics around the world [83]. These clinical documents can contain about 80% of unstructured data [84]. While conveying various kinds of information, from videos to images, as well as other monitored biosignals, a consistent amount of information in the EHRs is in the form of free text. Besides allowing physicians and the other figures involved in the clinical process to provide a more comprehensive description of the patients' health status [85, 86], free texts often require them a longer documentation time [87, 88], both for understanding or redact those reports. Not a case, the development of NLP in the health-related area was concurrent with the increasing adoption of EHRs in the clinical practice [89, 90, 91], in search of tools to efficiently manage the unstructured data.

Analyzing patients' medical records with NLP techniques can ease physicians' access to a lot of information by providing, for example, summaries contained in the notes daily produced during the care process [92, 93]. In this way, physicians would be allowed to examine and extract more quickly relevant information such as diagnoses, medications, treatment plans, and so on. Also, automatically compressing the information in summaries of medical records may be used to develop tools for automatic diagnoses [94] that can be used as Computer-Aided Diagnosis (CAD) systems by physicians. In many cases, the classification of a diagnosis or a procedure is conveyed by standard codifications. The International Classification Diseases (ICD) nomenclature, for example, is an important tool used worldwide, maintained by the World Health Organization (WHO) to simplify the comparison of health data within and across populations and ease epidemiology analyses. This classification procedure is, nowadays, entrusted by trained staff or medical personnel without professional training. Apart from physicians, healthcare companies may benefit from an automatic coding system, too. Hospitals use the ICD codes to group patients' stays to Diagnoses Related Group (DRG) codes [95],

which are used assign to determine the remunerative reimbursements from the patients' health insurance or the national health systems. Such a classification procedure is often entrusted to medical personnel. Considering its overwhelming and time-consuming characteristics, it is a task extremely complex even for professionally trained staff, and highly error-prone, which may lead to financial losses and potential legal consequences as well. Not a case, several researchers have focused their efforts on automatic systems for ICD coding [96] from clinical notes, e.g., discharge summaries. Recurrent [97] and convolution [98] approaches have been proposed, using attention mechanisms to highlight the most important *word n-grams* as an explanation. More recent works used Transformer-based models [99], even in hierarchical configurations [100]. Similarly, the approaches we present in Chapter 7 may be extended to multilabel ICD classification while providing extracts from the documents as an explanation of the decisions.

Interpretability is an essential aspect in modern research in NLP for healthcare, together with customizability and integration of heterogeneous information [9]. The lack of interpretation, in particular, undermines the adoption of NLP systems in this sensitive domain, leading patients and physicians to have low trust in these tools. However, NLP researchers have to tackle other issues in developing new methodologies: more than in other domains, the unfolding of NLP in healthcare has not been (and still is not) without hurdles. Concerns regarding patients' privacy restrict access to shared data, inhibiting co-operations and reproducibility among NLP researchers' teams. Narrative reports are full of sensitive information regulated by legislation, as the U.S. Health Insurance Portability and Accountability Act (HIPAA) of 1996 [101, 102]. The high costs and reliability issues of de-identifying such reports is one major barrier [8] for the NLP community in healthcare. Not a case, the first shared task for clinical NLP proposed in 2006, the first *Integrating Biology and the Bedside* (i2b2), focused on automatically removing *Private Health Information* (PHI) from medical discharge records [103]. Interestingly enough, the task was proposed by the Clinical Informatics (CI) community [104], while we had to wait until 2013 for the first shared task in healthcare proposed by the NLP community, in *CLEF eHealth Task 2 Disorder Mention* [105].

Many researchers put their efforts working on publicly available databases of clinical notes. For example, the MIMIC (Medical Information Mart for Intensive Care) databases[1] [106, 107] allowed researchers to work on a large amount of (not only textual) data overcoming regulatory obstacles. Many NLP researchers used such data for plenty

[1] https://mimic.mit.edu/

of tasks, such as identifying ICD codes from discharge summaries [97, 98, 108, 109], recognizing entities of medical interest such as drug names and their dosage [110] or other concepts [111], as well as producing clinically, semantically meaningful word representations exploiting the most recent models as Transformers [58].

Others have managed to create facilities like lexicons and ontologies for medical and clinical informatics that can be exploited by NLP practitioners. Such resources can help in tackling the challenges of the medical and clinical language [112], like several different clinical, biological, and medical domains, each with its lexicon and its acronyms, abbreviations, ambiguous names, entities, and variants. The Unified Medical Language System[2] (UMLS) is probably the most known compendium of the field. Started in 1986 by the National Library of Medicine to promote the development of interoperable biomedical information systems, e.g., EHRs, it *integrates and distributes key terminology, classification and coding standards, and associated resources* in the form of:

- a metathesaurus of biomedical concepts, their definitions, and relations;
- a semantic network of groups of these concepts and their semantic relations;
- a specialist lexicon [113].

NLP researchers in health-related fields can exploit data contained in this kind of resource to *improve the ability of computer programs to "understand" the biomedical meaning* [114, 115, 116].

Apart from EHRs and compendiums, with the increasing growth of the *World Wide Web* (WWW), the NLP community is exploiting data from resources like forums and social media intending to help health care and services improve. The analysis of social media texts, in particular, brings several challenges, i.e., noisy and full of domain-specific data [117], which often shows to medical-domain tools, e.g., MetaMap[3] [118] and cTakes[4] [119], to fail [120]. However, such an analysis is still useful for a set of public health applications [121]. Social media allows us to identify the trends in the prevalence of certain diseases like influenza [122, 123, 124] and, very recently, *COVID-19* [125], or for pharmacovigilance [126, 127, 128]. Among the several scopes, mental health surveillance is one of the most faced topics in literature [129, 130, 131, 132], with particular focus on the identification of suicidal intentions [133, 134]. Apart from

[2] https://www.nlm.nih.gov/research/umls/index.html
[3] https://lhncbc.nlm.nih.gov/ii/tools/MetaMap.html
[4] https://ctakes.apache.org/

the general-domain social media, other platforms are arising from the WWW, grouping users with similar interests [135]. Known examples of such online health communities are *PatientsLikeMe*[5],*DailyStrength*[6], and *Our Data Helps*, in which users donate their "friends and family"-visible social media data and annotated numbers and dates of past suicide attempts through the platform[7].

One typical application that can exploit patients' annotated data is the Sentiment Analysis (SA) of online reviews. As it happens in other fields where consumers check for reviews before purchasing a product or service, patients search for opinions from others with similar health-related experiences [136, 137, 138]. Healthcare companies can benefit from the automatic mining of patients' opinions to individuate strengths and flaws of their care services and treatments. Automatic tools would allow them to abandon the traditional structured surveys and questionnaires, which limit patients' expressiveness and are costly and time-consuming to design and then analyze, while also approaching larger amounts of reviews. Besides several sentiment analysis methods developed for many domains, their application to the healthcare domain was not largely exploited initially, especially for languages different but English [139]. In Chapter 5, we present our efforts in developing SA systems for the Italian language after collecting the first dataset for Italian health care.

The opportunities offered by the WWW can be exploited directly by patients too. One of the main merits of the Internet is to "democratically" provide easy access to a large amount of information. Among all the information, the Internet is full of medical knowledge that can be reached even by people without a medical background. When these resources are designed for medical professionals, other people may suffer from the so-called *curse of knowledge* [140] of the medical information present on the web, which may lead unskilled people to misinterpretations [141]. In Chapter 6, we present our approach to addressing such issues and reducing the expertise gap between doctors and patients. Issues in understanding medical information may also be a result of the presence of misinformation on the Internet, especially on social media. Several researchers have put their efforts into recognizing health-related fake news. For example, Sager et al. [142] collected and annotated a small database containing misinformative posts from the *Reddit* dermatology forums, while Patwa et al. [143] developed systems in an attempt to fight the infodemic, a spreading of (potentially harmful) false information,

[5] https://www.patientslikeme.com/
[6] https://www.dailystrength.org/
[7] https://ourdatahelps.org/

on *COVID-19* with a manually annotated database consisting of posts and articles from *Twitter*.

Coming back to the analysis of the EHRs, NLP represents an unprecedented opportunity for biomedical researchers. Researchers can exploit NLP techniques to ease the recruitment of patients for studies by identifying people meeting the study criteria from the clinical notes [144, 145], perhaps with appropriate adjustments to handle variations in clinical documentation between different institutions [146]. It leads to building larger cohorts with fewer efforts, which is extremely useful for researchers in biomedical informatics, too. Using NLP to annotate patients' health status from their reports, they can train other systems on different kinds of data, like images, in a supervised way. This strategy allows the researchers to exploit a large amount of data to train their CAD systems, which is often a strict requirement for data-driven methods (especially deep learning ones). Wang et al. [147], for example, used NLP techniques for detecting the pathology keywords in radiology reports and labeling chest X-ray images. Then, they generated *silver labels* to train a multi-label classification system to detect thoracic diseases from such images.

The NLP solutions in healthcare are not limited to research but are already being seen in some industrial applications. A popular one is represented by chatbots, also known as chatterbots [148], digital agents designed for conversing with specific users, e.g., patients. Such systems can provide useful information in a quick and personalized way, leading to better outcomes for patients' health. Chatbots offer a friendly and entertaining way to educate patients suffering from a chronic or mental health disease, in particular, to entail better adherence to care treatments. For example, the education given to patients with type 2 diabetes by their care providers has shown to be more efficient proportionally with the number of interactions [149]. In this sense, chatbots show promise in helping patients build new healthy habits at a scale, reducing hospital admissions and healthcare costs and times. An example of this kind of application is the AIDA tool[8], a free web software that allows users to ask questions in Italian regarding type 1 diabetes. Conversational agents can also help patients modify their behaviors by providing real-time support to enhance their coping skills in decision-making. Chatbots are a growing technology in the mobile health (mHealth) landscape: Parmar et al. [150] reviewed the *healthbots* from two major digital stores for mobile apps (i.e., *Google Play*

[8] www.aidachatbot.it

Store[9] and *iOS App Store*[10]). However, not surprisingly, they found most healthbots relying on rule-based approaches and finite-state dialogue management, directing the user through a predefined path rather than exploiting the latest, data-driven techniques.

The panorama of applications of NLP in healthcare is extremely vast and unfeasible to review in its entirety. Houssein et al. [151] focused on the machine learning techniques used in the literature for biomedical NLP. To help other researchers overcome privacy issues for their studies, Gao et al. [104] summarized the main publicly available tasks in clinical NLP, i.e., tasks involving publicly available EHR data. Gonzalez et al. [135], instead, reviewed works in mining either EHRs or social media posts. Most past investigations of the literature focused on particular subdomains. For example, recently, Zhang et al. [152] narratively described works in the scope of detecting mental illness, while Chen and Baxter [153] focused on studies in ophthalmology, highlighting new potential applications and limitations. Following their example, given the unfeasibility to discuss in deep the whole landscape of applications of NLP in healthcare, we decided to dive into the NLP in the care of *Low Back Pain* (LBP) and the related spine disorders. We present our systematic analysis in Chapter 4.

The reminder of this chapter is that NLP in healthcare is a thriving field of research, with applications in the real world that can help improve the current care of patients. As NLP techniques continue improving, many opportunities are behind the corner for researchers, leading to new resources to exploit and new tasks and benchmarks (e.g., [154]) to be faced.

[9] https://play.google.com/store/apps
[10] https://www.apple.com/app-store/

4. A Case Study of NLP in Healthcare: an In-Depth Analysis for Low Back Pain and Spine Disorders

To provide more insights into how NLP is striking the world in healthcare, we performed an in-depth analysis of the literature involving a specific case study: applications in Low Back Pain (LBP) and spine disorders[1]. In particular, we collaborated with the physicians of the *Department of Orthopaedic Surgery at University Campus Bio-Medico of Rome*. We decided to focus on this particular topic because of the high impact these diseases have on the patients' health and quality of life, other than the economic burden they bring.

The prevalence of such a musculoskeletal condition is increasing worldwide. A recent study [156] has reported the number of people experiencing LBP at some point in their lives increased from 377.5 million in 1990 to 577.0 million in 2017 globally. Even if the prevalence increases with age, people experience LBP not only in their earlier adulthood but also during adolescence [157]. In particular, chronic LBP is often considered the main reason for disability in a large portion of the population [158]. Even in cases pain does not imply disability, this condition often causes activity limitation and work absence [159, 160], leading to a high economic burden on workers, industries, and governments [161].

Although preliminary, studies concerning low back pain and other related spine disorders with relevant applications of NLP methodologies have been reported in the literature over the last few years. It motivated us to systematically review the literature comprised of two major public databases, PubMed and Scopus. To do so, we first formulated our research question following the PICO guidelines. Then, we followed a PRISMA-like protocol by performing a search query including terminologies of both

[1]The work presented in this chapter is an extract of our paper published in *Frontiers in Surgery* [155] entitled *Natural language processing in low back pain and spine diseases: A systematic review*

technical (e.g., *natural language* and *computational linguistics*) and clinical (e.g., *lumbar* and *spine surgery*) domains. We collected 221 non-duplicated studies, 16 of which were eligible for our analysis.

4.1. Materials and Methods

To perform an exhaustive overview of the applications of NLP in the management of LBP, we interrogated both PubMed and Scopus databases with similar queries. For both databases, we performed the search on November 6^{th}, 2021.

4.1.1. Research question

AI and CS systems have already shown to be a great support to physicians in diagnosing and treating LBP and related pathologies in humans [162, 163]. Here, we aimed to provide a comprehensive review of the literature regarding the described applications of NLP-related methods to the care of patients affected by LBP. Precisely, following the *PICO* guidelines, we aimed to answer the following research question:

- In human subjects, with any demographic information, affected by LBP and related spine disorders {**P**opulation/**P**roblem}
- may NLP methodologies, {**I**ntervention}
- compared with human operators and other already existing tools, {**C**omparison}
- help healthcare providers in the management of such conditions? {**O**utcome}

4.1.2. Research protocol

To exhaustively review the literature, we developed the following research protocol. First of all, we elaborated on a search query. We performed the query on two public databases, namely PubMed[2] and Scopus[3]. For both databases, we considered the title and abstract of the articles. For the Scopus database, in addition, we also took into account the keywords assigned to the papers. Then, we formalized the inclusion/exclusion criteria. We excluded papers not meeting the inclusion criteria from further analyses. After

[2] https://pubmed.ncbi.nlm.nih.gov/
[3] https://www.scopus.com/

conducting the first screening by removing the duplicated articles, two authors carried out a preliminary screening after reviewing the titles and abstracts (and, eventually, the keywords) of the total of the gathered papers. After that, the same authors went deeper by analyzing the full texts. Whenever a discordance happened, the two authors discussed it until reaching a consensus. Finally, we reported in the present review the works retrieved. The developed protocol is resumed in Figure 4.1, reporting the flow-chart diagram realized according to the *PRISMA* protocol employed.

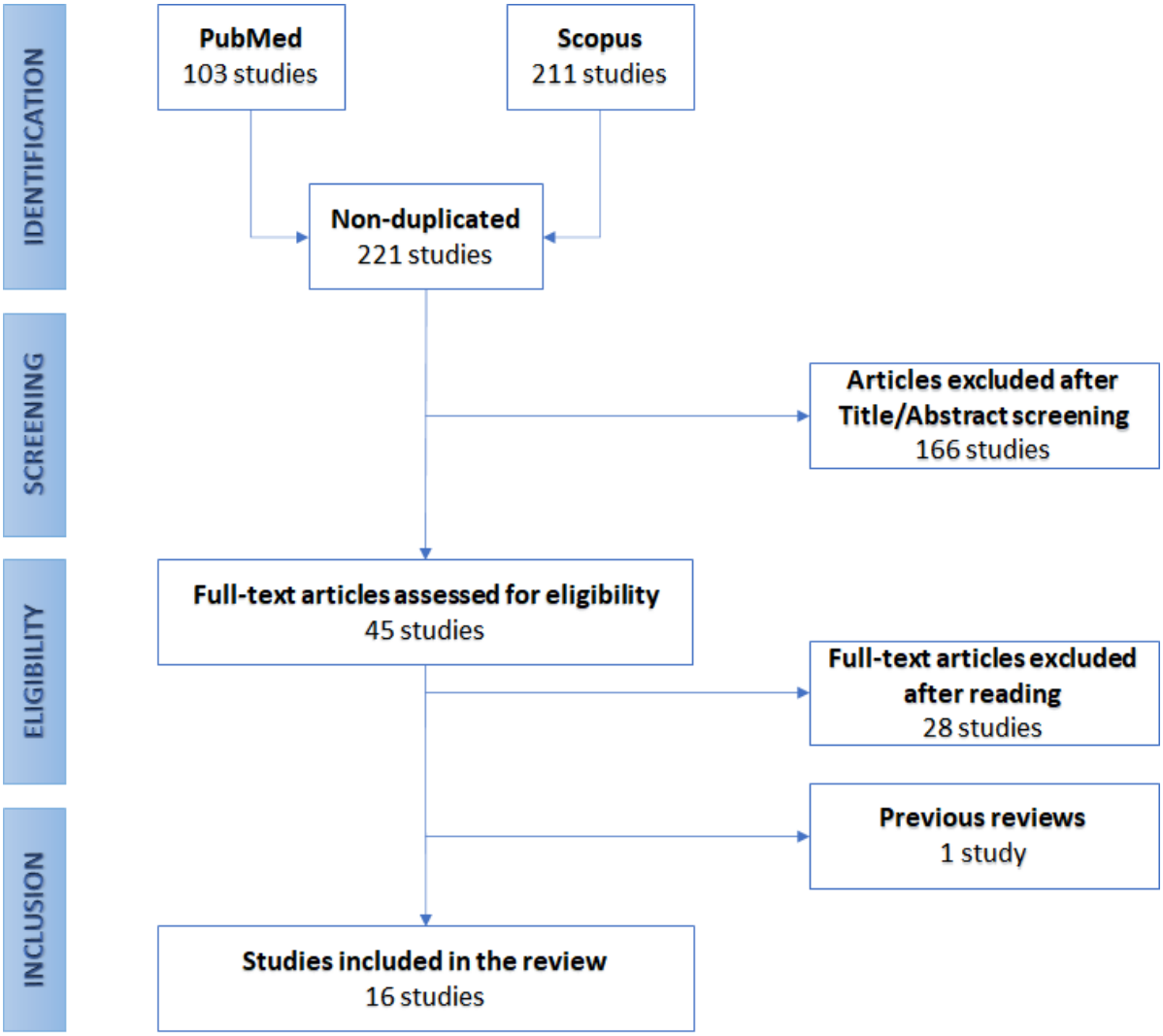

Figure 4.1.: Preferred Reporting Items for Systematic reviews and Meta-Analyses (PRISMA) flow diagram.

4.1.3. Search query

The proposed search query consists of two parts, one including terms from the NLP terminology and the other including terms related to LBP. In each of the two query sections, the terms have been linked by the logical OR operation, while the inter-relation between the two parts has been represented by the logical AND operation, meaning that the papers resulting from the interrogation had to present at least one of the terms for both query sections.

The NLP part contained several terms, each belonging to a particular characteristic of the NLP methodologies. Of course, terms as *natural language*, *NLP*, *NLG*, and *NLU* were directly inherent to the scope. Terms like *computational linguistics* and *text mining* were included because directly related to the NLP field, and are often utilized as interchangeable synonyms. For both of them, there are only slight differences. Sometimes, field practitioners disagree about those differences. Usually, computational linguistics concerns the development of computational models to study some linguistic phenomenon, also concerning other fields such as sociology, psychology, and neurology. For example, a successful approach in computational linguistics may be designing a better linguistic theory of how two languages are historically related. NLP, instead, is mainly oriented toward solving engineering problems by analyzing or generating natural language text. Here, the success of the NLP approach is quantified by how well the developed system resolves the specific task. Text mining, instead, usually refers to turning unstructured text into structured data to further exploit it, e.g., through statistical analysis (data mining).

Instead, terms as *tokenization*, *word embedding*, *rule based*, *regex*, *regular expression*, *bert*, and *transformers* refer to the methods to pre-process, extract features and models used to elaborate unstructured text, while *automated reporting*, *summarization*, *named entity recognition*, and *topic model* refer to specific tasks that can be performed on the text and are typical in the medical domain. Furthermore, we included some other generic terms: *text analysis*, *free text*, *biomedical text*, *medical text*, *clinical text*, *biomedical notes*, *medical notes*, *clinical notes*; and *linguistics*.

The medical part, instead, contains all terms related to the LBP and spine disorders conditions: *low back pain*, *lumbar*, *intervertebral disc degeneration*, *intervertebral disc displacement*, *spondylarthritis*, *spondylolisthesis*, *disc herniation*, *spine surgery*, *spondylarthrosis*, and *durotomy*.

4.1.4. Inclusion and exclusion criteria

This systematic review aimed to gather all the studies concerning the utilization of NLP in the diagnosis, prevention, and treatment of LBP. Straightforwardly, all the selected articles had to meet the following inclusion criteria:

- LBP must have been between the main topics of the articles;
- NLP techniques must have been used in the studies;
- Subjects of the studies: all the articles must have been based on studies of the human spine pathology;
- Language: all articles must have been written in English.

Conversely, we excluded articles that did not meet the inclusion criteria for one of the following reasons:

- Low Back Pain or spine diseases were not considered;
- No automatic tool of text analysis were exploited;
- Animal studies.

4.1.5. Quality of evidence

The methodological quality of included studies was assessed independently by two reviewers. Any disagreement was solved by the intervention of a third reviewer. The risk of bias and applicability of included studies were evaluated by using customized assessment criteria based on the Quality Assessment of Diagnostic Accuracy Studies (QUADAS-2) [164]. This tool is based on four domains: patient selection, index test, reference standard, and flow and timing. Each domain is evaluated in terms of the risk of bias. The first three domains are also assessed in terms of concerns regarding applicability. Sixteen studies were rated on a 3-point scale, reflecting concerns about the risk of bias and applicability as low, unclear, or high, as shown in Figure 4.2.

4.2. Results

The search queries performed on PubMed and Scopus resulted in 103 and 211 papers, respectively. Nonetheless, many of these articles were duplicates. So, as a first screening,

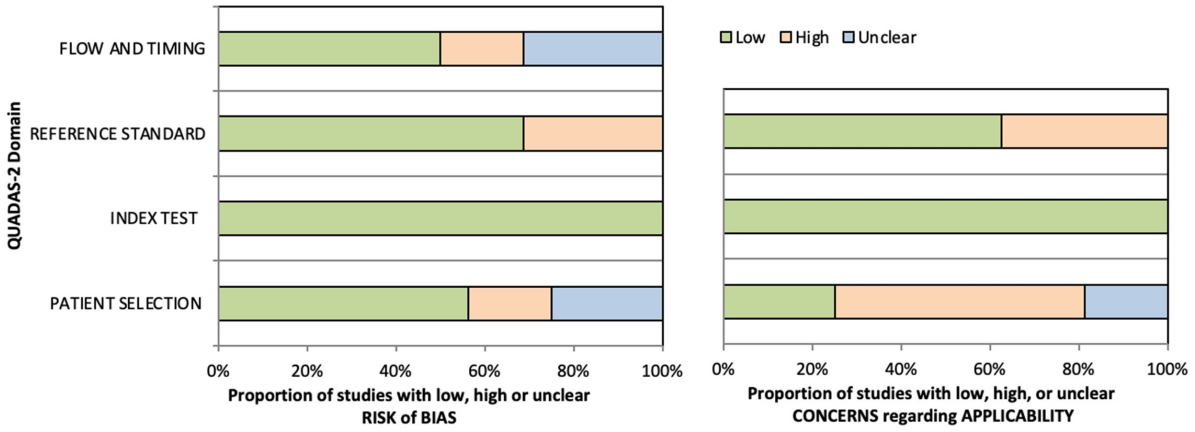

Figure 4.2.: Summary of the methodological quality of included studies regarding the four domains assessing the risk of bias (left) and the three domains assessing applicability concerns (right) of the QUADAS-2 score. The portion of studies with a low risk of bias is highlighted in green, the portion with an unclear risk of bias is depicted in blue, and the portion with a high risk of bias is represented in orange.

we removed the repeated studies, resulting in 221 unique papers. Then, we analyzed the remaining articles' titles and abstracts. In this phase, we excluded the works not meeting the inclusion criteria. This operation reduced the number of eligible articles to 45. Among them, we encountered one narrative review [165], in which *Groot et al.* recently focused on the role of the NLP in spine surgery in six studies from the recent literature. However, since these papers are extensively reported here, we did not further focus on their work here. So, the final screening was performed by reading the full text of each paper, leading to retaining 16 of them. Figure 4.1 graphically shows the described selection process through a flow-chart diagram according to the PRISMA protocol.

In the following paragraphs, we analyze included studies by focusing on the tasks and models in which NLP is involved.

4.2.1. Tasks

We identified three main NLP methodologies, i.e., classification, annotation, and prediction. Both first two approaches concern the identification of a category (class) to which a document belongs, differing for what the NLP methods are applied. In the former, the NLP system associates a label to each testing example (i.e., the patients' document). A classification system may be employed:

- as a CAD system, which the physicians may exploit to decide, for example, whether or not to operate on a patient,
- as a Decision Support System (DSS) which healthcare providers may utilize such a system to improve quality control,
- or to gather a large cohort of patients for some research study.

In the annotation approach, NLP is used to label the documents too. However, it is implemented as a part of the entire system, thought to provide the classification outcome from another kind of data, such as radiological images. From this point of view, the NLP system is a way to automatize the annotation of a large amount of data by identifying specific phenotypes related to a disease condition. The second part of the entire system may be trained and evaluated on a significantly wider amount of data than using only human annotations. This approach is used to develop successful predictors of clinical outcomes from clinical data and better define indications for surgery. It may improve clinical outcomes, thus avoiding invasive spine care and reducing healthcare costs.

The third approach can be referenced as the identification of some categories too. However, here the scope is to predict some outcomes by exploiting previously acquired data (free-text notes, in this case). Healthcare providers may use such a system to predict some outcomes from the patients and thus arrange in advance the resources necessary for their care. Moreover, we further classified included studies based on the timeframe regarding surgical interventions. Thus, papers may also fall in the pre-, intra-, and post-operative task categories, whether the task interests something before, during, or after surgery, respectively, as shown in Figure 4.3.

Classification

Pre-operative tasks We identified several studies in which the authors exploited pre-operative notes to identify useful diagnostic clues and findings. In detail, we retrieved:

- 1 paper focusing on the identification of multiple imaging findings;
- 1 paper focusing on the diagnosis of acute LBP;
- 2 papers focusing on the identification of spinal stenosis;
- 3 papers focusing on the identification of axial SpondyloArthritis (axSpA);

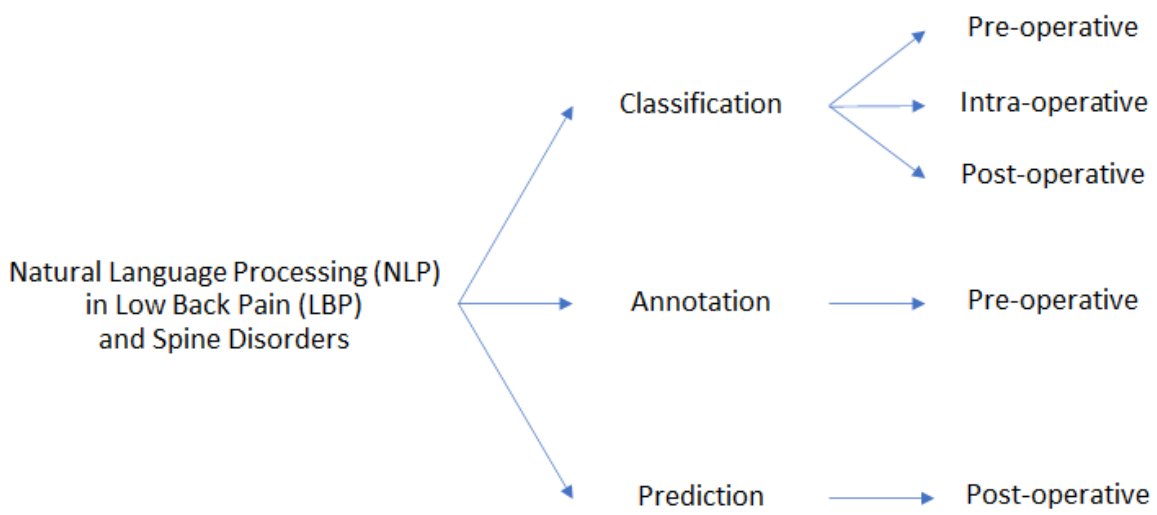

Figure 4.3.: Schematic partitioning of the works concerning the application of NLP in LBP and related spinal disorders.

- 1 paper focusing on the identification of type 1 Modic endplate changes;

Following, we describe the tasks.

Imaging findings identification

To advance the care of patients suffering from LBP, discovering distinct subgroups with similar prognoses and intervention recommendations is a relevant task. Spine imaging findings alone are often insufficient to diagnose the underlying causes of LBP. In addition, they are often not of clinical significance since their frequent occurrence in asymptomatic individuals [166]. To understand the relationships between imaging findings and LBP, an important step is the accurate extraction of the findings, such as spinal stenosis and disc herniation, from large patient cohorts. NLP may help identify lumbar spine imaging findings related to LBP in large sample sizes. *Tan et al.* [167] worked on this task.

Acute LBP identification

LBP events can be classified either as acute or chronic. While the former is usually treated with anti-inflammatories, with the recommendation of returning to perform daily activities soon, care of the latter often involves physical therapy, spinal injections [168], and even spine surgery. Thus, different conditions lead to distinct treatment recommendations and costs for the healthcare systems. *Miotto et al.* [169] faced this task.

Identification of axSpA

AxSpA is a serious spinal inflammatory disease characterized by the additional involvement of peripheral joints, entheses, and other systems (including the eye, the gut, etc.) [170]. As patients with axSpA often present with peculiar imaging features, developing a tool to facilitate the identification of this subset of patients is a key step to achieve in improving the care of this condition. To exploit large datasets, NLP may be used to identify concepts related to axSpA in text, and thus create a cohort of patients with (a high probability of having) the disease. *Zhao et al.* [171] and *Walsh et al.* [172] dealt with this this task. The last team also exploited their previous work in their [173] to identify axSpA patients.

Stenosis identification

Spinal stenosis is a condition of narrowing of the spaces within the spine, which can compress the spinal canal (spinal canal stenosis, SCS) and the nerve roots exiting at each intervertebral level (neural foraminal stenosis, NFS). Such conditions often develop in the lumbar spine. Here, NLP was used to classify both SCS and NFS, also with a severity grading scale (*Caton et al. [174, 175]*).

Type 1 Modic Endplate Changes identification

Modic changes consist of magnetic resonance imaging (MRI) signal alterations affecting the endplates of the lumbar spine and are particularly frequent in patients with LBP [176]. For this reason, *Huhdanpaa et al.* [177] employed NLP to identify the Type 1 Modic changes from radiology reports.

Intra-operative tasks We identified a few studies in which authors exploited operative notes to find evidence of some surgery complications. In detail, we retrieved two papers focusing on incidental durotomy (ID) identification and another on vascular injury (VI) identification. Such complications have potential implications for recovery, causing the length of stay and costs to increase. Thus, an automated system for surveillance of these events is relevant to healthcare providers.

Incidental durotomy (ID) identification

Incidental durotomy (ID) is a common intra-operative complication during spine surgery, occurring up to 14% of lumbar spine surgeries [178]. It is defined as an inadvertent tearing of the dura during surgery with cerebrospinal fluid (CSF) extravasation or bulging

of the arachnoid [179]. The group of *Karhade and Ehresman* faced the problem of automatizing detection of ID events from operative notes [180, 181].

Vascular injury (VI) identification

Vascular injury (VI) refers to the trauma of blood vessels (either an artery or a vein). It is a common event during spine surgery, often resulting in serious bleeding, thrombosis, and additional complications. *Karhade et al.* [182] dealt with the problem of detecting VI events from operative notes.

Post-operative tasks Classification in post-operative tasks serves to identify events occurring after the surgical intervention, i.e., venous thromboembolism (VTE). VTE results from the formation of a blood clot that may obstruct the blood flow locally (thus causing edema and pain) or travel to distant sites causing local blood flow arrest (such as in pulmonary embolism). *Dantes et al.* [183] attempted to identify from post-operative radiology reports the occurrence of VTE in patients who underwent various kinds of surgeries, including spine surgery.

Annotation Among the included papers, two implemented NLP to annotate radiology images. *Lewandrowski et al.* [184] classified findings related to spinal stenosis (both SCS and NFS) from pre-operative reports, while *Galbusera et al.* [185] trained the NLP model to identify several spinal disorders. In both cases, the authors retrieved the annotations for radiology reports and then used them to label the related images. However, in the study by *Galbusera et al.*, it was not possible to identify the timing with respect to surgery, since they included several types of disorders, as well as patients undergoing post-operative radiological examination and follow-up.

Prediction Prediction tasks focus on predicting post-operative outcomes. In their first paper, *Karhade et al.* [186], they attempted to identify required re-operations due to wound infections arising after lumbar discectomy, while in a subsequent study [187] they identified unplanned re-admissions of patients who underwent posterior lumbar fusion. Both tasks were intended to refer to a period of 90 days.

4.2.2. Data

Data used in the analyzed studies is the free text from clinical notes. However, the kind of notes exploited by the authors may vary depending on the task the authors aimed to cover. A large proportion of papers used radiology reports, aiming at identifying imaging findings [167] and diagnosing a specific condition [171, 172, 183, 177, 174, 175], or at annotating images [184, 185].

Other examples include operative notes, obviously used for the intra-operative tasks [180, 181, 182, 183], and post-operative ones too [186, 187]. Furthermore, the article from *Karhade et al.* [187] compared different kinds of clinical notes, including discharge summaries [171], and physicians and nursing notes. With the exception of [185], in which *Galbusera et al.* exploited notes in Italian, all other studies referred to texts written in English.

4.2.3. Models

The studies analyzed in this review used various kinds of NLP models. Referring to Figure 4.4, we identified such models as belonging to the rule-based, the ML-based, or for pipeline exploiting both, the *hybrid* approaches (see Section 2.2). Furthermore, the ML-based approach may be further split into classical ML and DL models.

Also, models may be categorized as belonging to:

- *Supervised* approach, which exploits labeled data to train the model;
- *Unsupervised* approach, in which the algorithm is not provided with any labeled data.

By taking into consideration the above definitions, it is reasonable to consider the rule-based models as belonging to the unsupervised class of algorithms, while the ML-based models may fall into both categories. Nonetheless, the supervised approach is usually more performant because the model learns directly from input-output pairs, while the unsupervised ones leverage only the input data. However, the former approach may require a lot of labeled data, a process that can be extremely time-consuming, requiring several human resources (annotators), especially for large datasets. Annotators in the healthcare field should necessarily have a certain degree of expertise in the domain, which is also one reason for automatizing the annotation process.

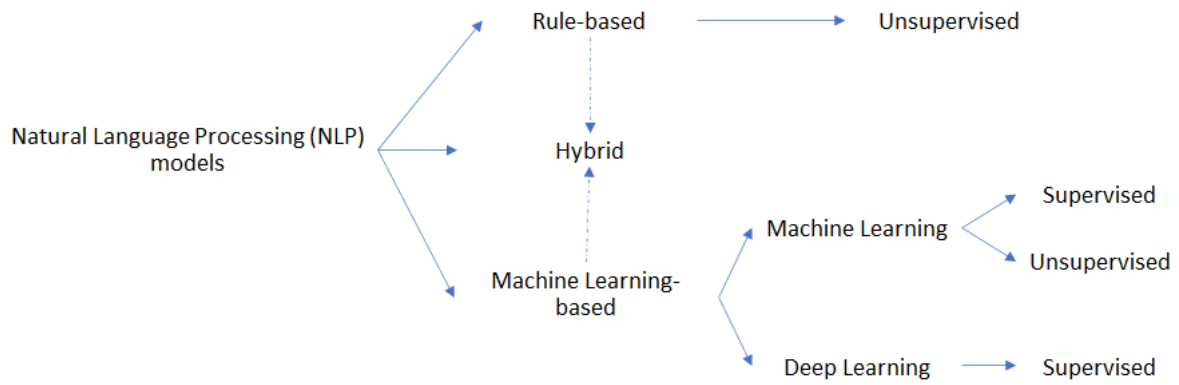

Figure 4.4.: Schematic partitioning of the NLP models applied in LBP and related spine disorders.

Rule-based models Rule-based models are concerned with simple searches of keywords in the text of clinical notes, often by developing regular expressions (regex). These rules may consist of both syntactic and semantic rules, leveraging knowledge from both linguistics and the application domain (knowledge-driven approach). To identify (and remove) negated occurrences, authors usually exploit algorithms such as NegEx [188]. Such an approach was implemented in [169] to identify acuity in LBP, and in [177] to identify Type 1 Modic changes, while in [167, 174, 175] to identify several findings related to LBP and stenosis from MRI and x-ray reports.

Machine learning-based models ML models are algorithms that leverage their experience on previously seen data to automatically improve their performance on some task. Thus, they leverage a data-driven approach, by learning discriminative content from a statistical representation of the input data. The authors of the paper encountered focused particularly on two models from the machine learning literature: Logistic Regression (LR) and eXtreme Gradient Boosting (XGBoost). The former was implemented in [169] for the acuity identification task and in [171] to identify axSpA. In both cases, the model was implemented together with a Least Absolute Shrinkage and Selection Operator (LASSO) regularization. The latter was particularly employed by *Karhade et al.* in several tasks [180, 186, 187, 181, 182]. Another used algorithm was the Support Vector Machine (SVM), employed in [172] to identify clues of axSpA and in [173] both to directly identify axSpA and to extract a feature for a multimodal random forest. Furthermore, authors in [183] exploited *IDEAL-X*, a tool introduced in [189] which exploits the online ML paradigm, to identify VTE following orthopedic surgery.

Deep learning models

The DL paradigm is a subfield of ML regarding algorithms partly inspired by the brain structure and functioning, the so-called artificial (deep) neural networks. Besides DL models are well known to perform better than classical ML ones, to be competitive they require a higher volume of training examples. Plus, the training phase may be expensive in terms of time, especially when researchers do not have access to performant hardware facilities (i.e., Graphics Processing Units, aka GPUs). Probably for these reasons, only a few papers investigated the use of DL models. In [169], the authors compared a convolutional neural network (ConvNet) with classic ML and rule-based models. More recently, in [185] the authors fine-tuned a BERT [23] model pre-trained on general-purpose Italian text ("bert-base-italian-uncased"). Models like BERT are based on the Transformer's architecture [190], introduced a few years ago. Exploiting a pre-trained Transformer-based model to initialize the weights and then train on some downstream tasks has become a standard practice within the NLP community.

Unsupervised models

All the above-reported studies leverage the supervised paradigm to train their models. The authors in [169] investigated the use of unsupervised models to identify acute LBP. They implemented a Latent Dirichlet Allocation (LDA) [191] to perform topic modeling, an unsupervised ML technique that captures patterns of word co-occurrences within documents to determine words' sets clusters (i.e., the topics). They identified a set of keywords among the topics and then manually reviewed them to retain only those that seemed more likely to characterize acute LBP episodes. In other words, they selected the topics including most of the keywords with high probabilities. Then, they considered the maximum likelihood among these topics as the probability that a report referred to acute LBP. Furthermore, the authors in [171] exploited the so-called multimodal automated phenotyping (MAP) [192] to identify axSpA from related concepts and coded features.

Hybrid models For what concerns the hybrid paradigm, we encountered only one paper [167] exploiting it. Here, the authors implemented a logistic regression with elastic-net penalization leveraging several kinds of features. In particular, they also used features extracted with a combination of regex and NegEx.

4.2.4. Explainability

As mentioned in past chapters, current methods may achieve high performance of a specific task but often lack interpretability. The absence of more interpretable feedback together with the output from the model is a great inconvenience, especially in the clinical field. For what concerns the explainability, only *Karhade* and colleagues have addressed it at both global and local (for the single subject) levels among included studies. It was possible thanks to the implementation of the XGBoost. Such an algorithm can provide the importance of each feature in a particular task. For example, in [180] patient-level explanations were provided by highlighting the most important words used by the algorithm to detect ID inside the text. Global explanations were provided by averaging the importance scores of each feature across all patients (the documents) to demonstrate the generally most relevant factors for detection. Analogous reasoning was applied in their other works [186].

4.2.5. Domain-Specific Knowledge

Perhaps unusual when reviewing works from the literature, we conducted a typical NLP analysis of the papers included in this review to extract some domain-specific knowledge from the articles included in this review. In particular, we treated the collection of abstracts as a corpus from which we extracted domain-specific entities to build its glossary.

We then retrieved the relations between them to create the knowledge graph of the domain we can call *Natural Language Processing in Low Back Pain and Spine Disorders*. To do so, we applied the $T2K^2$ suite of tools [193] to obtain the glossary in Figure 4.5, reporting the prototypical form of the entity (the term form most frequently attested in the corpus), its lemmatized form, and its frequency of occurrence. It is worth noting that these domain-specific entities may consist of single nominal terms but also complex nominal structures. For ease of visualization, we report only the first part of the glossary (containing the most relevant terms) in the figure: the ranking follows the domain relevance of the entities, computed based on their *C–NC* value [194]. By looking at the obtained glossary, it is easy to notice that the entities *NLP* (and its variations) and *lumbar spine* are the most relevant ones together with *patients*. We then selected these words as the most representative of the domain (we excluded the term *patients* because too generic) to compute their relations with the other entities in the glossary.

Prototypical Form	Lemma of Term	Frequency
NLP	NLP	46
patients	patient	37
NLP algorithm	NLP algorithm	18
natural language processing	natural language processing	13
lumbar spine	lumbar spine	14
lumbar	lumbar	21
algorithm	algorithm	21
specificity	specificity	20
incidental durotomy	incidental durotomy	12
Methods	method	18
sensitivity	sensitivity	16
axial SpA	axial spa	9
spine	spine	14
axSpA	axSpA	13
deep learning	deep learning	8
free-text operative notes	free-text operative note	6
anterior lumbar spine surgery	anterior lumbar spine surgery	5
models	model	11

Figure 4.5.: Glossary extracted from the abstracts of the papers included in this work. Entities are ranked following their domain relevance. For ease of visualization, only the first part of the glossary (containing the most relevant terms) is reported.

In particular, the relations are computed on the basis of the co-occurrence of the entity in the core sentence (the one in which appear the entity under consideration) and the ones immediately before and after. We report the knowledge graph obtained with such entities and their relations in Figure 4.6. For ease of visualization, we filtered out terms with a frequency lower than 3 and the relations not occurring at least twice.

As interpretable from the figure, the *NLP* entity represents the core of the graph (and thus, in some sense, of the articles' domain). It is worth noting the presence of several diseases related to the *NLP* part (*incidental durotomies*, *axSpa*, *modic changes*, etc.), suggesting the obvious importance of these terms for the domain, and of the terms

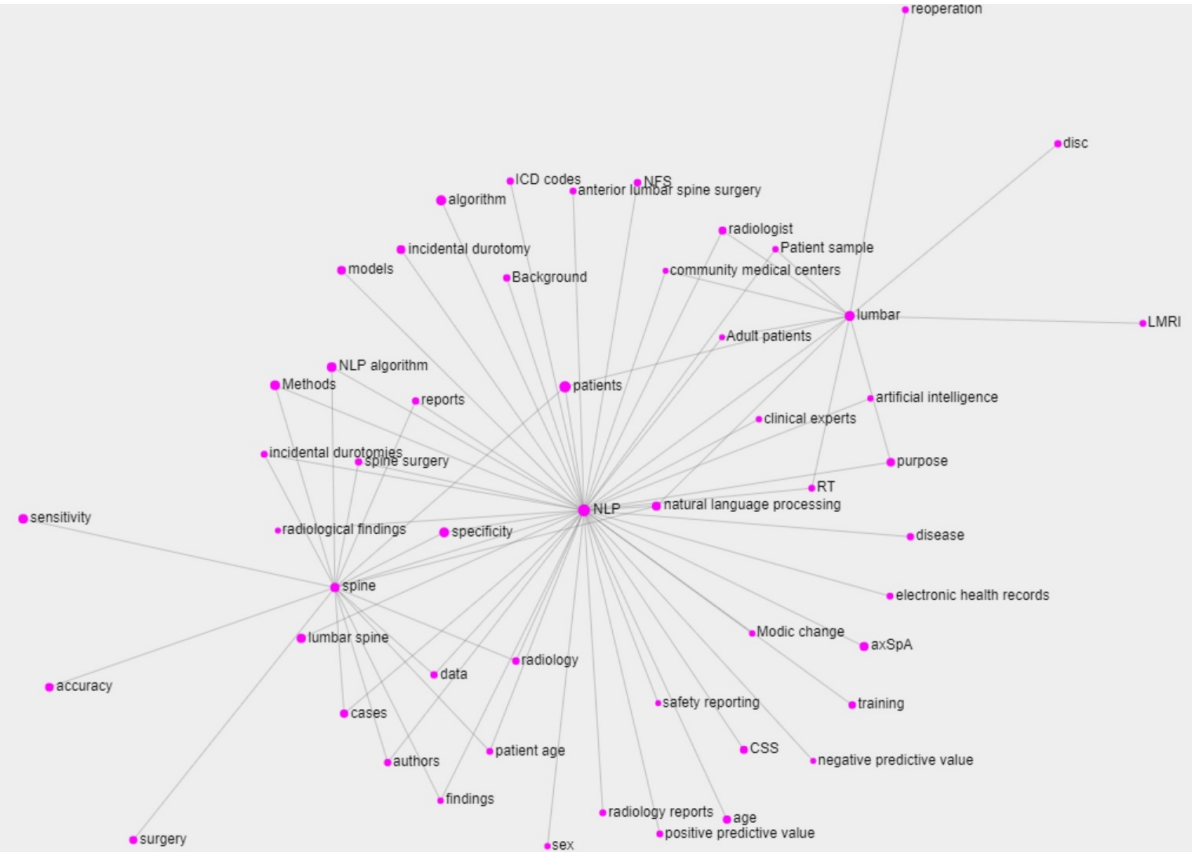

Figure 4.6.: Knowledge graph built for the main entities of the domain extracted from the abstracts of the papers included in this work. For ease of visualization, only the terms with a frequency lower than 3 and the relations occurring at least twice are reported.

related to the computational part (*algorithm*, *models*, *artificial intelligence* etc.) and the data sources (*radiology reports*, *electronic health records*, etc.). However, both the *lumbar* and *spine* entities show a few prerogative relations, such as with *disc* and with *surgery*, respectively, that are not shared with the *NLP* core. Also, apart from the entity *natural language processing* that is just a variant of the *NLP* one, the only relation shared by all the three main entities is the one with *patients*. Besides being a very generic term, this result suggests the focus the authors put on the patients of their works, which also reflects the findings of the glossary.

Part II.

Tackling Healthcare with NLP

5. Capturing the Patients' Perspective: a Case Study for Italian

As more and more content is shared by people on the web, the use of automated sentiment analysis (SA) tools has become increasingly present. When people want to buy a product or service, they often rely on online reviews of other buyers/users (think of online sales giants like Amazon). Likewise, patients increasingly rely on reviews on social media, blogs, and forums to choose a hospital where to be cured. This behavior is occurring abroad [137, 138] as well as in Italy, as demonstrated by the increasing amount of reviews in QSalute[1], one of the most popular Italian-ranking websites in healthcare. Hospital companies often ignore these sources of information, not exploiting the potential of such data to understand patients' experiences and consequently improve their services. Instead, they usually rely on inefficient and time-consuming structured surveys. Due to a large amount of data, there is a need for automatic analysis techniques. To meet this need, we introduced a sentiment analysis system to classify whether a review has positive or negative sentiment[2]. We compared a classical NLP pipeline based on a Support Vector Machine and an end-to-end pipeline based on a BERT model.

While there exist several works on affective computing in several domains for the Italian language [196, 197, 198], at the time of the study, there were no references in the literature addressing this particular domain in Italian. Thus, to the best of our knowledge, ours was the first study of sentiment analysis on Italian reviews in healthcare. Because of this, the first step was to build a brand-new annotated dataset that we publicly shared, helping other researchers handling SA in Italian for this specific domain. Our experiments show the SVM-based system slightly outperforming the BERT-based one, which, in particular, required the employment of techniques for adjusting the class

[1] www.qsalute.it

[2] The work presented in this chapter is an extract of our paper published in *Proceedings of 7th Italian Conference on Computational Linguistics (CLiC-it)* [195] entitled *A Machine Learning approach for Sentiment Analysis for Italian Reviews in Healthcare*

distribution of the training set.

5.1. Web scraping

To collect a large annotated dataset, we employed web scraping techniques to collect data from the aforementioned QSalute website, an Italian portal where users share their experiences about hospitals, nursing homes, and doctors. Web scraping is a technique whose validity in collecting big data is widely acknowledged [199, 200], even for healthcare domains such as epidemiology [201]. Such a technique allows researchers to relatively easily collect data spanned across multiple pages/sections of a website, even overcoming issues of many websites, which do not allow saving on local storage the data displayed [202, 203]. It can be seen as an integrated, preliminary part of the NLP pipeline since most NLP approaches exploit numerous amounts of texts. Web scraping consists of two steps: formatting a request for acquiring resources from a target website/urls, and extracting the desired information from the obtained resources simulating user navigation through the contents [200].

Among the wide landscape of web scraping tools [203], we employed the *Beautiful Soup* [204, 205] Python package. By exploiting its toolkit, we scraped in a programmatic and pythonic way a relatively high-volume of web pages in a relatively short time. We made the back-end code publicly available on github[3].

5.2. Data analysis

The dataset, collected on $May, 26^{th}2020$, consists total of 47224 documents (i.e., reviews). Each document consists of the free text of the review and other metadata such as document id, disease area to which the document belongs, and title. In addition, among the provided metadata, there is the average grade, i.e., the mean over the votes in four categories: Competence, Assistance, Cleaning, and Services.

Here, we assigned documents with an average grade less than or equal to 2 to the negative class (-1) and greater than or equal to 4 to the positive one (1). We labeled the remaining documents as neutral (0). The dataset is strongly unbalanced towards the positive class: 40641 positive, 3898 for the neutral, and 2685 negative reviews.

[3] www.github.com/lbacco/Italian-Healthcare-Reviews-4-Sentiment-Analysis

After a manual inspection, the reviews belonging to the neutral class have shown to be ambiguous borderline cases, which are discernible into either the positive or the negative class by humans. Thus, we discarded neutral reviews to work with highly polarized data, thus resulting in a binary sentiment classification task and a total of 43326 reviews. The following analyses are then referred to this subset: in Table 5.1, we report some characteristics of the dataset for each site (i.e., the disease area), while the distribution of tokens over their length is reported in Figure 5.1.

Site	**Positive / Total**	**Lexicon**	**Overlap(%)**
Nervous System	9984 / 10595	34827	69.93
Hearth	5297 / 5491	22677	79.27
Haematology	353 / 377	5336	93.91
Endocrinology	630 / 699	7417	92.40
Endoscopy	1342 / 1484	12046	88.31
Facial	757 / 791	7686	92.13
Genital	2365 / 2552	15605	85.33
Gynaecology	2115 / 2293	14438	90.57
Infections	187 / 220	4001	94.98
Ophthalmology	2167 / 2339	13449	85.43
Oncology	5732 / 6033	25178	79.70
Otorhinology	1156 / 1227	9738	89.91
Skin	763 / 883	8442	90.43
Plastic Surgery	766 / 795	8026	92.04
Pneumology	824 / 982	9454	90.09
Rheumatology	528 / 598	7239	92.14
Senology	3644 / 3783	17497	87.99
Thoracic Surgery	1131 / 1225	10214	90.59
Vascular Surgery	900 / 959	9157	90.05

Table 5.1.: Dataset attributes for each site. The first column reports the names of the sites (disease areas), and the second one reports the number of positive reviews with respect to the total number. The third column reports the lexicon values in terms of the number of unique words, whereas the last one reports the lexicon overlap (in percentage) of each site to all the others.

5.3. Machine Learning approaches

We developed two systems based on two state-of-the-art classifiers from the state-of-the-art for sentiment analysis, Support Vector Machine (SVM) and BERT. In this Section, we present the implemented classifiers.

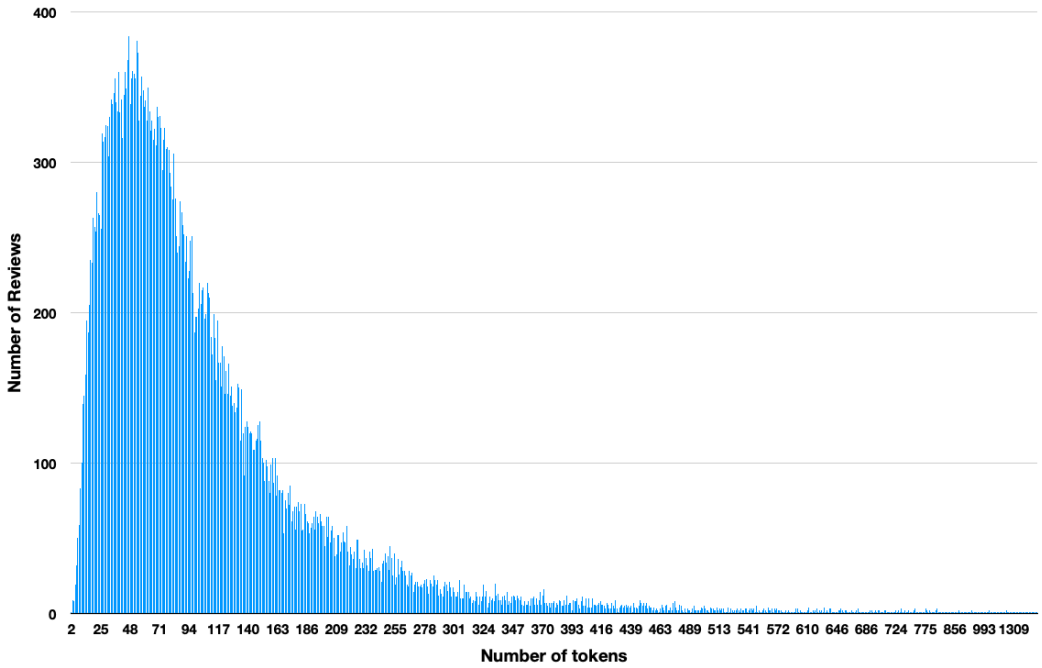

Figure 5.1.: Distribution of documents according to their length, in terms of the number of tokens. The shortest document has only two tokens, while the longest has 3571 tokens. On average, the reviews are 106.41 tokens long, with a standard deviation of 102.18 tokens.

5.3.1. SVM-based System (1)

Here, we followed the approach proposed by [206] for the sentiment analysis of English tweets, which we adapted for Italian reviews in healthcare. More precisely, we implemented a Support Vector Machine (SVM) classifier with a linear kernel, in terms of the *liblinear* [207] library rather than the *libsvm* one, to scale better to large numbers of samples, as also reported in the documentation[4] of the model employed.

Firstly, all documents pass through a pre-processing pipeline, consisting of a sentence splitter, a tokenizer, and a Part-Of-Speech (POS) tagger (all of these tools have been previously developed by the *ItaliaNLP*[5] laboratory). Then, documents pass through a step of feature extraction.

Feature Extraction All features were chosen due to their effectiveness shown in several tasks for sentiment classification for Italian [208]. We refer to these features under the

[4] www.scikit-learn.org/stable/modules/generated/sklearn.svm.LinearSVC.html
[5] www.italianlp.it

name of either *hand-crafted* and *embedding* features.

Raw and Lexical Text Features

- **(Uncased) Word *n*-grams**: presence or absence of contiguous sequences of n tokens in the document text, with n=*{1, 2, 3}*.
- **Lemma *n*-grams**: presence or absence of contiguous sequences of n lemmas occurring in the document text, with n=*{1, 2, 3}*.
- **Character *n*-grams**: presence or absence of contiguous sequences of n characters occurring in the document text, with n=*{2, 3, 4, 5}*.
- **Number of tokens**: total number of tokens of the document.
- **Number of sentences**: total number of sentences of the document.

Morpho-syntactic Features

- **Coarse-grained Part-Of-Speech *n*-grams**: presence or absence of contiguous sequences of n grammatical categories, with n=*{1, 2, 3}*.
- **Fine-grained Part-Of-Speech *n*-grams**: presence or absence of contiguous sequences of n (fine-grained) grammatical categories, with n=*{1, 2, 3}*.

Word Embeddings Combination: this set of features consists of three vectors. Each vector was calculated by the mean over word embeddings belonging to a specific fine-grained grammatical category: adjectives (excluding possessive adjectives), nouns (excluding abbreviations), and verbs (excluding modal and auxiliary verbs). We used word embeddings of 128 dimensions, extracted from a corpus of more than 46 million tweets, already used in [209] and available for download at the *ItaliaNLP*[6] website. Furthermore, we added three features to indicate the absence of word embeddings belonging to such categories, for a total of 387 $(128 * 3 + 3)$ features.

5.3.2. BERT-based System (2)

We implemented a multilingual version of Bidirectional Encoder Representations from Transformers, better known as BERT, to classify the sentiment of the reviews. BERT is a pre-trained language model developed in [23] at Google AI Language. Pre-trained

[6] www.italianlp.it/resources/italian-word-embeddings

BERT (available at its GitHub page[7]) may be fine-tuned on a specific NLP task in a specific domain, such as the sentiment analysis for reviews in the healthcare domain. Befor that, we tokenized the original text with its tokenizer.

5.4. Experiments

We conducted two types of experiments. In the first one, we wanted to evaluate which of the systems was the best. For each configuration, we trained and tested the system using a stratified k-fold cross-validation with $k = 5$. At each iteration the model was trained using four of the folds and evaluated with the remaining one. In the second part, we wanted to evaluate the robustness of the best system in a context out-domain, dividing the folders by disease sites. In these ways, in both kinds of experiments, we ensured that no information from the test sets flew through the training sets, thus avoiding any sort of over estimation of the performance.

5.4.1. System 1

We tested three different configurations of our SVM-based system, depending on the sets of features used in the experiment: only hand-crafted features (more than 626 thousand features), only embeddings (387), and a combination of both. The features that have shown to not bring improvements to the performance (numbers of tokens and sentences), or even to lower it (fg-POS n-grams, Lemmas n-grams with *n={2, 3}*) during a preliminary experimental phase were excluded from the hand-crafted features set. Thus, it turns out that such a set is composed only of Uncased Word and cg-POS *n*-grams with *n={1, 2, 3}*, in addition to Lemmas. To reduce the dimensionality of the set but also to improve the performance of our system, the features pass through a step of filtering: we assumed each one appearing less than a certain threshold *th* within the training set to be not relevant and was, thus, discarded. After searching for the optimal value during the preliminary experimental phase, we set the threshold equal to 1 (*th=1*), which means a token is retained whether it appeared more than once in the training set.

[7] www.github.com/google-research/bert

5.4.2. System 2

We conducted the experiments with BERT using the same partition into the *5* folds used during the experiments with the SVM-based classifier. This division allowed us to compare the results achieved by the two classifiers. The BERT model used in our experiments is the multilingual cased pre-trained one.

We tested two different approaches. These experiments have followed two pipelines. In the former, we fine-tuned the model with folds from the original dataset described in Section 5.2. In the second one, each fold was obtained by oversampling the minority class (i.e., the negative one) in the original fold. The oversampling was obtained by multiplying each negative sample in the fold by a factor of 4. It increased the ratio of negative out of the positive samples from about 1:16 to about 1:4. We conducted other experiments by further increasing the ratio to about 1:2 without significant improvements in performance at the expense of computational time for training. For both approaches, the model was fine-tuned for five epochs on a 12 GB *NVIDIA* GPU with *Cuda 9.0* with the following hyperparameters:

- maximum sequence length of 128 tokens,
- batch size of 24 samples,
- and a learning rate of $5 * 10^{-5}$.

The maximum length and the batch size were tuned with a light preliminary phase. For the former, in particular, the number of tokens seems reasonable since it is close to the average length of the documents in the dataset while allowing to retain the vast majority of documents without truncation (see Figure 5.1).

5.5. Results

Table 5.2 resumes the results of the experiments in stratified 5-fold cross-validation in terms of the macro average of *F1-score*. After analyzing these results, we took the best model for the leave-one-site-out cross-validation experiments to test its reliability in an out-domain (site) problem. Table 5.3 resumes these results.

First, we can notice that such performances are much higher than the baseline system, i.e., the performance achieved by a hypothetical model that classifies all the samples as belonging to the majority class (that is, the positive class). Due to the strike dataset

	$\mathbf{F1_{(1)}}$ (%)	$\mathbf{F1_{(-1)}}$(%)	**F1 (%)**
SVM			
Hand-crafted	98.90 ± 0.07	82.73 ± 1.03	90.81 ± 0.55
aEmbeddings	96.16 ± 0.15	62.37 ± 0.74	79.27 ± 0.44
Both	**98.94 ± 0.04**	**83.47 ± 0.72**	**91.21 ± 0.48**
BERT			
w/o oversampling	/	/	/
w/ oversampling	98.60 ± 0.04	77.56 ± 0.81	88.08 ± 0.42
Baseline	96.80	0.00	48.40

Table 5.2.: Results of the experiments in the stratified 5-fold cross-validation. Performances are reported in terms of F1-score (%) on each class and the (macro) average between the two. The best results are shown in bold.

imbalance and the small batch size, training BERT without oversampling the dataset leads the system to classify all samples as belonging to the majority class, i.e., the positive class. It leads to often obtaining bad performance, i.e., the baseline performance. Oversampling the minority class has shown to partially cope with such problems, leading to an improvement in terms of repeatability and performance.

For what concerns the experiments with the SVM-based system, hand-crafted features have greater relevance for the task than the embedding features. However, the resulting best model is the one with both sets of features, outperforming the BERT-based system best configuration by about three percentage points.

Given the high degree of overlap of the lexicons between domains and a larger training set, the leave-one-site-out experiments with this model result in very good performance, showing the system to be reliable in an out-domain (site) context.

In addition to the two main phases of experiments, we further investigated the confidence of the best model developed in making decisions. The motivation behind this study is that it may have application in real-world cases, where an automated system is required to filter the documents on which it is highly confident (i.e., above a certain threshold) and then passes the most complex documents to a human operator. To do so, we applied the Platt scaling [210] method on top of the trained SVM model. This step is needed to convert the output of the model from a decision score $d \in (-\infty, +\infty)$, i.e., the distance of the test sample from the trained boundary, to a probabilistic score $p \in [0, 1]$, representative of the system confidence in making the decision. Figure 5.2 resumes the results of this analysis. As expected, the number of documents on which

Site	F1 (%)	$\mathbf{F1_{Baseline}}$ (%)
Nervous System	89.91	48.52
Hearth	90.20	49.10
Haematology	91.10	48.36
Endocrinology	87.79	47.40
Endoscopy	94.34	47.49
Facial	88.31	48.90
Genital	92.12	48.10
Gynaecology	93.64	47.98
Infections	91.09	45.95
Ophthalmology	90.74	48.09
Oncology	90.85	48.72
Otorhinology	89.56	48.51
Skin	93.86	46.35
Plastic Surgery	93.63	49.07
Pneumology	92.76	45.63
Rheumatology	90.75	46.90
Senology	91.29	49.06
Thoracic Surgery	92.62	48.01
Vascular Surgery	90.01	48.41
Average	**91.24**	**47.92**

Table 5.3.: Results of the experiments in leave-one-site-out cross-validation. The first column shows the site used for testing, while the next two columns are the values of performance and baseline in terms of the (macro) average of *F1-score* of each test set.

the system makes a decision falls as the confidence threshold required of the system increases. However, this trend does not have such a negative slope and still classifies more than 91% of the documents with 99% confidence. At the same time, the performance advantage is clear, leading to an increase of the F1-score on negative samples by more than ten percentage points.

5.6. Discussion

As a Natural Language Understanding task in healthcare, we tackled the sentiment analysis of Italian reviews. For the best of our knowledge, we are the first ones facing SA in healthcare for the Italian, which led us to build the first dataset of this kind. Despite the striking imbalance of such a dataset, we have obtained very good results, especially with the SVM-based system, which outperformed the BERT-based one while

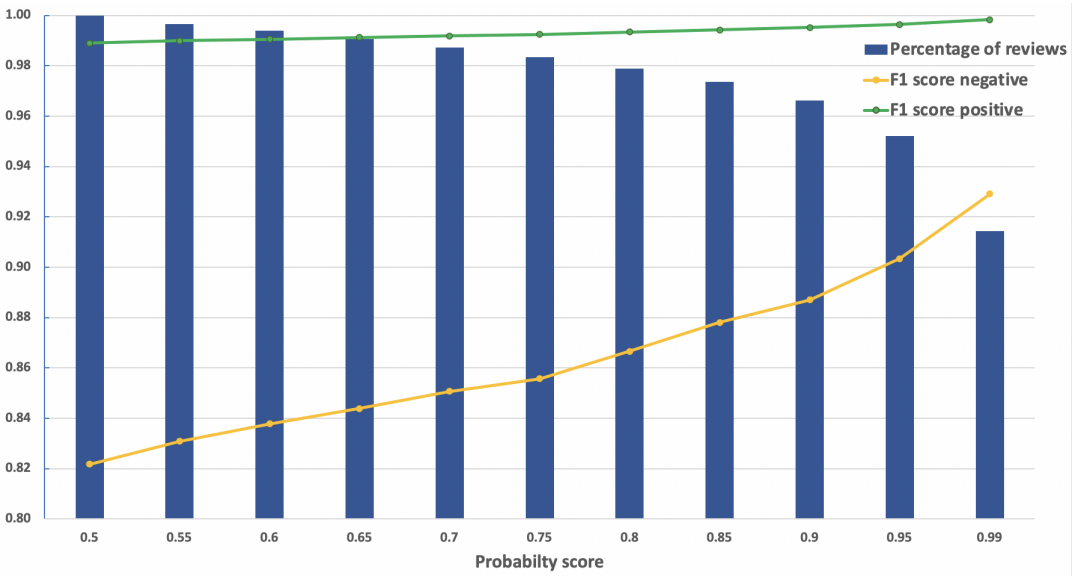

Figure 5.2.: Results in terms of percentage of classified reviews and F1-score over threshold values on the probabilistic score $p \in [0, 1]$ returned by the Platt scaling method applied on top of the SVM-based system. All the results refer to the k-fold cross-validation (with $k = 5$) fashion. Note that for $threshold = 0.5$, even if the percentage of classified documents is 100%, the value of the macro average of the F1-score is lower than the one reported in Table 5.2. tI is due to the inherent inconsistency between the probabilities calculated through the Platt scaling method p and the decisive score of the SVM model (i.e., the distance of the sample from the trained boundary, $d \in (-\infty, +\infty)$).

maintaining a low computational burden during training. To achieve competitive results with the BERT-based system, we had to perform oversampling strategies on the training set, increasing, even more, the computational costs in training. These results show the difficulties of BERT handling unbalanced datasets and that classical NLP pipelines are still capable to capture more useful information for certain types of tasks and domains. At the time when the study was conducted, Nozza et al. [211] analyzed the contribution of language-specific models, showing general improvements over multilingual BERT for a wide variety of NLP tasks. Using specific models for Italian, such as GilBERTo[8], UmBERTo[9], and AlBERTo[10]can contribute increasing the BERT-based system performance. The latter, in particular, was already used for a sentiment classification task [212]. It was also employed in a recent work that followed ours (see next section). Furthermore, even if the sentiment detection may not particularly rely on the

[8] www.github.com/idb-ita/GilBERTo
[9] www.github.com/musixmatchresearch/umberto
[10] www.github.com/marcopoli/AlBERTo-it

domain lexicon, using domain-adapted models in healthcare could improve the performance. However, at the present day, no biomedical or clinical models for the Italian language have been proposed in the literature.

Anyway, with our work, we developed a system that has resulted reliable enough to be already employed in real-world practice, especially when taking into account the confidence of the classification. Future works in this direction may focus on collecting more documents, especially for the minority, negative class.

5.7. Subsequent Works

The task we presented here was tackled by other subsequent works. Ranaldi et al. [213] used the dataset we provided, managing it following our indications. They implemented KERMIT for HealthCare ($KERMIT_{HC}$), based on the KERMIT model [214], the complementing architecture for Transformers that explicitly encodes syntactic interpretations. Compared to other BERT-based models pre-trained on the Italian language, such as AlBERTo [212], they demonstrated that encoding the syntactic components usually leads to better performance. Plus, the KERMIT-viz [215] visualizer helps to interpret the internal decision-making mechanism. However, their approach still presents macro-averaged F1-scores far below our SVM-based system. Again, such an issue could be due to the high imbalance: the authors have not stated to manage by implementing any resampling strategy.

Furthermore, motivated by the high computational costs of modern neural language models such as BERT, Martinis et al. [216] recently exploited a subset of our data to evaluate their rule-based approach *VADER-IT*, proposing the Italian version of the lexicon-based algorithm *VADER* [217]. With their method, they achieved a micro-averaged $F1$-score of 81%.

6. Reducing the Expertise Gap for Patients

To communicate information to each other, humans use natural language in speech or text. However, the information is not only conveyed by means of the content itself, but also by the linguistic form in which the semantics is carried, such as formal/informal or speech/writing attributes. Such attributes, also referable as *style*, are used to highlight the intent of the writer (e.g., politeness) or reveal their characteristics (e.g., gender). For example, to be perceived as more professional by our interlocutor, we prefer to exploit a more formal lexicon than the one we rely on in daily life (i.e., formality). Or, if we are trying to explain a difficult concept to someone unfamiliar with the subject, we tend to use a more understandable vocabulary and sentence construction rules (i.e., simplicity).

As a Natural Language Generation task, we faced *Text Style Transfer* (TST). TST concerns the rewriting of a source text by preserving its content while changing its style (target text). Among this panorama, it may help in reducing the so-called *curse of knowledge* [140] in the medical domain between physicians (experts) and their patients (laymen), which is a well-known cognitive problem leading to misunderstandings and, therefore, potential mistakes in treatments [218, 219]. These issues may affect the care process in both directions of communication. On one side, patients may find it hard to understand the messages from their doctors (expert-to-layman), which may lead to non-adherence to therapies. On the other hand, doctors may struggle to provide accurate diagnoses on the basis of the patients' jargon (layman-to-expert). Thus, improving physician-patient communication is vital to enhance the outcomes of the healthcare process [220]. In particular, in the former case, we can refer to the task of *Text Simplification*. Furthermore, given the spreading of health-related mobile apps and social networks, it is not hard to imagine the beneficial effects of implementing systems for reducing the expertise gap in existing and future applications, overcoming the language barriers, and allowing access to health resources to the consumers [141].

Recently, Cao et al. [221] proposed the *Expertise Style Transfer* (EST) task between medical experts and laymen. They introduced a large non-parallel set of sentences (to be used for training models) and a relatively small set of parallel texts annotated by domain experts (to be used for evaluation purposes). Since non-parallel sets of text may represent sub-optimal solutions for training some model for the task at hand, we explored methods to collect parallel sets from a large, non-parallel corpus for training a generative model. We implemented BART (Bidirectional and Auto-Regressive Transformers [222]) as the generative model, which has already been proven to be suitable for style transfer tasks [223, 224]. To collect parallel sets from the original EST training set, we exploited both datasets and models from the literature concerning the (clinical) *Semantic Textual Similarity*. In particular, we implemented *Sentence Transformers* [225] models, bi-encoder Transformers particularly suitable for the task of similarity search required to collect the parallel datasets. Also, we evaluated both the collected parallel sets and the style transferred texts not only with automatic but also with manual annotations performed by domain lay and experts, from both content preservation and degree of style changes points of view. Our main contributions may be listed as follows:

- we proposed a *Text Style Transfer* system to effectively reduce the expertise gap in the communication between physicians and patients;
- we analyzed and evaluated several *Semantic Textual Similarity* methodologies to automatically collect parallel datasets;
- we analyzed how different qualities (in terms of varying similarity) in the collected datasets affect the downstream task (*Text Style Transfer*);
- we conducted an extensive human evaluation phase with expert and lay people, highlighting issues and characteristics of datasets and models as well as evaluation metrics involved;
- we collected the experts' annotations, which may be employed in future works to deploy or evaluate *Semantic Textual Similarity* and *Text Style Transfer* systems in the medical domain.

In particular, with our approach, we have overtaken the state-of-the-art performance on the *Expertise Style Transfer* task on a variegated set of automatic metrics, computed on both the input (*self*) and the target (*ref*). However, we also argue that achieving higher performance in terms of automatic metrics does not necessarily imply better accomplishing the task. Our system, anyway, outperformed state-of-the-art models based

on both human experts' and laymen's judgments.

In this chapter, we present our work as follows. First, we report the investigated literature in the context of TST (the expertise one, in particular) and then the datasets we exploited for both *Semantic Textual Similarity* and *Expertise Style Transfer*. After that, we describe in detail the methods implemented to tackle the *Expertise Style Transfer* task, the strategies for collecting the pseudo-parallel datasets, and the evaluation protocols we adopted to assess their qualities. Finally, exploiting plots and tables, both automatic and human results regarding content preservation, style strength, and fluency of the models' outputs and/or the collected parallel training sets are discussed.

6.1. Related works

Style Transfer (Neural ST, in particular) is the task of reproducing some input content in a different style. Researchers investigated it for several media, from images and videos [226, 227, 228, 229] to music [230, 231]. Some of the developed systems have already seen their application in industrial solutions[1]. *Text Style Transfer* (TST) shares the same principle as the other media: rewriting some textual input with a different attribute while minimizing the information loss. Researchers have investigated TST for various attributes such as formality [232, 223], politeness [233, 234], and sentiment [235]. Past works in TST have focused on these attributes as the related resources are easier to obtain. The sentiment Style Transfer, commonly known as polarity swap, has been questioned as a TST task [224] as it does not preserve the original meaning of the source text, i.e., a positive sentence is changed to negative and vice-versa. However, in order to get a more comprehensive overview of TST tasks, we refer the reader to recent reviews [236, 237].

Existing TST approaches can be grouped into three main categories, i.e., disentanglement, manipulation, and translation:

- *Disentanglement* methods attempt to learn separate representations for content and style [235, 238, 239], so that one can be manipulated without affecting the other. However, the success of disentanglement is difficult to assess, and some studies have shown that the latent representations may not actually be disentangled, being possible to recover information of style from the other [240, 241].

[1] https://prisma-ai.com/; https://www.pikazoapp.com/; https://deepart.io/; https://groove2groove.telecom-paris.fr/

- *Manipulation* methods work by identifying specific words in the text that contribute to its style, such as professional language or clinical abbreviations (e.g., *qd*), and replacing them with synonyms or explanations (e.g., *once per day*) that are more appropriate for lay people [242, 243]. In the biomedical and clinical domain [244, 245], these methods often use *Consumer Health Vocabularies* [141, 246, 247, CHVs]. Weng et al. [245], in particular, used CHVs as a preliminary step to align embedding spaces and then used a translation-based technique to generate simplified sentences.

- *Translation* methods often use unsupervised training to learn style-specific translations [241] with back-translation or cycle reconstruction strategies. Back-translation [248] involves translating the source text to another language and back again. Pabrhumoye et al. [249] proposed it on the basis of the evidence shown by Rabinovich et al. [250] to reduce the style properties of the source text. Such strategy has already shown its efficiency [224] Cycle reconstruction, instead, involves training a model to reconstruct the source text from the transferred output [251, 252]. Parallel corpora can also be used for supervised training, but they can be expensive and time-consuming to collect.

For the *Expertise Style Transfer* task at hand, Cao et al. [221] evaluated models belonging to the three macro-categories of TST discussed earlier (see also Sec. 8.3 for an overview), while our approach falls into the latter category. In particular, we exploited the collection of pseudo-parallel corpora, built on the basis of a definition of similarity criterion between sentences, which has shown advantages over unsupervised training [253], while being cost-effective if compared with the collection of human-annotated corpora.

In one related work, Luo et al. [254] collected gold corpora from *MIMIC-III* database [255], which was a time-intensive process requiring a certain degree of expertise. To overcome this issue, like us, Xu et al. [256] collected a large, pseudo-parallel corpus from the MSD training set. While sharing the same intent to collect pseudo-parallel corpora, there are some crucial differences. They used a language- and topic-agnostic LASER [257] framework to extract the embeddings and collected the largest number of training pairs above a fixed threshold on their similarity criterion. Our approach differs in the use of general and domain-specific monolingual Transformer-based models and in the investigation of the impact of different threshold ranges on the final TST system.

Disposing of parallel corpora can be an effective solution for these issues [253]. When such data is not available, the automatic collection of pseudo-parallel data has proven

to be effective, including in neural machine translation tasks [258, 259]. Style transfer, similarly to machine translation, is a rewriting task and shares similar modeling approaches. While machine translation deals with cross-language content, style transfer is typically within the same language. In some cases, it is approached from a multilingual perspective [260].

However, the collection of high-quality parallel datasets is a challenging task, especially in a specialized domain like healthcare, where human efforts and costs are significant. To address this issue, van den Bercken et al. [261] proposed using the BLEU score [262] to automatically collect a parallel dataset for a medical simplification task by utilizing texts from Wikipedia and Simple Wikipedia. However, this approach was found to be unsuitable for our use case due to the presence of many texts in both the expert and lay training corpora, and the significant differences between the expert and lay test samples. Another common technique for collecting parallel datasets is to train a classifier that can distinguish sentence pairs from two different corpora [263, 264]. However, using a classifier for large corpora is often infeasible, especially when using Transformer architectures. Our approach addresses these limitations by employing bi-encoders. To the best of our knowledge, the use of bi-encoders in style transfer tasks, particularly in the technical domain of medicine, has not been explored previously.

Furthermore, Xu et al. [256] focused mainly on human evaluation and compared their outputs only with inputs using (self-)BLEU, ignoring reference sentences in their analysis (ref-BLEU). The interpretation of high self-BLEU scores is not trivial: a score close to 100% between input and output only means that the model has learned to reproduce the input without making any changes to the style. Moreover, it has been established that surface-based metrics like BLEU are not ideal for TST tasks, as they exhibit low correlation with human judgments [265, 266]. For these reasons, we evaluated our outputs and those of the models presented by [221] using several other metrics, referred to both input and target sentences. Computing these metrics for the gold source and target texts allowed us to highlight the degree of content changes in the test set, as suggested in previous works [221, 267, 268], and confirmed through our human evaluations. This issue may stem from the loss of contextual information when working at a sentence level. As a result, a few studies have taken a paragraph-level approach to the medical-style transfer task from the perspective of Plain Language Summarization (PLS [269, 270]), also in languages other than English [271].

Our study makes a unique and significant contribution to the field of Text Style Trans-

fer by presenting an extensive examination of the collection of parallel data and offering unique insights specific to its application. Furthermore, our human expert evaluations set our work apart from previous studies, providing a valuable and rare perspective on the performance and quality of our system. Despite building upon previous works, our approach offers a fresh and innovative perspective on the task of Text Style Transfer. The combination of automated and human evaluations, coupled with the in-depth analysis of parallel data collection, makes our study a valuable addition to the current state-of-the-art.

6.2. Datasets

In this work, we exploited and combined three datasets for the similarity and the style transfer tasks.

ClinicalSTS2019 (CSTS) Wang et al. [272] collected a total of 2054 pairs annotated by two clinical experts for the track on *Clinical Semantic Textual Similarity* in the *n2c2/OHNLP* challenge of 2019. The training set, in particular, is an extension of the dataset presented in the previous year's challenge [273]. For the annotation phase, the authors asked the experts to independently annotate each pair, on the basis of their semantic equivalence, on a scale from 0 to 5, where 0 indicates a completely dissimilar sentence (i.e., no overlap in their meanings) and 5 indicates a perfect semantic match. We refer the reader to Table 6.3 and to [274] for more in-depth analysis and data examples.

Medical Question Pairs (MQP) McCreery et al. [275] collected a dataset of 1524 of random *COVID-19* related questions[2]. For each question, a doctor provided one positive and one negative example. For the former, doctors have rewritten the original question by restructuring it as much as possible while maintaining the content. In the latter one, doctors have rewritten the original question in a manner that the answer of that question would be resulted being wrong/irrelevant while maintaining the same structure and keywords. In this way, positive question pairs can look very different, whereas negative question pairs can conversely look very similar at the surface. Each pair is then labeled as either similar or dissimilar (1 or 0).

[2] https://huggingface.co/datasets/medical_questions_pairs

MSD Cao et al. [221] collected data from human-written medical references from the *Merck Manuals* (also known as the *MSD Manuals*) website[3], one of most world-widely trusted reference in health. They collected more than 245k non-parallel sentences in expert and layman styles (∼130k and ∼115k sentences, respectively). Also, they hired three doctors who annotated a total of 675 pairs of parallel texts. This additional set may be too small to train a system but can be used to evaluate the generative models. Furthermore, the authors provided a list of related concepts and terms related to each sample, obtained with the QuickUMLS [276] tool to link medical entities in the text to the Unified Medical Language System (UMLS) concepts [277].

We performed an empirical analysis of the parallel set provided as the test set. Among all the pairs, we found various samples sharing some problematic patterns that can compromise the evaluation of the models. Some pairs report the same text for both styles, a poor fluency for one or both texts, or information not mentioned in the counterpart text (and thus unlikely to be reproduced in the transfer task). Also, we encountered different gold target references for the analogous (or even the same) source texts. Furthermore, sometimes domain knowledge seems to be essential for the model to perform well, e.g., in the case of acronyms. Even if this case cannot be marked as an "erroneous" sample, it can help understand the difficulties of the task. Another serious problem is the presence of different meanings in the two texts of the test pair. Table 6.1 reports an overview of some examples. These problematic patterns can compromise the evaluation of the models, and a more thorough evaluation is needed to better understand the difficulties of the medical style transfer task.

Regarding the training dataset, we discovered that there were overlapping texts in both styles, particularly in instances of fixed word patterns. We regarded these instances as irrelevant and filtered them out by removing sentences that were short (less than 10 tokens) or displayed specific patterns using simple regular expressions. This preprocessing stage reduced the number of samples to approximately 110k for the expert style and 97k for the layman style.

6.3. Text Style Transfer System

From a mathematical standpoint, the aim of a TST system is to model the probability $p(y|x)$ where $x(c, a)$ is the source sentence and $y(c, b)$ is the target sentence, with the

[3] https://www.msdmanuals.com/

	Expert	Layman
(i)	The change in LDL levels may partly explain why atherosclerosis and thus coronary artery disease become more common among women after menopause. [...]	The change in LDL levels may partly explain why atherosclerosis and thus coronary artery disease become more common among women after menopause. [...]
(ii)	Treatment of underlying disorder	Treatment of cause
(iii)	The most common causative organisms of occult bacteremia are Streptococcus pneumoniae and Haemophilus influenzae. [...]	Children **under 3 years old** who develop a fever **(particularly if their temperature is 102.2°F [39°C] or higher)** sometimes have bacteria in their bloodstream (bacteremia). [...]
(iv)	Clinical evaluation Clinical evaluation Clinical evaluation.	Physical examination A doctor's evaluation A doctor's examination.
(v)	**IV** fluids.	Fluids given by vein
(vi)	[...] **It occurs predominantly in men** practicing receptive anal intercourse and can occur in women who participate in anal sex.	**It occurs mainly in women**. Anal sex with an infected partner may result in gonorrhea of the rectum.

Table 6.1.: The analysis of the *MSD* test dataset has revealed some problematic pairs. Most of them belong to one of the following patterns: (i) duplicate texts for both styles, (ii) poor fluency, (iii) missing information, (iv) different gold target references for the same source text, (v) acronyms, and (vi) different meanings between source and target texts. The truncated texts are indicated with "[...]" to accommodate them in the table.

same content c but different attributes (styles) a and b. If the system can also model the reverse direction $p(x|y)$, it is referred to as bidirectional [236]. If the transformation is from a more complex source text to a simpler one, such as from expert to layman style, it is also referred to as *Text Simplification*.

For our system, we exploited the collected pseudo-parallel training sets (Sec. 6.3.1) to fine-tune a Bidirectional and Auto-Regressive Transformers (BART) model [222]. *BART* is a denoising autoencoder for pre-training sequence-to-sequence model. Given a source sentence $\boldsymbol{x} = \{x_1, \cdots, x_n\}$ and a target sentence $\boldsymbol{y} = \{y_1, \cdots, y_m\}$, its loss function is the cross-entropy between the decoder's output and the target sentence:

$$L_(\phi) = -\Sigma_i \log(p(y_i|y_{1:i-1}, \boldsymbol{x}; \phi)) \tag{6.1}$$

The entire system pipeline is depicted in Figure 6.1 and consists of four steps.

i) Initializing the *Sentence Transformers* with the pre-trained *BERT*-based models' weights.

ii) Fine-tuning the Sentence Transformers with the datasets described in Sec. 6.2.

iii) Using the bi-encoders to perform a similarity search on the expert and layman corpora from the MSD training data.

iv) Fine-tuning the *BART* model for the Text Style Transfer task using pseudo-parallel data collected by setting a similarity threshold.

The resulting model can then be used during inference to simplify medical texts for a lay audience.

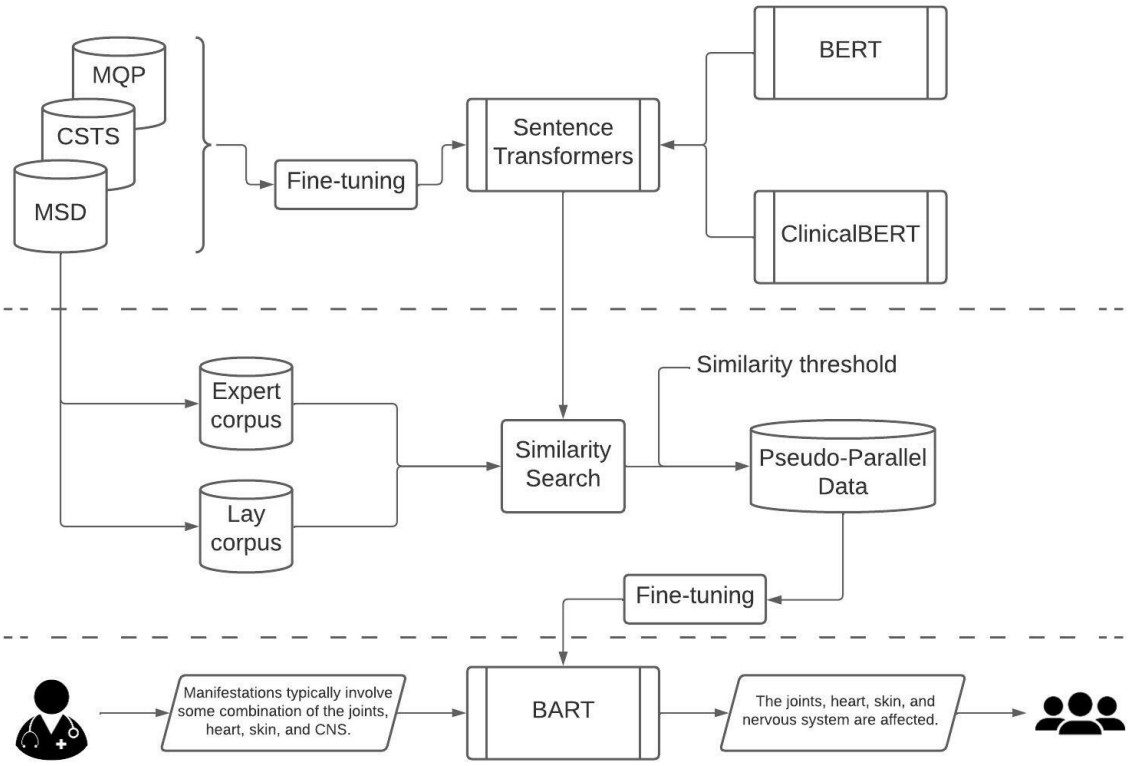

Figure 6.1.: Our approach consists of the following pipeline: (i) retrieving pre-trained Transformers as a bi-encoder and (ii) fine-tuning them with *Semantic Textual Similarity* datasets or *MSD* training set; then, (iii) using the fine-tuned bi-encoder to perform a similarity search on the expert and layman corpora derived from the MSD training set. By setting a similarity threshold to collect pseudo-parallel data, (iv) fine-tuning the style transfer model using the collected pseudo-parallel data. In the end, the fine-tuned model is used during inference time to simplify medical texts from physicians to patients.

6.3.1. Pseudo-parallel Data Collection

To implement our supervised approach, we first tackled the preliminary task of collecting pseudo-parallel data from the two large, expert and lay corpora, to be used for the training process of the model. To collect these parallel training sets, we implemented the *Sentence-Transformers* [225] architecture. Such architecture is a bi-encoder, a siamese network in which one Transformer encoder gets trained to produce semantically meaningful embeddings. It means that the outputs of semantically similar sentences are closer to each other in the vector space for some distance definition than dissimilar ones.

Besides being known to achieve suboptimal performance compared with a cross-encoder architecture (i.e., the classical Transformer-encoder network), using a bi-encoder is advantageous for large-scale semantic search, as the case in hand. It reduces the computational complexity of retrieving representations for each paired combination in the dataset to the task of obtaining one embedding for each sentence and computing some similarity metric between paired embeddings combination. Implementing this kind of technique allowed us to conduct a similarity search through *FAISS* [278], based on the cosine similarity between paired embeddings, to retrieve with a GPU-optimized strategy the nearest layman neighbor of each expert sample.

Pre-Trained Models We evaluated several Transformers encoders to collect the sentence embeddings of the training dataset. Since we were interested in analyzing the behaviors of a domain-specific encoder with respect to a general-topic one, we chose chose *BERT* [23] and *(Bio-)ClinicalBert* [58]. We considered these two models because they share the same architecture, thus excluding influences derived from different architectures. The latter was initialized starting from the former and then pre-trained on large medical and clinical domain data. As expected, it led to better similarity performance (Table 6.2). Thus, we retrieved this model to conduct our training strategies.

In a preliminary phase, we evaluated the *all-mpnet-base-v2*[4] model, which is the best model reported in Sentence Transformers specifically trained for producing semantically significant sentence embeddings[5]. It performed better than both models before training in terms of the performance on the CSTS test set (see paragraph "*STS Evaluation*"). However, after training, it was outperformed by the *(Bio-)ClinicalBert* and thus discarded for further analyses.

[4] https://huggingface.co/sentence-transformers/all-mpnet-base-v2
[5] https://www.sbert.net/docs/pretrained_models.html

Training Strategies We implemented the *Multiple Negatives Ranking (MNR) loss* [279] as the loss for the *contrastive (representation) learning* [280]. It pushes the model to create closer representations in the vector space for similar sentences and more distant for dissimilar ones, based on some distance/similarity metric. At each step, the training process aims to minimize the following equation:

$$L_{MNR} = -\frac{1}{K}\sum_{i=1}^{K}[s(x_i, y_i) - log\sum_{j=1}^{K} e^{s(x_i, y_j)}] \tag{6.2}$$

in which (x_i, y_i) indicates any i-th anchor-positive (premise and hypothesis) pair and (x_i, y_j) indicates any anchor-negative pair in the batch of size K; $s(.,.)$, instead, indicates the score based on the defined metric (cosine similarity in our case).

We trained our models on *MQP* and *CSTS* training datasets, in some cases exploiting only the positive pairs during the training process. For the latter, we considered positive the pairs with a semantic equivalence score greater or equal to 4. For the model trained on the *MSD* training dataset, not disposing of any content equivalence labels, we exploited the strategy proposed in [281] for the unsupervised *SimCSE* (Similarity Contrastive Sentence Embedding) framework. We used the Transformer encoder with anchor-positive pairs consisting of the same input sentence. Due to the randomness of the *dropout* [282, 283] masks in the encoder's layers, the model generates two different (noisy) representations of the same sentence, and the model learns to generate closer embeddings from the noisy representations while distancing anchor-negatives pairs in the batch.

STS Evaluation To be consistent with past literature, we evaluated the models using two common metrics for semantic textual similarity, Pearson and Spearman correlations between the similarity scores $\boldsymbol{x} = \{x_1, \ldots, x_n\}$ produced by the sentence embeddings and the *CSTS* official test set labels $\boldsymbol{y} = \{y_1, \ldots, y_n\}$. Equation 6.3 reports the formulas of these metrics,

$$pearson = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \overline{x})^2(y_i - \overline{y})^2}}; \quad spearman = 1 - \frac{6\sum d_i^2}{n(n^2-1)} \tag{6.3}$$

where $\overline{x}$ and $\overline{y}$ indicate the mean of vectors $\boldsymbol{x}$ and $\boldsymbol{y}$, respectively, n is the number of elements, and d_i the pairwise distance of the ranks of the i-th elements (x_i and y_i). In particular, we defined the score x_i between two sentences $\boldsymbol{a_i}$ and $\boldsymbol{b_i}$ as the cosine

similarity of their embeddings, as reported in Equation 6.4.

$$cos(\boldsymbol{a_i}, \boldsymbol{b_i}) = \frac{\boldsymbol{a_i} \cdot \boldsymbol{b_i}}{||\boldsymbol{a_i}|| \cdot ||\boldsymbol{b_i}||} \tag{6.4}$$

In addition, we also assessed the models' performance by calculating the average cosine similarity between expert-layman pairs ($\boldsymbol{a_i}$ and $\boldsymbol{b_i}$) in the *MSD* test set as

$$similarity = \frac{1}{N} \sum_{i}^{N} cos(\boldsymbol{a_i}, \boldsymbol{b_i}) \tag{6.5}$$

where N is the number of pairs. Table 6.2 reports the results of these evaluations. As previously mentioned, *(Bio-)ClinicalBert* outperformed the other pre-trained model, *Bert* (**cb** and **bert** in the table), for all the metrics. This was expected since the former passed through a pre-training (domain adaptation) phase in the biomedical and clinical domains. The **cb__mqp__csts1** model achieved the highest correlation scores. It is a *(Bio-)ClinicalBert* we first fine-tuned on the *MQP* dataset and then on the positive samples of the *CSTS* dataset. Interestingly, the second fine-tuning step only slightly improved the performance (see **cb__mqp**). Apart from the pre-trained models, for what concerns the evaluation on the *MSD* test set, the model fine-tuned only on the *CSTS* positive samples (**cb__csts1**) achieved the highest averaged cosine similarity. The model fine-tuned using only the *MSD* data, instead, performed poorly on the averaged cosine similarity. Such a result may indicate that the training strategy employed for this model was not suitable for the test set at hand, which presents a high degree of aggressiveness in the changes between source and related target texts.

Datasets Creation As the last step, we collected the pseudo-parallel datasets. We conducted the next analyses with a selected set of the implemented models. Based on their performances, we retrieved the pairs collected using **cb__mqp__csts1** and **cb__csts1** models. To analyze the impact of the fine-tuning strategies, as well as the pre-training domain adaptation step, we included the **cb__msd**, **cb** and **bert** models, too. Furthermore, we also computed the similarity search between the lay corpus and a "corrected" expert one, for which we switched expert terms with their lay-related terms. To do so, following a similar approach of Xu et al. [256], we first collect all the *Concept Unique Identifier* (CUI) codes in the *MSD* training set, as well as the number of occurrences of the terms appearing in the texts for each style. Then, we switched each expert term with the most represented one in the lay texts that share the same CUI(s). From now

Id	Model	Training set	Pearson (%)	Spearman (%)	Similarity (%)
bert	Bert-base-uncased	/	21.64	25.03	87.70
cb	Bio-ClinicalBert	/	30.07	31.84	**93.99**
cb_mqp1	Bio-ClinicalBert	$mqp^{(pos)}$	68.26	71.29	69.72
cb_mqp	Bio-ClinicalBert	mqp	80.27	77.41	67.12
cb_csts1	Bio-ClinicalBert	$csts^{(pos)}$	43.91	47.44	**79.28**
cb_csts	Bio-ClinicalBert	csts	61.61	56.08	66.33
cb_mqp_csts1	cb_mqp	$csts^{(pos)}$	**81.12**	**78.29**	69.93
cb_mqp_csts	cb_mqp	csts	66.51	62.33	65.17
cb_msd	Bio-ClinicalBert	msd	53.22	53.67	47.93

Table 6.2.: Semantic Textual Similarity model performance. Each model identified by the first column got fine-tuned starting from a pre-trained model on a specific training set: mqp (medical question pairs), csts (clinicalSTS2019), and msd, where $^{(pos)}$ indicates that only the positive pairs were included in the training process. The first two rows report basic pre-trained models without a further training phase. We evaluated the performance in terms of Pearson and Spearman correlation coefficients computed on the clinicalSTS2019 dataset and on the average cosine similarity computed on the parallel samples of the msd test set.

on in the paper, we refer to this dataset as **cb_msd_swap**.

To analyze the impact collected training sets may have on the final task at different similarity thresholds, we retrieved several datasets at different threshold ranges based on the quantiles they separate in the entire training set. We thus selected the following ranges between the following quantiles: {99%, 95%, 90%, 85%, 80%, 75%, 70%, 50%}. To minimize the impact of the training set size, we used the same number of samples for each interval (with the exception of the ones above 99% and between 99% and 95%, which contained a smaller number of samples). We then evaluated the overlap between the datasets collected by the several models by averaging their overlap at each quantile. As shown in Figure 6.2, datasets were more or less dissimilar, on average. The pre-trained models are more similar to each other than to the fine-tuned ones. Also, the two datasets collected with the **cb_msd** model look mostly overlap.

6.4. Automatic and Human Evaluation

We evaluated the collected training sets and the outputs of the style transfer systems through both automatic metrics and manual annotation. Each metric refers to one of the text's aspects, i.e., the style strength (degree of style transfer) of the target text, the content preservation between source and target texts, and the fluency of the

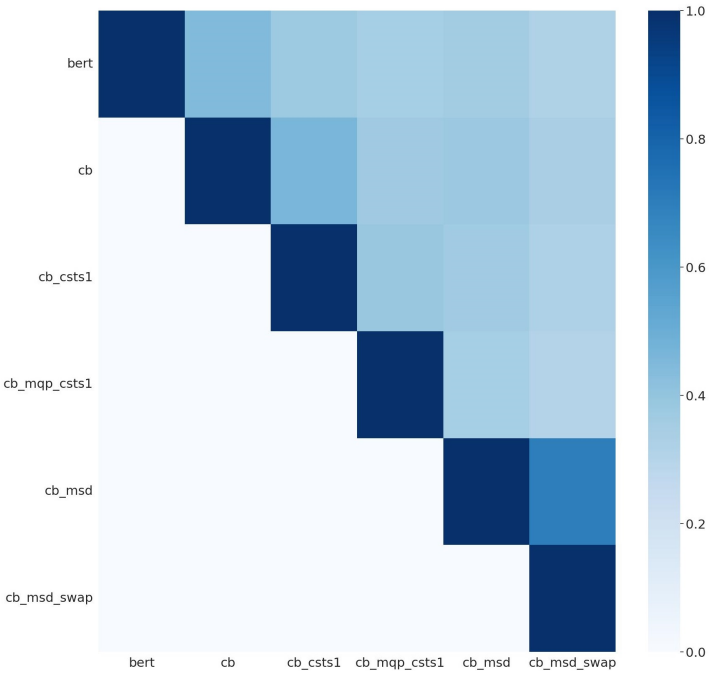

Figure 6.2.: Average overlaps between collected datasets.

generated text. In particular, we compared our results with some baseline models, i.e., an *unsupervised BART* model (having the same architecture as in our system) and all the models provided by [221], which include two text simplification models and three style transfer models. We fine-tuned the *unsupervised BART* model without parallel data by employing an iterative back-translation approach [284]. Two models for the two transfer directions get trained (almost) simultaneously. Specifically, each model generates synthetic parallel data for the other. In this way, the models get trained in a pseudo-supervised fashion, each in one direction.

The baselines provided by Cao et al. [221] include:

- *OpenNMT+PT* [243], an OpenNMT-based [285] supervised model that replaces complex words with their simple synonym based on a phrase table;
- *UNTS* [286], an unsupervised neural model consisting of a shared encoder and a pair of attentional decoders; it is trained with discrimination-based losses and denoising;
- *ControlledGen* [238], a neural generative model combining variational auto-encoders and style attribute discriminators for the effective imposition of semantic structures;

- *DeleteAndRetrieve* [242], an editing-based method that first deletes style-related words, then retrieves new phrases associated with the target attribute and uses a neural model to combine them as the final output;
- *StyleTransformer* [251], a Transformer-based model that uses cycle reconstruction to learn content and style representation without parallel data.

6.4.1. Automatic Evaluation

Following previous works [287, 288, 223, 224], we used the following strategies. To assess the **content** aspect, we computed BLEU [262] and BERTScore [289] between the generated sentence and the human source and reference. BLEU counts the n-gram matches in the candidate text with the reference one, this can be roughly formulated as

$$\text{BLEU-}n = \frac{\sum_{C \in \{Candidates\}} \sum_{\text{n-gram} \in C} Count_{match}(\text{n-gram})}{\sum_{C \in \{Candidates\}} \sum_{\text{n-gram} \in C} Count(\text{n-gram})} \tag{6.6}$$

where C represents the candidate text and *match* means that a n-gram appears in both the candidate and either the source (*self-BLEU*) or the reference (*ref-BLEU*). In particular, we used the *overall* BLEU by averaging the scores obtained with $n = \{1, 2, 3, 4\}$. BERTScore uses greedy matching to maximize the matching similarity score for each token in the candidate sentence with each token in either the source (*self-BERTScore*) or the reference (*ref-BERTScore*), and combines recall (R) and precision (P) to compute an F_1 measure. This can be formulated as

$$R = \frac{1}{|x|} \sum_{x_i \in x} \max_{\hat{x}_j \in \hat{x}} \mathbf{x}_i^\top \hat{\mathbf{x}}_\mathbf{j} \,, \quad P = \frac{1}{|\hat{x}|} \sum_{\hat{x}_j \in \hat{x}} \max_{x_i \in x} \mathbf{x}_i^\top \hat{\mathbf{x}}_\mathbf{j} \,, \quad F_1 = 2\frac{R \cdot P}{R + P} \tag{6.7}$$

where $\hat{x}$ and x represent the candidate and reference, respectively. We also included two learnable metrics, BLEURT [290] and COMET [291], as they have shown promising correlation results with human judgments in the evaluation of machine translation, as well as style transfer tasks as formality [266].

For what concerns the evaluation of the **style** of texts, we used a TextCNN-based [292] classifier (trained on the entire training set) to evaluate the target style accuracy of the transferred texts.

Regarding the **fluency**, we assessed the perplexity in an analogous way as in [221],

computing a (pseudo-)perplexity with a masked language model. As in [293], for each text W^i, for each of its tokens w_t, we first computed the conditional log probability $P_{MLM}(w_t|W^i_{\setminus t})$ obtained by the model giving in input the sentence $W^i_{\setminus t}$, that is the same as W^i but with the t-th token masked. Then, we computed the pseudo-likelihood PLL^i of each text by summing the contribution for each token. Finally, we added the scores of all the corpus S of sentences together, normalized the result with respect to the total number of tokens N in the corpus, and exponentiated the result to obtain a measure of pseudo-perplexity $pPPL$. The process is summed up in the following equation:

$$pPPL(S) = exp(-\frac{1}{N}\sum_{i=1}^{|S|} PLL(W^i)) =$$
$$= exp(-\frac{1}{N}\sum_{i=1}^{|S|}\sum_{t=1}^{|W^i|} logP_{mlm}(w^i_t|W^i_{\setminus t}))$$

To compute such scores, we used *(Bio-)ClinicalBert* as well as its versions fine-tuned on the training sets for lay or expert styles. To balance the data sizes and remove any influence given by the different corpus dimensions, we reduced the number of experts' texts during fine-tuning.

6.4.2. Human Evaluation

The human evaluation was conducted with two different protocols to capture both the lay people and professional physician's perspective, thus we elaborated two different protocols. The lay people were asked to judge only the style perception of the samples, while the physicians were asked to judge both the style and the content preservation. For the evaluation of the pseudo-parallel datasets collected with the *Semantic Textual Similarity* models, only the evaluation by the physicians was conducted. The goal was to ensure that the content preservation was evaluated by experts in the field, as lay people without the right field expertise were deemed unreliable in assessing it.

Due to the high cost of hiring professional healthcare personnel, we carried out the annotations in only one direction, from expert to layman. The reason for this choice is that text simplification tasks have been more widely studied in the past and are considered more important for real-world applications. Due to cost constraints, we selected only one of our TST models for evaluation. We made this decision based on the results of our automatic evaluation (Section 6.5.1). We examined the results of the

pseudo-parallel datasets collected with respect to similarity quantiles and focused on models trained on datasets collected at the 85% quantile. This was because the parallel sets at the 85% quantile score were closer to the results obtained on the gold test set and the TST models trained on them generally showed good balance between content preservation and style evaluation. Among all the TST models trained on the pseudo-parallel sets (Sec. 6.3.1), we chose the one trained on the **cb_mqp_csts1** (at 85% quantile) set because achieving higher content preservation scores on average. We also manually inspected its outputs and the ones of the model trained on **cb_msd_swap** (at 85% quantile), which was close in terms of performance. We noted that the former tends in some cases (especially when the input sentence is relatively short) to generate explanations of medical terms. This behavior was shown by the other model, too, but with less frequency and accuracy (refer to Sec. 6.5.4 for some examples). To what concerns perplexity metrics instead, the trends were not clearly separable, thus they did not influence the final decision.

To compare our system with state-of-the-art models, we chose the **Style Transformer** as a competitor. This choice has a dual justification. Among the models previously proposed in the literature to tackle the *Expertise Style Transfer*, it showed a more stable behavior in terms of the content preservation/style strength trade-off, as also highlighted in [221]. Also, its architecture and training strategy are similar to our unsupervised model. It allowed us to show the improvement of our methodology against unsupervised ones.

We decided not to pick our unsupervised model into account for the human evaluation because of its (too) high content preservation scores with respect to the source, which indicates its outputs are created by mostly repeating the inputs. Furthermore, we added the gold lay references in the comparative analysis to assess the distance and goodness of the two models with respect to the gold references. The annotators were not aware of which text was written by which system.

From the evaluation process, we excluded samples with source texts of less than 5 and more than 32 tokens and samples for which at least one of the models presented an output that was the same as the source text. It allowed us to perform a fair comparison between models regardless of their hyperparameters (e.g., maximum input sequence length), as well as reducing the annotators' efforts while removing trivial examples.

Besides giving different annotation protocols to lay people and experts, we asked them to judge the same texts.

Annotation protocol for layman We hired ten lay people proficient in English without a background in medicine and annexed fields. We asked them to choose between two texts the easier to understand. Each pair included the source text and one of the systems-related output (or the reference text). Before being presented to the annotators, we shuffled the pairs to minimize the bias. Also, each subject annotated 30 samples (consisting of 3 pairs, one for each system). An overlap of 10 samples with another subject was present to assess the agreement between couples of annotators, resulting in 250 annotations. We assessed the agreement between the annotators with the *Cohen's Kappa* (K^{lay}) [294] and evaluated the style transfer as the ratio between the number of texts judged easier to understand than the related source text (Sty^{lay}).

Annotation protocol for expert We hired four physicians proficient in English from the Department of Orthopaedic Surgery of University Campus Bio-Medico of Rome, Italy. We divided them into two groups, one for judging the collected pseudo-parallel data and one for judging the output of style transfer systems. For both settings, we asked them to assess the content preservation. For the outputs, we also asked to evaluate the degree of style transfer, in terms of the quality of the changes made. Regarding content preservation, to be consistent with the past literature [274], we followed the same guidelines to assess the content preservation. In particular, we asked them to assess the style question without influence by the content, even if medically inaccurate, evaluated with the other question. We asked them not to take into account fluency issues as well. To assess the style strength, specifically, we asked them to take into particular account terminology and empirical evidence knowledge gaps, as also highlighted in [221]. Table 6.3 reports the questions and the answers included in the expert protocols, as well as the scores associated with each answer.

For what concerns the evaluation of the pseudo-parallel data, we presented to the annotators a total of 350 samples consisting of one expert sentence and its lay counterpart. Each sample was randomly extracted from one of the quantile-dependent sets collected with **cb_mqp_csts1**. We excluded the samples for the 99% quantile presenting pairs of the same texts. This protocol allowed us to analyze how the quality perceived by the physicians changes across the quantile ranges.

To evaluate the three systems' outputs, we presented to the annotators a total of 250 samples, consisting of one source text and three rephrased texts. For each sample, together with the source text, we presented all the outputs on the same annotation page.

	Content preservation
Q:	To what extent is the rewritten text still conveying the same content as the source text?
0:	The two texts are *completely dissimilar*.
1:	The two texts are not equivalent, but are on the *same topic*.
2:	The two texts are not equivalent, but share *some details*.
3:	The two texts are roughly equivalent, but some *important information differs/missing*.
4:	The two texts are mostly equivalent, but some *unimportant details differ*.
5:	The two texts are *completely equivalent*, as they mean the same thing.
	Style strength
Q:	To what extent the process of rewriting for a lay audience can be considered a good attempt?
0:	The rewriting process is not a good attempt, performing *no changes* from the source text.
1:	The rewriting process is not a good attempt, performing *some changes* that are *not good* for the scope.
2:	The rewriting process made some *minimal good changes* but the rewritten text still mostly targets an expert audience.
3:	The rewriting process made quite *substantial changes*, although there are some elements for an expert audience.
4:	The rewritten text *really* targets a lay audience.

Table 6.3.: Questions and answers, included in the expert protocols, for the evaluations of content preservation and style strength. On the left of each answer, the associated score is reported.

While not being a *relative rating* protocol [295] since we did not ask the annotators to perform a ranking between systems, it is not a pure *direct rating* approach either because of the possible influences of the other systems' outputs. While this choice could have led to this kind of bias, it was justified to decrease the cognitive effort of the annotators. Thus, we evaluated the content preservation (Cnt) and the style strength (Sty) by looking at their averaged scores and their ranking comparisons.

To assess the agreement between annotators, we presented a subset of 100 training samples and 50 outputs samples to both the physicians involved in the annotation phase. We exploited the quadratic weighted version of the *Cohen's Kappa* (K_w) [296]. The advantage of using it is to require just the distribution of the distance between two annotations to be ordinal [297]. Considering y_i and y_j as the annotation performed by the i-th annotator and the j-th one, respectively, computing the weights as in the following equation

$$w_{i,j} = \frac{(y_i - y_j)^2}{(N-1)^2} \tag{6.8}$$

where N is the number of choices, allowed us to measure the agreement while taking into account the seriousness of the disagreement between annotators. Trivially, the disagreement between two annotators evaluating a pair of texts as *completely equivalent* and *unimportant details differ*, respectively, is weighted less than a disagreement between *completely equivalent* and *completely dissimilar* (refer to Table 6.3).

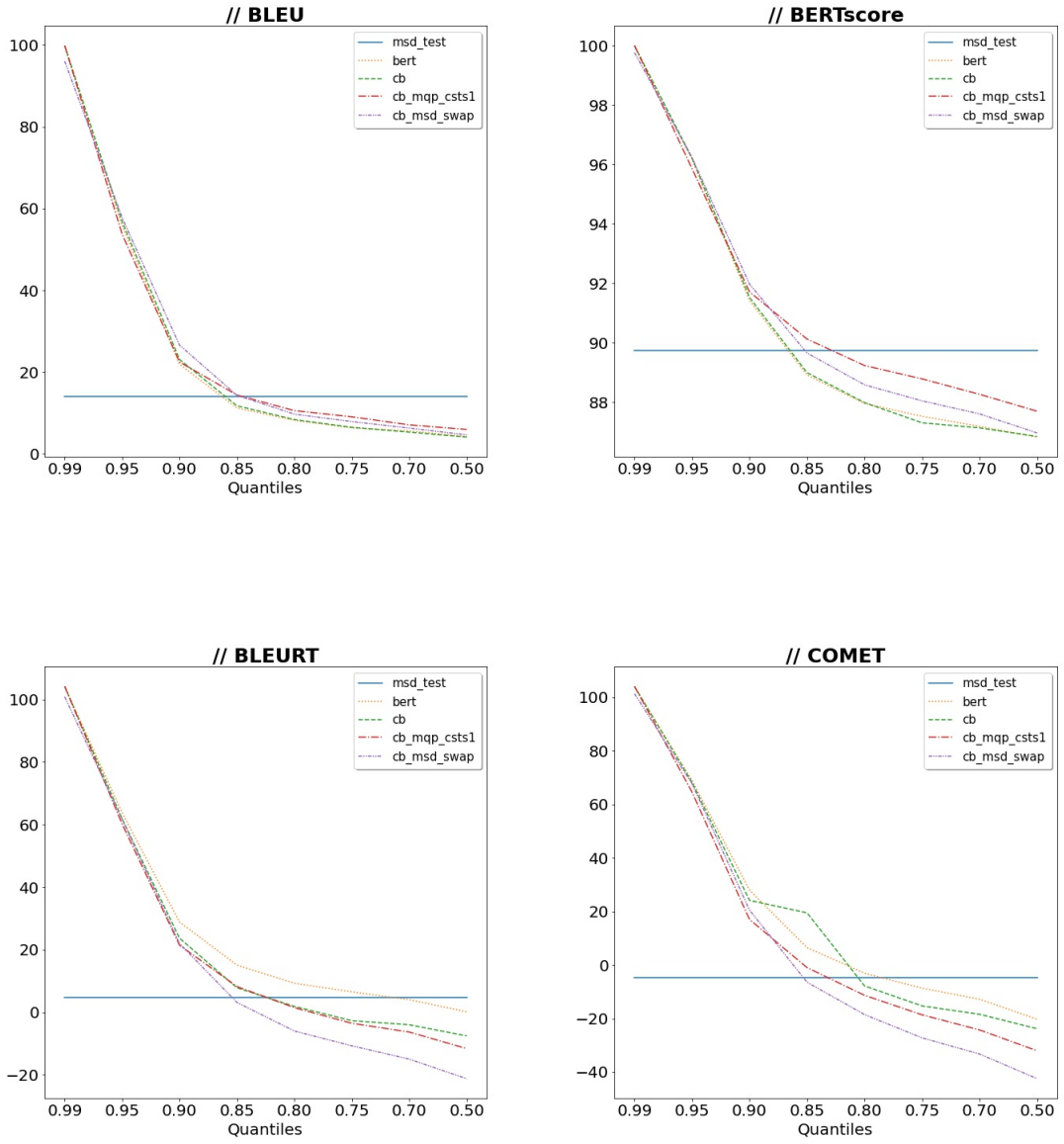

Figure 6.3.: Automatic content preservation metrics (in terms of %) for the collected parallel sets, indicated with the // symbol over the quantile ranges. The most relevant models are reported. The blue solid horizontal line indicates the score computed between the source and the gold reference.

6.5. Results and Discussion

In this section, we reported and discussed all the results, including automatic and human evaluations. The annotations allowed us to study the correlation between automatic

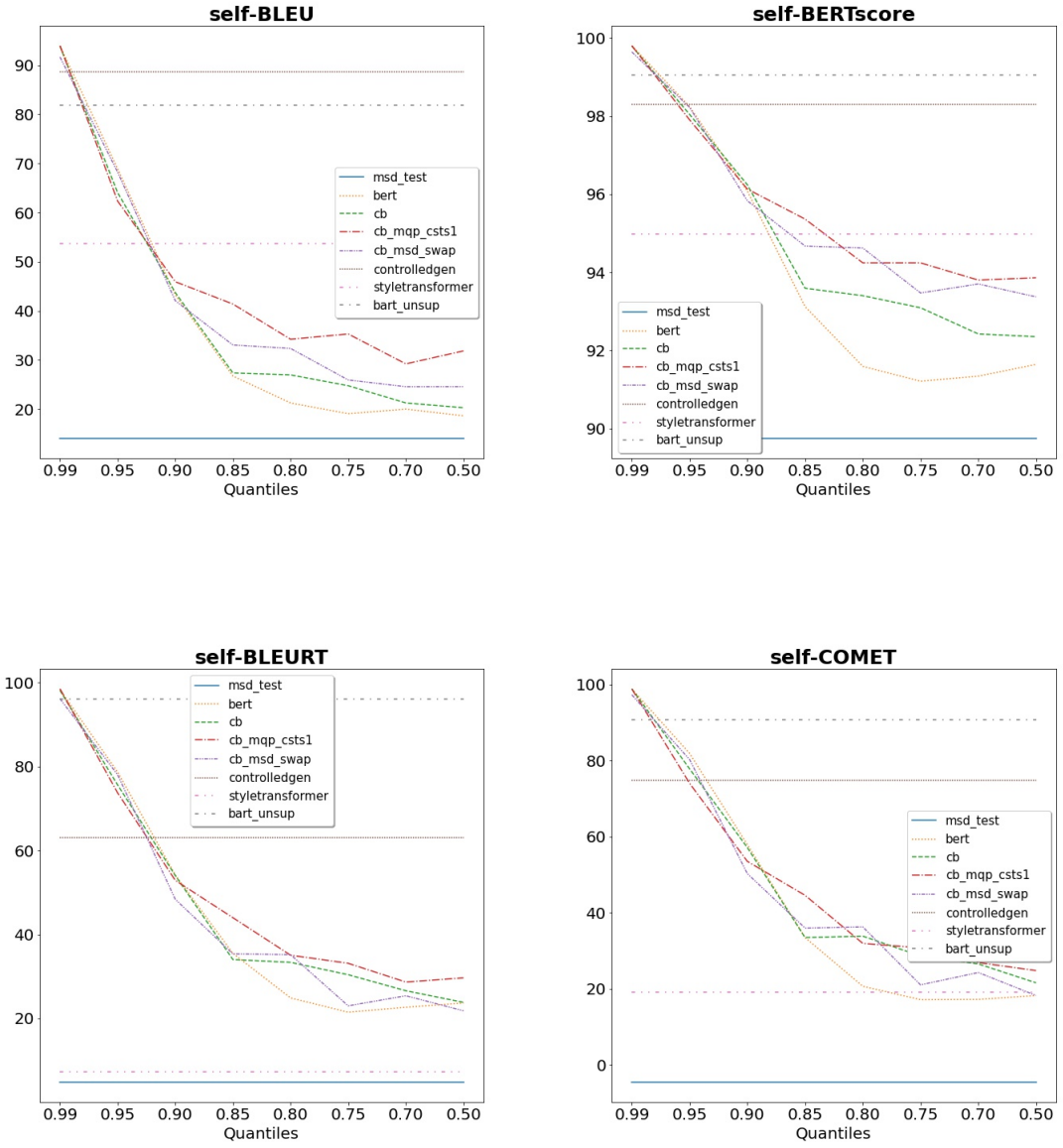

Figure 6.4.: Automatic content preservation metrics (in terms of %) over the quantile ranges for the most relevant models, computed with respect to the source (*self*-). The blue solid horizontal line indicates the score computed between the source and the gold reference, while the other horizontal lines refer to the competitors.

scores and human judgments, and their feedback helped us to perform a qualitative analysis. We pointed out the critical aspects of the expertise style transfer task and the

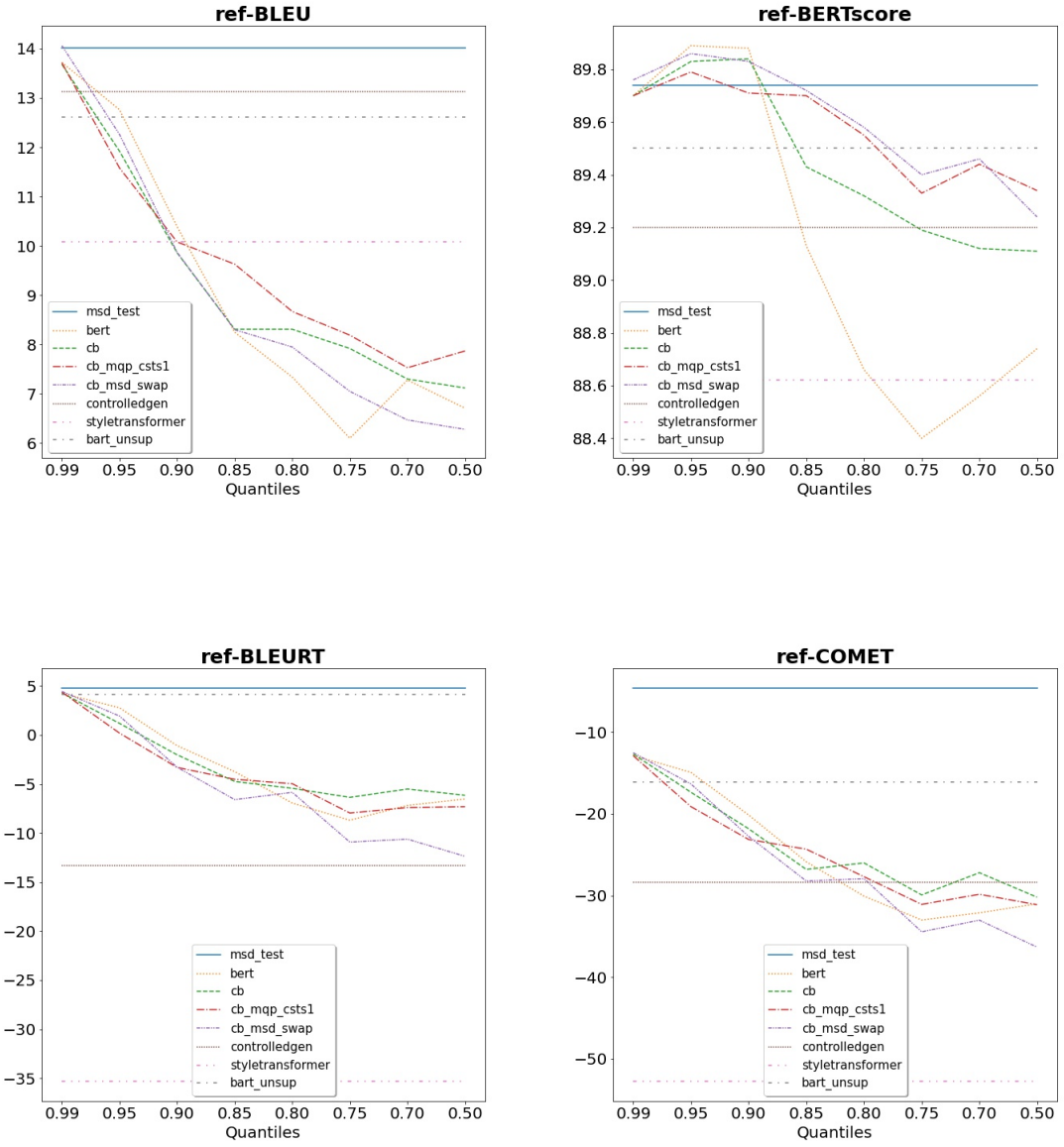

Figure 6.5.: Automatic content preservation metrics (in terms of %) over the quantile ranges for the most relevant models, computed with respect to the gold reference (*ref-*). The blue solid horizontal line indicates the score computed between the source and the gold reference, while the other horizontal lines refer to the system competitors.

models' results.

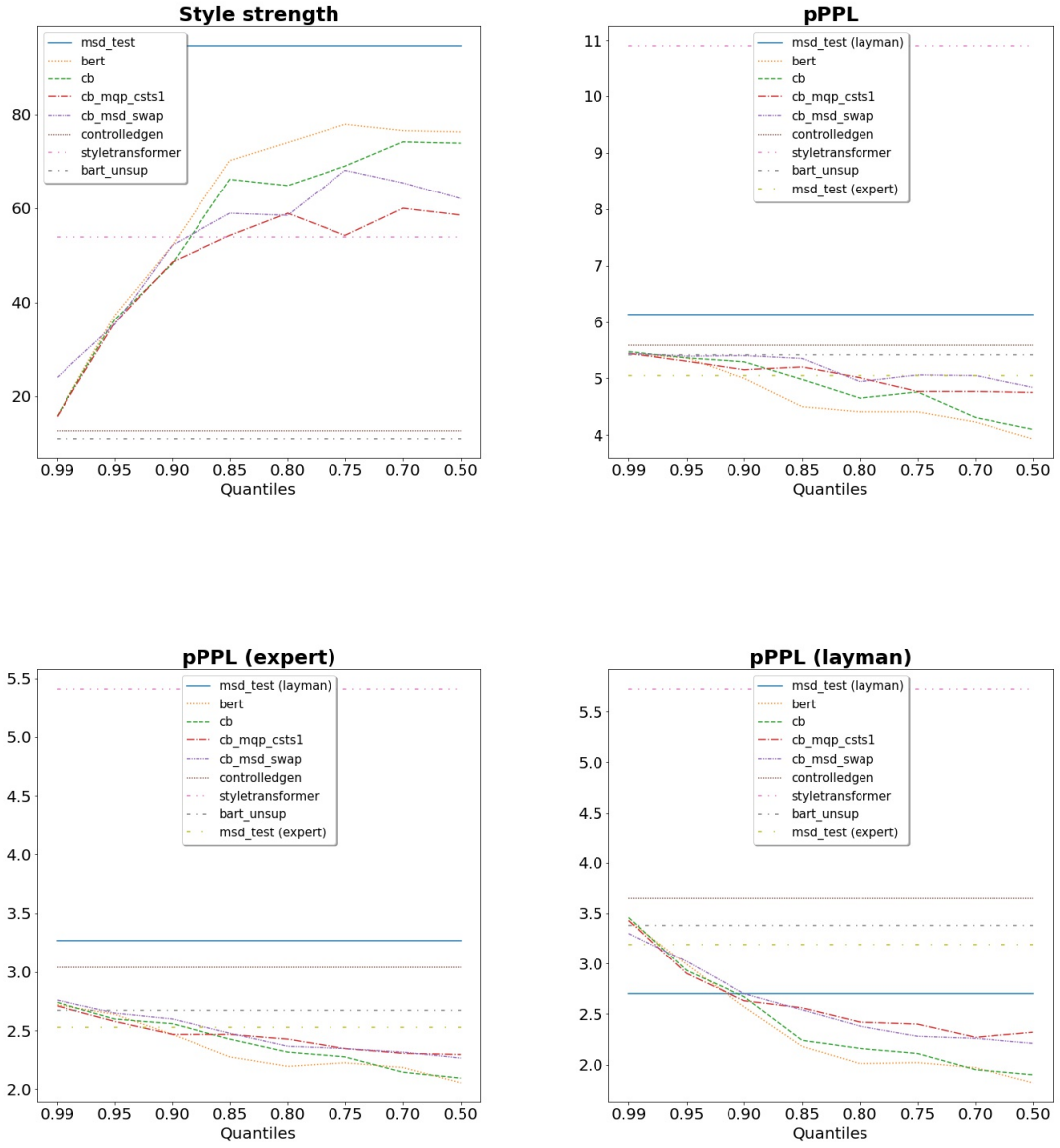

Figure 6.6.: Automatic style strength metric (in terms of accuracy percentage) and (pseudo)perplexity metrics. The latter were computed using a *(Bio-)ClinicalBert* masked language model ($pPPL$) and its fine-tuned versions on expert and lay corpora ($pPPL_{exp}$ and $pPPL_{lay}$, respectively).

6.5.1. Automatic evaluation

For what concerns the automatic evaluation, to help in understanding the impact of the different training sets on the TST performances, we report plots showing the metrics

scores over the quantile value used for collecting the parallel training set. Figure 6.3 reports the content preservation metrics assessed for the collected parallel training sets, indicated with the symbol //. The same metrics were assessed for the TST system outputs in relation to both the source and the target, denoted with the *self-* and *ref-* prefixes, respectively. We reported them in Figure 6.4 and Figure 6.5. Figure 6.6, instead, reports the automatic metrics assessing the style strength of the outputs and their perplexity. We reported the (pseudo-)perplexities using either the original *(Bio-)ClinicalBert* and its fine-tuned versions on the lay or expert corpora. For ease of visualization, we reported only the models that are most relevant to discuss the results. Each subfigure presents as the baseline(s) the scores assumed by the *MSD* test set, and as competitors the state-of-the-art model achieving the best content preservation performance (*ControlledGen*), the state-of-the-art model showing to be more stable across content preservation and style strength performance (*StyleTransformer*), and our unsupervised *BART* model. We report the results comprehensive of all systems in Table 6.4 and Table 6.5.

Figure 6.3 shows that the different automatic metrics share analogous trends, even if values may differ. *// BERTscore* in particular shows a range way limited in comparison with the others, while, given its definition, *// BLEU* does not show negative values as *// BLEURT* and *// COMET*. Pseudo-parallel datasets collected with different models share similar trends across the several metrics. In the beginning, they share similar values, which confirm that, at high quantiles, the sets overlapped more regardless of the model used for the collection. Then, after the 90% quantile, apart from *// BLEU*, the metric starts to capture the differences in the collected datasets. As already discussed in Sec. 6.4.2, in general, for the 85% quantile, the sets are closer to the test set, more or less without regard for the metric in the exam.

The characteristics of the parallel training sets influence the style transfer systems outputs. The same metrics computed to assess the content preservation in the outputs (Figure 6.4 and Figure 6.5) share a similar trend. It is more evident for the *self-*metrics, for which the values ranges are close, too. In particular, *self-* and *ref-*metrics show large differences in values, with a largely lower starting value and a gentler slope. Showing an increase at first, *ref-BERTscore* represents an exception to the trend. Although, its variations are nonetheless extremely modest. At lower quantiles, different metrics report different models' rankings. However, the model trained on the sets collected with the **cb_mqp_csts1** model is more prone to achieve higher ranking positions for both *self-* and *ref-*metrics. Anyway, the low content scores obtained by the gold references

highlight the inherent difficulty of the task. In general, our models obtained far larger *self*-scores, which may either suggest that they are better than the human references or that the (*self*-)metrics are flawed for the task at hand.

TST Model	STS Model	Quantile	// BLEU	self-BLEU	ref-BLEU	// BERT	self-BERT	ref-BERT	// BLEURT	self-BLEURT	ref-BLEURT	// COMET	self-COMET	ref-COMET	SS	pPPL	pPPL (lay)	pPPL (exp)
Test set	-	-	14.01	14.01	14.01	89.74	89.74	89.74	4.77	4.77	4.77	-4.62	-4.62	-4.62	94.67	6.14 / 5.05	2.70 / 3.19	3.27 / 2.53
BART (unsup)	-	-	-	81.78	12.62	-	99.04	89.50	-	96.04	4.14	-	90.75	-16.07	10.96	5.41	3.38	2.67
BART	bert	0.99	99.84	**93.94**	**13.72**	100.00	**99.80**	**89.70**	104.21	**98.26**	**4.26**	104.07	**98.79**	**-12.86**	15.85	5.43	3.43	2.72
		0.95	55.87	68.94	12.76	96.17	98.24	89.89	63.79	78.67	2.77	68.93	81.86	-14.93	37.19	5.36	2.99	2.64
		0.90	21.92	43.27	10.39	91.42	96.08	89.88	28.82	54.14	-1.09	27.97	57.90	-20.14	52.15	5.00	2.57	2.47
		0.85	**11.22**	26.75	8.25	**88.91**	93.12	89.13	15.04	35.50	-3.74	6.47	33.46	-25.88	70.22	4.50	2.18	2.28
		0.80	8.20	21.27	7.34	87.95	91.59	88.66	9.27	24.88	-6.94	**-3.08**	20.65	-30.08	74.07	4.41	2.01	2.2
		0.75	6.36	19.09	6.10	87.52	91.21	88.40	6.51	21.49	-8.69	-8.64	17.13	-33.00	**77.93**	4.41	2.02	2.23
		0.70	5.60	20.02	7.28	87.18	91.34	88.56	**4.01**	22.66	-7.18	-12.79	17.20	-32.12	76.59	4.23	1.97	2.19
		0.50	4.23	18.66	6.71	86.82	91.64	88.74	0.12	23.68	-6.53	-20.27	18.21	-31.02	76.30	**3.93**	**1.82**	**2.06**
BART	cb	0.99	99.88	**93.85**	**13.68**	99.99	**99.79**	**89.70**	104.21	**98.09**	**4.29**	104.03	**98.85**	**-12.69**	15.85	5.47	3.46	2.74
		0.95	56.77	64.08	11.92	96.22	98.04	89.83	61.81	75.60	1.18	68.26	77.80	-17.36	36.30	5.36	2.93	2.60
		0.90	23.09	43.73	9.86	91.52	96.24	89.84	23.65	54.02	-2.02	24.15	57.13	-21.83	48.30	5.29	2.67	2.56
		0.85	**11.75**	27.38	8.31	**88.99**	93.59	89.43	7.80	34.02	-4.74	19.48	33.46	-26.81	66.22	4.98	2.24	2.43
		0.80	8.38	26.98	8.31	87.98	93.40	89.32	**1.85**	33.35	-5.44	**-7.81**	33.84	-26.01	64.89	4.65	2.16	2.32
		0.75	6.49	24.78	7.92	87.30	93.09	89.19	-2.73	30.41	-6.36	-15.29	28.59	-29.93	69.04	4.76	2.11	2.28
		0.70	5.31	21.28	7.30	87.13	92.42	89.12	-4.03	26.59	-5.51	-18.41	26.53	-27.21	**74.22**	4.31	1.95	2.15
		0.50	4.10	20.30	7.12	86.84	92.35	89.11	-7.60	23.83	-6.15	-23.82	21.56	-30.21	73.93	**4.10**	**1.90**	**2.10**
BART	cb_csts1	0.99	99.87	**93.95**	**13.72**	100.00	**99.80**	89.70	104.21	**98.15**	**4.26**	104.05	**98.82**	**-12.81**	15.85	5.49	3.47	2.75
		0.95	54.57	66.77	12.40	95.76	98.08	**89.89**	59.39	75.53	1.38	63.62	77.70	-16.64	35.56	5.39	2.98	2.64
		0.90	21.52	41.00	9.83	90.92	95.28	89.72	19.22	47.11	-2.35	14.09	47.09	-23.02	56.74	5.10	2.53	2.48
		0.85	**12.12**	30.31	8.50	**89.10**	93.83	89.50	**5.44**	32.65	-5.00	**-5.98**	33.17	-26.65	63.41	4.92	2.32	2.38
		0.80	9.02	30.89	7.90	88.34	94.20	89.58	0.17	35.33	-4.11	-15.09	34.54	-26.45	61.48	4.96	2.36	2.38
		0.75	7.66	34.14	8.20	87.90	94.28	89.47	-4.21	35.80	-5.01	-20.30	34.12	-27.26	58.81	4.88	2.35	2.36
		0.70	6.45	29.24	7.95	87.59	93.78	89.56	-6.30	32.14	-5.11	-24.76	29.44	-27.41	62.52	4.60	2.22	2.25
		0.50	5.02	29.01	7.62	87.11	93.49	89.36	-10.12	29.32	-7.19	-30.48	2 4.49	-29.44	**65.63**	**4.57**	**2.16**	**2.25**
BART	cb_mqp_csts1	0.99	99.86	**93.97**	**13.71**	100.00	**99.80**	89.70	104.22	**98.48**	**4.37**	104.06	**98.81**	**-12.89**	15.70	5.44	3.43	2.71
		0.95	53.70	62.35	11.58	95.86	97.90	**89.79**	60.17	73.71	0.19	64.64	74.00	-19.12	35.56	5.30	2.90	2.58
		0.90	22.48	45.94	10.08	91.71	96.13	89.71	21.41	52.87	-3.31	17.01	53.49	-23.14	48.59	5.15	2.63	2.47
		0.85	**14.39**	41.40	9.63	**90.12**	95.36	89.70	8.19	43.99	-4.52	**-0.97**	44.56	-24.33	54.22	5.20	2.56	2.47
		0.80	10.57	34.23	8.67	89.23	94.24	89.55	**1.44**	35.06	-4.97	-11.32	31.93	-27.70	58.96	5.01	2.42	2.43
		0.75	9.06	35.31	8.19	88.78	94.24	89.33	-3.56	33.13	-7.96	-18.61	30.75	-31.09	54.22	4.77	2.40	2.35
		0.70	7.09	29.22	7.53	88.26	93.8	89.44	-6.37	28.66	-7.40	-24.27	26.88	-29.85	**60.00**	4.77	2.27	2.31
		0.50	5.97	31.86	7.87	87.69	93.86	89.34	-11.65	29.66	-7.31	-32.05	24.78	-31.13	58.57	**4.75**	**2.32**	**2.30**
BART	cb_msd	0.99	99.86	**93.80**	**13.75**	100.00	**99.79**	89.70	104.22	**97.97**	**4.23**	104.09	**98.70**	**-12.90**	16.15	5.52	3.47	2.78
		0.95	58.51	67.87	12.50	96.33	98.22	**89.87**	62.04	78.72	2.08	69.00	80.77	-15.20	35.26	5.35	2.95	2.62
		0.90	27.24	47.88	10.16	92.13	96.58	89.75	22.64	55.73	-2.91	22.19	58.80	-22.43	44.59	5.27	2.73	2.55
		0.85	**14.11**	35.38	8.60	**89.76**	95.00	89.68	**3.93**	39.08	-5.17	**-4.69**	39.50	-26.45	56.74	5.13	2.51	2.48
		0.80	9.89	30.67	7.79	88.71	94.41	89.53	-5.42	33.48	-7.99	-17.40	33.51	-30.28	59.26	5.26	2.44	2.41
		0.75	7.90	29.98	6.87	88.10	94.39	89.41	-10.69	29.32	-9.91	-25.89	31.58	-32.46	52.89	5.03	2.44	2.39
		0.70	6.45	29.26	7.44	87.67	93.90	89.46	-14.34	28.48	-8.68	-31.65	29.17	-30.13	**61.63**	4.96	2.33	2.37
		0.50	4.71	24.18	6.44	87.00	93.42	89.31	-21.07	23.04	-10.62	-41.42	20.34	-35.33	60.30	**4.66**	**2.19**	**2.21**
BART	cb_msd_swap	0.99	96.05	**91.66**	**14.06**	99.76	**99.64**	89.76	100.94	**96.11**	**4.44**	101.33	**97.27**	**-12.52**	24.00	5.44	3.30	2.76
		0.95	57.78	68.24	12.26	96.23	98.22	**89.86**	61.69	78.02	1.93	67.74	80.25	-16.34	35.26	5.39	3.02	2.65
		0.90	26.58	42.08	9.87	91.97	95.83	89.83	21.94	48.37	-3.29	20.52	50.27	-22.76	52.15	5.40	2.70	2.60
		0.85	**14.31**	33.06	8.30	**89.65**	94.67	89.72	**3.11**	35.37	-6.59	**-6.42**	35.96	-28.23	58.96	5.35	2.54	2.48
		0.80	9.70	32.36	7.95	88.58	94.62	89.58	-6.04	35.19	-5.85	-18.53	36.26	-27.94	58.52	4.94	2.38	2.37
		0.75	7.89	25.95	7.05	88.04	93.47	89.40	-10.77	23.00	-10.92	-27.23	21.02	-34.44	**68.15**	5.06	2.28	2.35
		0.70	6.30	24.58	6.47	87.60	93.70	89.46	-15.02	25.40	-10.62	-33.26	24.27	-33.02	65.48	5.05	2.26	2.32
		0.50	4.65	24.59	6.28	86.96	93.37	89.24	-21.27	21.82	-12.37	-42.46	18.23	-36.32	62.07	**4.84**	**2.21**	**2.27**

Table 6.4.: Results of the automatic evaluations of our models with respect to the collected parallel training sets. Both sets and models were evaluated at various quantile thresholds. For the // metrics, with the **bold** font we indicate the values closer to scores obtained on the test set. For the others, we used it to indicate the best scores obtained. In particular, the values in red indicates the state-of-the-art scores. For the test set, the perplexity scores of both lay and expert corpora are reported in this order.

TST Model	STS Model	Quantile	// BLEU	self-BLEU	ref-BLEU	// BERT	self-BERT	ref-BERT	// BLEURT	self-BLEURT	ref-BLEURT	// COMET	self-COMET	ref-COMET	SS	pPPL	pPPL (lay)	pPPL (exp)
Test set	-	-	14.01	14.01	14.01	89.74	89.74	89.74	4.77	4.77	4.77	-4.62	-4.62	-4.62	94.67	6.14 / 5.05	2.70 / 3.19	3.27 / 2.53
OpenNMT+PT	-	-	-	59.89	9.92	-	97.16	89.13	-	60.75	**-10.96**	-	56.43	-33.62	21.04	5.49	**3.24**	**2.64**
UNTS	-	-	-	20.49	3.94	-	87.87	83.25	-	-47.99	-74.96	-	46.61	-96.89	37.48	31.78	18.46	14.89
ControlledGen	-	-	-	**88.61**	**13.13**	-	**98.29**	**89.20**	-	**63.10**	-13.35	-	**74.81**	**-28.38**	12.74	5.58	3.65	3.04
DeleteAndRetrieve	-	-	-	6.66	2.95	-	85.05	83.97	-	-78.48	-83.77	-	-91.43	-110.99	**79.56**	**5.46**	4.44	4.51
StyleTransformer	-	-	-	53.66	10.09	-	94.98	88.62	-	7.38	-35.33	-	19.02	-52.77	53.93	10.9	5.73	5.41

Table 6.5.: Results of the automatic evaluations of the state of the art models are reported, as in Table6.4: with the **bold** font we indicated the best scores obtained and the red color the state-of-the-art scores. For the test set, the perplexity scores of both lay and expert corpora are reported in this order.

As can be easily noted in the first plot of Figure 6.6, the trend is reversed for the style strength accuracy metric, computed with the TextCNN style classifier. The differences in this behavior are due to the dataset used to train the models. For the top quantiles, the parallel texts may be considered very similar, or even equal at the extreme case at the 99% quantile, as demonstrated by the close to 100% // *BLEU* scores. On the contrary, the lowest quantiles contained pairs of (too) dissimilar sentences. In one extreme case, the model got trained to mostly reproduce the input, while, on the other extreme, the model got trained to generate outputs too dissimilar from the source but more towards the lay style.

Note that, only in the first quantiles, we outperformed the content preservation state-of-the-art performances. In the subsequent quantiles, instead, we have overcome the style strength of state-of-the-art systems (if we exclude the score achieved by the *DeleteAndRetrieve* system, which is the worst one in preserving the content, as can be seen in Table 6.4 and Table 6.5).

Interestingly, the models trained on non-fine-tuned models (**cb** and **bert**) achieved higher style strength, in general, than our other models. Conversely, they performed worse on the content preservation metrics (the *self-* ones, in particular). It suggests that non-specialized *Semantic Textual Similarity* models tended to retrieve pairs of less related texts (at lower quantiles), which led to an increase in variability in the collected pseudo-parallel datasets, resulting in higher style results on the TST task.

About the other models' baselines, both *ControlledGen* and the unsupervised *BART* show great content performances but the worst style strengths, suggesting that those models mostly reproduced the input without significant changes. It justified choosing the trade-off of the *StyleTransformer* as the competitor in the human evaluations. Anyway, the test set style strength score was definitely unachievable for all the models. At the same time, the high test accuracy confirms the ability of the style classifier model to distinguish between expert and lay styles.

Figure 6.6 also reports the (pseudo-)perplexities computed with *(Bio-)ClinicalBert* ($pPPL$) and its fine-tuned versions on the lay or expert corpora ($pPPL_{lay}$ and $pPPL_{exp}$, respectively). For all three metrics, the models show a trend similar to the one seen with the content preservation metrics. The decrease in perplexity with the quantile ranges reflects the variability increase of the related training set. Such behavior is more pronounced with the perplexity model fine-tuned on the lay corpus.

Curiously, each model achieved greater $pPPL_{lay}$ scores in comparison with the $pPPL_{exp}$ ones. It seems to suggest that the lay outputs generated by the models present more similarities with the expert training corpus than with the lay one. The lay test corpus is the exception for it (while the expert test corpus shares this behavior as was expected), reflecting that the test set was extracted from the same corpus from which the training dataset was collected. Interestingly, at lower quantiles, the perplexity metrics between models trained on sets collected with fine-tuned models are closer to the ones trained on sets collected with non-fine-tuned models and vice-versa (even if such differences in the perplexity are modest). The reason for this behavior has to be searched in the variability of the training sets, again: using non-fine-tuned models to collect the parallel sets leads to obtaining more dissimilar sentences in the pairs, thus increasing the variability of what the model has seen during training.

Our models have shown, in general, lower perplexities than the state-of-the-art ones and our unsupervised BART baseline, regardless of the quantile range, showing the goodness of our models.

6.5.2. Human evaluation

Table 6.6 reports the results of the analysis of our model (based on the training set collected with **cb_mqp_csts1** at the 85% quantile), compared with the state-of-the-art model (*StyleTransformer*), and the gold references. Before analyzing the results, we first assessed the agreement between annotators to establish the quality of the annotations processes. In the lay case, pairs of annotators evaluated the style as a binary task by choosing the easier-to-understand text between the source and one system output. They achieved an averaged Cohen's Kappa (K^{lay}) of .32 ($\pm$ .15), which may be considered a fair agreement as suggested by previous literature [298]. However, the large standard deviations suggest that some pairs may be easier to annotate (low disagreement degree) while others are harder (high disagreement degree). We also evaluated the agreement on the single systems. It is worth noting that the *StyleTransformer* (*ST*), for which the annotators agreed most on average, is also the system that achieved the lowest style-related score for lay people. It means that their concordance reflects in low performance (the annotators mostly agree in saying that its output is not easier to understand for them compared with the source). We discussed this aspect in the section dedicated to qualitative analysis. Our model achieved higher results, comparable with the gold reference, showing that it was actually able to perform some changes in the simplification

direction.

Unlike lay annotators, we asked the experts to judge by following a measuring scale. Thus, the original Cohen's Kappa was not suitable to assess their agreement. We, therefore, applied its quadratic weighted version (as described in Sec. 6.4.2). Plus, we asked the physicians to judge not only the style but the content preservation too. Thus, we assessed the agreement score for both content (K_w^{cnt}) and style (K_w^{sty}), separately. However, since the weighted Cohen's Kappa interpretation is debated, with its results influenced by the weight scale [297], we also assessed the Spearman correlation indices (ρ^{cnt} and ρ^{sty}). The annotators achieved a Kappa agreement of .42 for content preservation on the outputs and .50 on the style strength, which can be considered moderate agreements. The Spearman score confirmed the fairness of those agreements [299]. We evaluated the agreements on the single systems for this setup, too. It is noticeable that the annotators shared more agreement on the content preservation of the system results while agreeing only slightly on the gold reference (Ref). It suggests a higher aggressiveness in changes for the reference with respect to the source text, which leaves room for more interpretation for the annotators. Regarding the style analysis, the expert annotators showed outcomes analogous to the ones obtained with lay people.

Moving to the proper evaluation analysis, our model ($Ours$) obtained larger scores on average with respect to the state-of-the-art model, highlighting once again the improvements brought by our approach. However, while its content preservation scores are even greater than the reference, its style scores are still worse. It indicates that our model prefers to change less instead of messing with the meaning of the input. Even if not optimal, avoid to lose information is preferable, even at the cost of not changing or changing just minor things to the layman's direction. In this sense, it can hardly compete with the abstraction level of the gold references. Anyway, we compared the three systems in relation to the others, measuring the number of times that a system outperformed another on the same sample. The heatmaps in Figure 6.7 confirm that our model outperforms the reference in content preservation, while it was not considered as good as the reference in changing the style. However, both of them largely outperformed the *StyleTransformer* for both content and style, and a combination of the two (overall). In particular, besides often performing only minimal changes, the outputs of our model were still perceived as easier than the source texts by lay people, which is a great advancement of the state-of-the-art. Indeed, in the lay evaluation, our model outputs resulted extremely close to the gold references.

System	#	Agreement K^{lay}	K_w^{cnt}	K_w^{sty}	ρ^{cnt}	ρ^{sty}	Human Evaluation Sty^{lay}	Cnt	Sty	Automatic Evaluation $BLEU$	$BERT$	$BLEURT$	$COMET$	SS
Ref	50	.26 ± .36	.24	.21	.41	**.31**	**69.00**	65.12 ± 29.23	**71.60 ± 27.59**	14.01	89.74	4.77	-4.62	**94.67**
ST	50	**.31 ± .45**	**.63**	**.34**	.62	**.31**	28.50	62.00 ± 25.35	31.35 ± 17.15	**53.66**	94.98	7.38	19.02	53.93
Ours	50	.16 ± .27	.57	.20	**.66**	.19	66.50	**79.48 ± 20.42**	48.80 ± 27.95	41.40	**95.36**	**43.99**	**44.56**	54.22
All	150	.32 ± .15	.42	.50	.50	.52	-	-	-	-	-	-	-	-

Table 6.6.: Evaluation results for the gold reference (Ref), the StyleTransformer (ST), and our model ($Ours$), as well as the three systems together (All). The first block regards the agreement between annotators assessed with a given number of samples (#). For lay annotators, the agreement is assessed with Cohen's Kappa (K^{lay}), while for the experts it is measured with the quadratic weighted version (K_w) and the Spearman correlation index (ρ), for both content preservation (cnt) and style strength (sty). The second block reports the human evaluation results (in terms of percentages) of the different systems for lay and expert annotations. For the former case, the style is evaluated as the ratio between the number of texts judged easier to understand than the related source text (Sty^{lay}). For the latter, both content and style scores are normalized with the range of the related scale. The third block is dedicated to the automatic (self-)metrics computed with respect to the source text and the style strength. The best results for each metric are shown in **bold**.

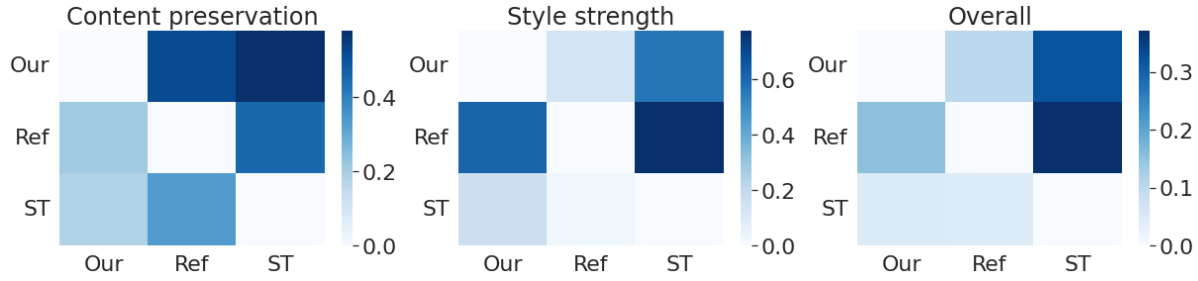

Figure 6.7.: Ranking comparison regarding human evaluations for content preservation, style strength, and a combination of the two (overall). The darker the color of the cell (i, j), the more times the i-th system (on the y-axis) was ranked better than the j-th model (on the x-axis) on the same sample. Note that the sum between the cell (i, j) and the cell (j, i) is lower than 1 because of cases of draws. For the same reason, the diagonal is represented by all zeros.

Furthermore, we conducted an expert evaluation to retrieve insights on the content preservation quality of the collected parallel sets at different quantiles. First, we calculated the agreement between annotators (using 100 samples) with the quadratic weighted version of the Cohen's Kappa and the Spearman correlation. The annotators shared an agreement between moderate and substantial ($K_w^{cnt} = .60$, $\rho^{cnt} = .64$). Figure 6.8 reports the results of the evaluation in terms of the normalized average and standard

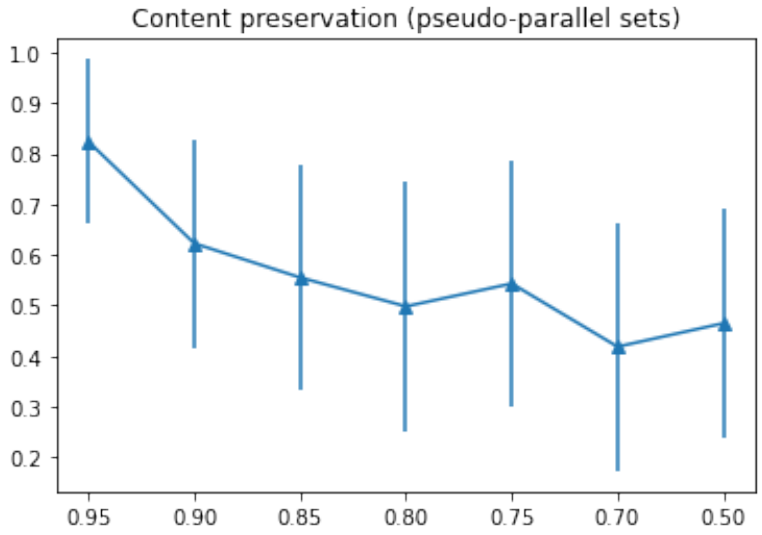

Figure 6.8.: Human evaluation results in content preservation for the collected pseudo-parallel datasets over the quantile thresholds. The results are reported in terms of the normalized average and standard deviation scores.

deviation of the scores. As expected, as for the automatic metrics, the averaged content preservation score generally decreases as the quantile threshold decreases. However, at the lower quantiles, the trend becomes more ambiguous, especially between 70% and 50%. It may be due to similar content preservation performance for lower quantiles, reflected by the automatic metrics (Figure 6.3), too, showing similar values between lower quantiles-related parallel sets. The same behavior is shown by the self- and ref-metrics on the style transfer task (Figures 6.4 and 6.5). These results suggest that decreasing the threshold below some value led to obtaining datasets of similar (low) qualities.

Even if not directly comparable with the same task on the outputs since different setups are involved (and different pairs of annotators took part at the two different setups), the obtained results point out interesting things. The larger agreements suggest that annotators may agree more on the content preservation quality of parallel sets automatically collected than the ones created with a generation model or even the references annotated by humans (gold). These results suggest that the implemented pipeline may be successfully applied, with respect to the similarity threshold, to retrieve parallel corpora for the style transfer task in the medical domain. Eventually, such a collection phase may be employed as a preliminary step to a human annotation phase. It may help for reducing the annotators' efforts while minimizing the aggressiveness of the changes between source and target texts.

6.5.3. Comparing automatic and human evaluations

The comparison of the automatic evaluation scores with the human evaluation scores in Table 6.6 reveals some interesting findings. Although the comparison of BLEU scores between the *StyleTransformer* and our model are not directly comparable, the automatic *self*-metrics for content preservation tend to have a similar behavior as the human evaluation scores. To further explore the agreement between the automatic and human evaluation metrics, we analyzed the correlation between them for both content preservation and style strength. Table 6.7 shows the Spearman correlations for *self*- and *ref*-content metrics (ρ^{self} and ρ^{ref}, respectively), as well as for style (ρ^{ss}). The results show that the correlation between the *self*-metrics and human judgments is higher compared to the correlation between the *ref*-metrics and human judgments. This is consistent with past literature [266]. Overall, BLEURT and COMET are the metrics that show the highest correlation with human judgments, both in the *self*- and *ref*-setting. It is also worth noting that the reference texts (Ref) have a lower correlation with *self*-metrics compared to the two models, which highlights the aggressive differences between the reference texts and the associated source texts. Furthermore, the *StyleTransformer* model (ST) shows higher correlation scores, suggesting that there is a stronger correlation between automatic metrics and human evaluations when judging a less-performing system.

When examining the results for the style aspect, a noticeable feature is the low correlation score for the *StyleTransformer*. This is likely due to the model's strategy of replacing complex terms with simpler but often unrelated words, which are evaluated as simplifications by the classifier that is less influenced by the outputs' meaning and fluency than humans. Additionally, the correlation score for reference texts, which the style classifier was able to identify well, is notably low. This highlights the difficulty for humans in assessing the style strength, separating it from the structure and semantics. These findings are in line with recent studies in the field [300].

6.5.4. Qualitative analysis

Our manual inspection was conducted on a significant number of examples and incorporated the feedback from the expert annotators. The three models we focused on were: our *cb_mqp_csts1* model at 85% quantile, the *StyleTransformer* system, and a model based on the *cb_msd_swap* dataset (at 85% quantile) which had similar performance to

System	#	Humans-BLEU ρ^{self}	Humans-BLEU ρ^{ref}	Humans-BERT ρ^{self}	Humans-BERT ρ^{ref}	Humans-BLEURT ρ^{self}	Humans-BLEURT ρ^{ref}	Humans-COMET ρ^{self}	Humans-COMET ρ^{ref}	Humans-SS ρ^{ss}
Ref	250	**.45**	-	.42	-	.41	-	.43	-	.10
ST	250	.59	.21	**.68**	.22	.67	**.43**	.67	.42	- .03
Ours	250	.64	**.34**	.60	.21	.62	.30	**.65**	.30	.20
All	750	.39	.26	.45	.27	**.60**	**.46**	.58	.44	.34

Table 6.7.: Spearman correlation scores between expert human judgments and automatic metrics for the gold reference (*Ref*), the StyleTransformer (*ST*), and our model (*Ours*), as well as the three systems together (*All*). The # column reports the number of samples used to assess the correlation scores. For content preservation scores, we reported correlation involving both self- (ρ^{self}) and ref- (ρ^{ref}) metrics. The last column instead assesses the correlation (ρ^{ss}) between the style annotations and the outputs of our trained style classifier. The best content-related correlations for each system are shown in **bold**.

our model. We did not take into consideration the models' outputs that were not mere repetitions of the input. The results of our analysis allowed us to draw some qualitative conclusions. Firstly, we observed that each model employed different strategies to simplify the text. In instances where our models couldn't substitute a complex term, they attempted to provide an explanation:

> **Source:** Pulmonary arteries are affected, sometimes causing pulmonary hypertension.
>
> **StyleTransformer:** Pulmonary arteries are affected, sometimes causing intravenously recurring.
>
> **cb_mqp_csts1 (85%):** Pulmonary hypertension is a condition in which blood pressure in the lungs is too high.

This behavior is particularly highlighted for short and/or incomplete sentences:

> **Source:** IV fluids.
>
> **StyleTransformer:** common fluids.
>
> **cb_mqp_csts1 (85%):** IV fluids are given intravenously.
>
> **cb_msd_swap (85%):** Blood and urine are given intravenously.

Of course, not all of the found examples are good attempts

> **Source:** Biopsy.
>
> **cb_mqp_csts1 (85%):** Biopsy is the most common type of bleeding disorder.
>
> **cb_msd_swap (85%):** Biopsy is given intravenously.

Despite that, these examples indicate that our models tend to provide explanations when

unable to substitute significant terms, and at times, exhibit good domain knowledge. Additionally, we discovered some domain knowledge related to gender in our models, which was interesting.

> **Source:** Most *patients* have pelvic pain (which is sometimes crampy), vaginal bleeding, or both.
>
> **StyleTransformer:** Most *people* have pelvic pain (which is sometimes crampy), vaginal bleeding, or both.
>
> **cb_mqp_csts1 (85%):** Most *women* have pelvic pain (which is sometimes crampy), vaginal bleeding, or both.

The above example helps in pointing out also another common tract: all the models tend to not reproduce the word *patient*, substituting it with a less domain-specific term such as *people* (or *woman* under the above particular conditions). It suggests that in the lay corpus people are not referred to as patients as in the expert corpus. However, being a common word, such a change was mostly not considered by the annotators as a valid simplification. We also found that our models were mostly able to deal with some abbreviations, like *hr* and *yr*, while the *StyleTransformer* model was not:

> **Source:** Jaundice usually peaks within 1 to 2 *wk*. Recovery phase: During this 2- to 4-wk period, jaundice fades.
>
> **StyleTransformer:** Jaundice usually peaks within 1 to 2 *relieving*.
>
> **cb_mqp_csts1 (85%):** Jaundice usually peaks within 1 to 2 *weeks*.
>
> **cb_msd_swap (85%):** Jaundice usually disappears within 1 to 2 *weeks*.

The example above allows us to point out another behavior that we found in common among the models. When the source text presents a complex structure, the models tend to remove part of the whole text to simplify its structure. In particular, when the source text consists of more than one sentence, the models prefer to truncate the output, removing either the left or the right context. We believe that this behavior is due to the nature of the training corpus, which mostly consists of one-sentence texts (and the limited token lengths accepted in input by the models, in some cases). As in the example above, when the *StyleTransformer* cannot cope with a stylistic change, its output looks messy. It is also partly demonstrated by larger (pseudo-)perplexities as well as by the annotators' results and feedback. The lay annotators always pointed out the presence of ill-structured sentences. While the model attempted to simplify the inputs, it often made them incomprehensible, causing the annotators to prefer the source (expert) texts.

This, however, put the annotators on the spot in cases where the source was actually hard to understand and/or the meaning of the output was easy to understand even if presenting messy terms. Furthermore, to some annotators, shorter sentences mostly felt easier to understand, while others felt that longer texts gave them more contextual information and reputed them more understandable. Often, the choice was complicated by differences in the meaning caused by additional information in one of the texts in a pair. Another issue highlighted by lay annotators is the presence of minimal changes (e.g., changing the capital letter of a common term in the middle of the sentence to its corresponding lower one), often judged randomly (which also led to an increase in the disagreement). Another common minimal change pointed out by both laymen and experts is the change of common words (such as *patients*) with other common ones (like *people*). In most of these cases, lay annotators declared to have often made their choices randomly, while expert ones usually annotated them as *no good changes* or even *no changes* where they were the only alterations made.

Part III.

Explaining Transformers

7. Hierarchical Transformers to the Rescue: Extractive Summaries as Explanation

Performing document classification of free text has already been shown its importance in literature, even in the healthcare domain. One example of document classification is sentiment analysis, which has shown benefits also in multimodal settings, where adding features extracted from free text is essential for good performance in sentiment classification of audio and videos [301]. However, today's systems often lack transparency, as they cannot provide an interpretation of their reasoning. In recent years, this has been a well-known problem in the scientific community. In fact, the contribution that *artificial intelligence* algorithms are making in shaping tomorrow's society is constantly growing. Given the high performance that today's models can achieve, their application is spanning an increasingly large landscape of fields. This is motivating a rapid paradigm shift in the use of these technologies. We are moving from a paradigm in which AI models are required to deliver the highest possible performance, to one in which such systems are required to provide information about taken decisions that is interpretable by humans.

We are referring to the *explainable artificial intelligence* (XAI) paradigm. As stated by the *DARPA*'s XAI program launched in 2017, the main goal of XAI is to create a suite of models that provide an explanation without affecting performance [73, 74, 75]. That is, to pass from the concept of black-box models, in which it is hard (or even impossible) to get any sort of explanation from them, to white-box ones, in which the model also provides results that are understandable by the final users, or at least by the experts in the application domain [72]. This may lead systems of the near future to address the needs of government organizations and the users who use them, such as the right to explanation, which can raise the reliability of users in the system, and the right to decision rejection, especially in applications where a human-the-loop approach

is expected (*Articles 13–15, 22 of the EU GDPR*).

Since the *Transformer* architecture was introduced by Vaswani et al. [47], the NLP research has made great strides. Additionally, its community is beginning to approach to this new paradigm [76]. However, the task of explaining NLP systems is certainly not an easy one, in a context where models based on deep neural networks, usually referred to as the least explicable models of machine learning, take the lead.

In an effort to investigate the behavior of these models and provide some sort of human-understandable interpretation, the weights of the attention mechanism inherent in these structures have often been taken into account (Section 7.1.4). In this work, we propose and compare two transformer-based models to perform tasks of sentiment analysis, while retrieving an explanation of the models' decisions through a summary built by extracting the sentences of the document that are the most informative for the task in hand[1]. That is, we exploited the *extractive* (*single document*) *summarization* paradigm (Section 7.1.2). In particular, for one of the two models, we made use of the attention weights of the transformer model to get insights on the most relevant sentences. To do so, we exploited hierarchical configurations (Section 7.1.3). We evaluated our models on a binary sentiment classification task. However, the underlying structures may be easily adapted for any document classification task.

In particular, to evaluate the classification performance, we shifted the focus from patients-related opinion to a benchmark extensively used in past literature, the IMDB movie reviews dataset [304]. Such a choice is justified by our intent to analyze our methodology regardless of the domain of application, gathering indications on the performance of our methods in comparison with past works.

To assess the explainability performance, we annotated some samples of the dataset to retrieve human extractive summaries from the training and test sets, and then assessed the overlap between these and the models' ones. The annotation phase was necessary since, to the best of our knowledge, this is the first kind of work trying to exploit model architectures to retrieve an extractive summary of a document while performing its sentiment classification (Section 7.1.1), even for a benchmark of the chosen dataset.

The main contributions of our work may be resumed as:

[1]The work presented in this chapter is an extract of our paper published in *Electronics* [302] entitled *Explainable Sentiment Analysis: A Hierarchical Transformer-Based Extractive Summarization Approach*, and our paper published in *Proceedings of European Semantic Web Conference (ESWC)* [303] entitled *Extractive Summarization for Explainable Sentiment Analysis using Transformers*

- **A new approach** to explain document classification tasks as sentiment analysis, by providing extractive summaries as the explanation of the model decision;
- **Exploring use of attention weights** of a hierarchical transformer architecture as a base to achieve extractive summaries as an explanation of the document classification task;
- **A new annotated dataset** for the evaluation of extractive summaries as an explanation of a sentiment analysis task. We shared the annotated dataset together with the algorithm code on our *Github* page[2];
- **Two different proposed models**, both based on transformer architectures, analyzed in terms of the performance in both the classification and explanation tasks.

Furthermore, we proposed an ablation study for the hierarchical model, to evaluate the impact of (sentence) masking and positional embedding, and the role of the first transformer when it is frozen during the training phase. Additionally, we implemented a new a posteriori metric to evaluate the models' summaries with no regard to prior annotations.

7.1. Related Works

7.1.1. Explainability in Sentiment Analysis

Sentiment analysis, also called opinion mining [305], consists of the classification task of the polarity of some text. In recent years, it has gained interest not only in research but also in industry. It is particularly true due to the advent of blogs and social media, and, thus, the impressive growth that shared content has shown. Organizations are currently using these kind of data for their decision making processes instead of conduct surveys, for example, to rank products or services from the users' reviews [306] and provide recommendations to the users [307], to predict changes in the stock prices [308], or, to give an example closer to the domain of our work, to predict incomes from movies at the box-office basing the prediction on the online movies' reviews [309].

In particular, sentiment analysis tasks may be performed at the word, sentence, and document levels. The latter is, of course, the more difficult one to perform because of the greater length of the text, which also may lead to the presence of noisy words or

[2] www.github.com/lbacco/ExS4ExSA

sentences. Longer documents more easily show words and sentences with a polarity that can be neutral or even opposite in respect to the overall polarity of the entire document. This kind of task presents limitations on the interpretability of the decision made by the models. However, in literature, there are not so many works dealing with this side of sentiment analysis models. One way the past literature dealt with the *explainable sentiment analysis* field was by exploiting a fourth degree of the task, the so-called *entity/aspect-based* level.

Originally called *feature-based* level [310], it consists of performing a finer-grained analysis by directly looking at the opinions themselves reported in the document. This concept is based on the assumption that each opinion may be seen as the combination of sentiment and its target (the entities and their attributes, the aspects). For example, for the sentence

> The photography is nice, but the movie is way too slow

we could say it is a negative comment, but not in its entirety. In fact, we can individuate two entities or aspects (*aspect extraction*), the *photography* and the *movie*, and we can also determine their sentiment (*aspect sentiment classification*), respectively, as positive and negative. The emphasis on the latter may indicate that the overall score of the sentence is more negative than positive. Thus, combining the aspects' polarity score, with this approach it is also possible to retrieve an overall polarity score [311], and consequently the document sentiment. This is done while also giving finer-grained insights, for example turning the free text into a structured list of entities and aspects and their associated sentiments. The main disadvantage of this approach is the effort to extract and annotate entities, attributes, and sentiment words or phrases, while also dealing with the presence of implicit aspects.

Another way to explore the explainability in sentiment analysis exploited by the past literature is to use the so-called *sentiment lexicons*. Such items are de facto dictionaries in which words (but also phrases) are associated with some polarity score. The main advantage of this kind of approach is the possibility to exploit already existing accessible resources, such as *SentiWordNet* [312] or *SenticNet* [313] and its newer versions. However, it is not rare that some words assume different connotations depending on the domain. Thus, in some cases is useful to build some custom lexicon by extracting aspects and opinions [314] and, eventually, to combine it with the other external resources. Aside from this, another advantage of this approach is being completely unsupervised, not requiring an annotated corpus for training: for each document, the polarity scores

are merely combined (e.g., by average) to provide its overall sentiment. In other cases, the external knowledge may directly inject domain information into the input text, for example, by leveraging a sentiment knowledge graph (SKG) to enable a BERT model to incorporate the external knowledge [315].

However, our approach does not make use of any external resources aside from the training dataset, and does not work at the aspect level but at the document level to extract the overall sentiment while extracting a summary from the original text as an interpretation of the models' decision. To the best of our knowledge, no previous works have proposed something close to our approach.

7.1.2. Automatic Text Summarization

The automatic text summarization (ATS) topic is gaining more and more interest in research, not only in the academic but also in the industrial field. This is due to the increasingly large amount of textual data on the various archives of the Internet. It is not difficult to imagine the value it may have to automatically summarize scientific papers, to give an example close to our world. Additionally, such an approach could be beneficial to analyze clinical documents (usually, kinds of documents that are very long), social media opinions, product reviews, etc. From these points of view, it becomes even more obvious how it would be worthy to automatize a summarization process if you think about how much a manual text summarization (MTS) may cost, in terms of both time and human efforts.

Not least, the ATS may be used as an explanation of a model decision, as in this work. However, ATS is not a monolithic topic of research, but it may be seen as spread in many sub-fields where researchers are putting their efforts in. Following the nomenclature in [316], we may distinguish the first and most important differences between ATS techniques presented in the literature.

First of all, ATS systems may be classified by the size of their input. We may have a system which target is to shorten a single document given in input (single document summarization, *SDS*) or to compress the important pieces of information from a set of multiple documents (multi-document summarization, *MDS*). Obviously, the MDS paradigm is not suitable for the case at hand, where we were interested in achieving an interpretation (the summary) on the classification of a single document.

Systems may also be divided by the nature of the summary. Some methods are

defined as *extractive* because they build summaries by extracting the most important sentences from the document. Others are called *abstractive* because they aim to generate a summary made by new (generated) sentences. Even if the abstractive paradigm can theoretically solve issues such as redundancy and information lost, because of the task complexity the research efforts focused more on the extractive kind. A third way is the *hybrid* one, which may be seen as a trade-off between the two paradigms. This kind of approach, as it can be guessed, combines the extraction of the most important sentences, on which the system will rely to generate the final summary. Since the abstractive phase relies only on those sentences, the quality of the summary may be of less quality than a pure abstractive summary, although it could be a good compromise.

Since our models focus on extracting sentences from the original document, it falls within the extractive paradigm. We could also define our models as deep learning-based (because, of course, transformers are deep neural networks models) and informative (because the extracted summaries contain important information of the original document). For an in-depth analysis of the nomenclature of the summarization systems, we suggest the reader to refer to [316].

7.1.3. Hierarchy in Transformer Models

One of the greatest limitations of the transformer-based models is to be limited to input of a fixed length of text, usually less than a few hundred tokens, even if they have the potential to learn longer-range context dependencies. This is due to the computational and memory requirements of the self-attention mechanism, which quadratically grows with the number of tokens in the sequence. The simplest approach to use for long document classification tasks with transformers is, therefore, the truncation of the document. This obviously may lead to a significant loss of information.

Trying to overcome this issue, some groups of researchers developed an extension of models like BERT. Such extensions usually exploit a hierarchical architecture, in which a classifier is built on the representations of some chunks of text obtained from a first transformer model. For example, in [317] two kinds of architecture were investigated: RoBERT and ToBERT. They build each model upon stacked representations retrieved in output from a first BERT layer. In RoBERT, a recurrency over BERT was implemented using an LSTM layer and two fully-connected layers. In ToBERT, another transformer was used over BERT, substituting the LSTM layer with a 2-layers transformer. At

a cost of a greater computational cost, ToBERT showed better performance on some evaluated tasks, especially on the one dataset consisting of longer documents. For both models, each document was divided into chunks counting 200 tokens, with an overlap of 50 tokens for consecutive chunks.

Inspired by this work, in [318] documents were divided into chunks of 512 tokens (with 50 overlapping tokens within consecutive segments), and an investigation on the merge method was conducted. In particular, the classification was based on the most representative vector (the one with the highest norm), on the average of all the vectors, and a representation built through a 1D convolutional layer.

Closer to our task, there is the work in [319], where HIBERT, a hierarchical transformer (again, based on BERT) was first pre-trained in an unsupervised fashion and then fine-tuned on a supervised extractive summarization task, where all the sentences of each document are labeled as belonging or not to the summary of that document.

Following this work, in [320] proposed to pre-train a hierarchical transformer model with a masked sentence prediction (in which the model is required to predict a masked sentence) and a sentence shuffling tasks (in which the model is required to predict the original order of the shuffled sentences). Then, also using the self-attention weights matrix (obtained by averaging over the heads for each layer and then averaging over the layers), the hierarchical pre-trained encoder is used to compute a ranking score for the sentences. The top-3 sentences are then used to constitute the summary. To the best of our knowledge, this last work is the closest to our, exploiting the attention weights of a hierarchical transformer model to generate a ranking useful to the extractive summarization. However, this last model was used with the aim to generate summaries in an unsupervised manner, while we aimed to collaterally generate summaries that explain the decision of a hierarchical model in a task of document classification.

7.1.4. Attention as Explanation

In the recent literature, various works proposed to analyze the attention patterns of the transformer architecture to have an insight on how such a model works. In [82], the author proposed a useful visualization tool, named $BertViz$. This tool provides an interactive interface to visualize attention weights between tokens for every attention head in every layer. Through this tool the author was able to find that some particular heads (in some particular layer) may capture lexical features, such as verbs and acronyms, or

may relate to the co-reference resolution, also showing the eventuality for such heads to encode gender bias. Another kind of visualization tool for the attention weights is the attention (heat-)map. Using these maps, the authors in [321] found patterns that are consistent with the previous ones. In detail, they divided the patterns into five categories: vertical (which mainly corresponds to attention to the delimiter tokens), diagonal (attention to previous or next word), a mix of these two, block (intra-sentence attention), and heterogeneous (said, no distinct structure). In this work, a heads and layers disabling study was also conducted, showing that in some cases a pruning strategy does not lead to a drop in performance (sometimes it even leads to an increase).

In addition to these two, other studies have been conducted showing that the self-attention heads allow BERT, as other transformer models, to capture linguistic features, such as anaphora [322], subject-verb pairings [323] (then extended by [324]), dependency parse trees in encoder-decoder machine translation models [325, 326], part-of-speech tags [327], and dependency relations and rare words [328].

However, in our study, we did not aim to reach an explanation of how the Transformer model deals with such features but to reach an interpretation of the document classification given by the model. Talking about this paradigm, various works focus on the weights of the attention layer in Transformers [329] or other kinds of networks, such as the recurrent or the convolutional ones, to highlight the words or n-grams in the text that are the most relevant for the decision.

Regarding the sentiment analysis task, authors in [330] observed a strong interaction between neighboring words visualizing the attention matrix of a Transformer-like network. Furthermore, in [331], the authors of the work discussed the use of attention scores from an attention layer as a good and less computationally burdensome alternative to external explainer models like *LIME* [77, 78] and *integrated gradients* [79] methods. However, the result of such a method is, again, to just highlight parts of the discourse for which the model seems to focus more. This kind of approach does not lead to an actual interpretative summary that may be more easily readable and, therefore, interpretable.

7.2. Materials and Methods

7.2.1. Data

To benchmark our models, we used the *IMDB Large Movie Review Dataset*. Such a dataset consists of $50K$ movie reviews written in English and collected by [304]. Those reviews (no more than 30 reviews per movie) were highly polarized, as a negative review corresponds to a $score \leq 4$ (out of 10), and a positive one has a $score \geq 7$. We downloaded the data through the *Tensorflow*[3] *API*. The data are already divided into two equivalent sets, one for training and one for testing (plus $50K$ unlabeled reviews that might be used for unsupervised learning, not used in this work). Each of the subsets presents a 50:50 proportion between negative and positive examples.

To assess the explainability of our methods we randomly extracted a total of 150 reviews, divided into two subsets, 50 from the training set and 100 from the test set. Documents were chosen by maintaining the proportion between the two classes, ensuring that both the models can correctly classify them. Four annotators were instructed to select the three most important (out of $N = 15$) sentences in each document. To make such a choice, the annotator is allowed to look at the sentiment of the document. To evaluate the agreement between the annotators, we calculated the so-called *Krippendorff's alpha*. First proposed by Klaus Krippendorff [332], to which it owes its name, it is a statistic measure of the inter-annotator agreement or reliability. The strength of this index is to apply to any number of annotators, no matter the missing data, and it can be used on various levels of measurement, such as binary, nominal, and ordinal. This measure may be calculated as in Equation (7.1)[4].

$$\alpha = 1 - \frac{D_o}{D_e} \in [0, 1] \tag{7.1}$$

where D_o is the disagreement *observed*, and D_e is the disagreement *expected* by chance.

Since Krippendorff's alpha is calculated by comparing the pairs within each unit, those samples presenting at most one annotation are eliminated. However, in this case, each sample (sentence) is automatically annotated as within the three most important sentences or not. Hence, such an elimination phase was not required. Values of α less then 0.667 are often discarded, while values above 0.8 are often considered as ideal [333,

[3] www.tensorflow.org/datasets/catalog/imdb_reviews
[4] https://github.com/foolswood/krippendorffs_alpha

334]. Anyway, except for $\alpha = 1$, we could say that there is no such thing as a magical number as a threshold for this kind of analysis, especially for tasks as much subjective as this one. In our case, $\alpha_{training} = 0.47$ and $\alpha_{test} = 0.61$.

7.2.2. Models

Here, we illustrate the two proposed architectures. To provide a visual explanation of them, we report the simplified schemes in Figures 7.1 and 7.2.

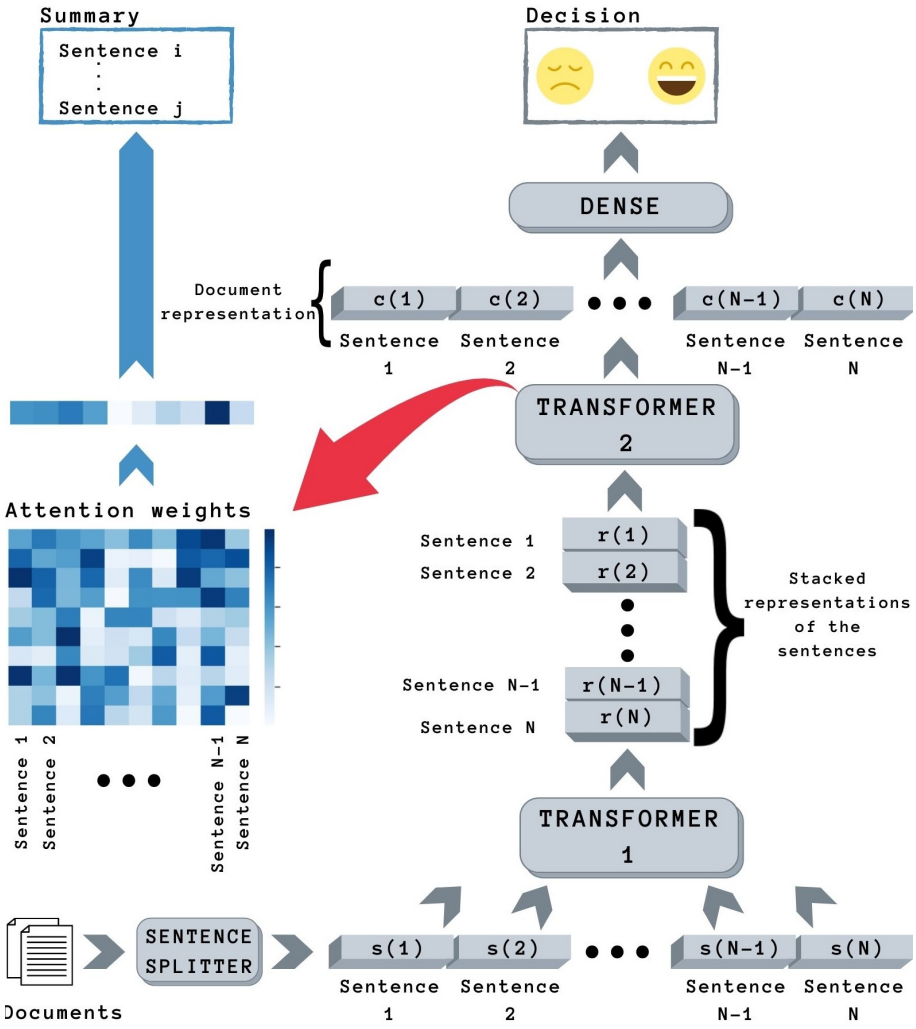

Figure 7.1.: Hierarchical transformers model

Explainable Hierarchical Transformer (ExHiT) The first model exploits a hierarchical architecture, consisting of two transformers ($T1$ and $T2$) in cascade (Figure 7.1). Be-

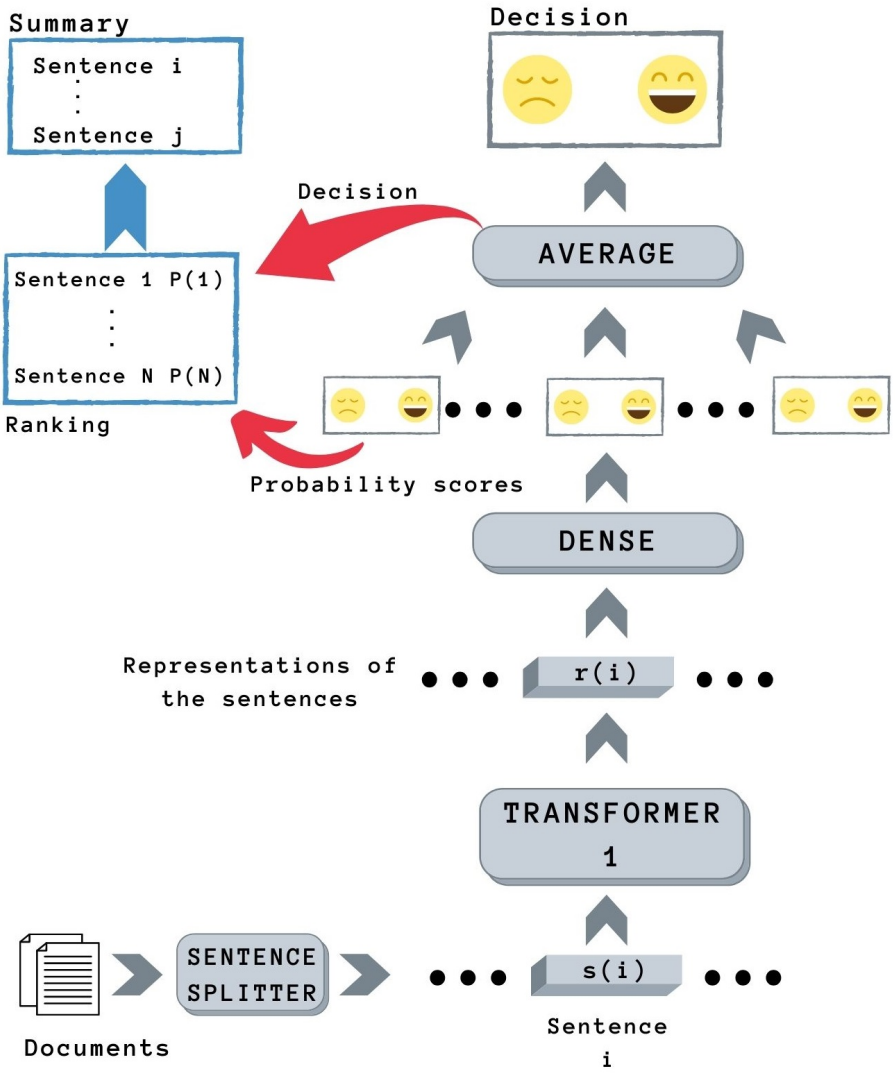

Figure 7.2.: Sentence classification combiner model

cause of its nature, we like to refer at this as the ***Ex**plainable **Hi**erarchical **T**ransformer* (*ExHiT*). The input of the first transformer is a sequence of t tokens, while the output is an embedding representation of that sequence. Each sequence represents one of the N sentences $\{s_1, \cdots, s_N\}$ in which the document is divided. If a document can be divided into just $m \leq N$ sentences, then $N - m$ empty sentences (just the special tokens) are added to the document. After $T1$ has elaborated the N sequences, the new generated representations $\{r_1, \cdots, r_N\}$ are stacked together to become the input of $T2$. $T2$ then outputs a contextual representation c_i for the i-th sentence that depends on the other sentences ($c_i = f(r_1, \cdots, r_N)$). By merging these contextual representations we obtain an unique document representation $d = U(c_1, \cdots, c_N)$. In this work, we investigated the

following merging strategies:

- By concatenation: $U(.) = Concat(.)$;
- By averaging: $U(.) = Avg(.)$;
- By masked averaging: $U(c_1, \cdots, c_N) = Avg(c_1, \cdots, c_m)$ with $m \leq N$, for which $\{s_{m+1}, \cdots, s_N\}$ is the set of the added empty sentences;
- By the application of a Bidirectional LSTM: $U(.) = BiLSTM(.)$.

Then, vector d is given as input to a classification layer. In this work, such a layer consists of a two-units fully-connected dense layer with the softmax activation for the binary classification task. Other than the contextual representations, we were able to retrieve from $T2$ also the self-attention weights for each head of each layer inside the transformer itself. To give more importance to the interpretability of the model instead of the performance, $T2$ consists only of two layers and just one head per layer. In this way, it is easier to extract valuable information. By averaging the attention weights associated with a specific sentence, we extracted the score of that sentence. The sentences are ranked through such a score, and the most important ones are then selected to provide an extractive summary of the document. Such a summary serves then as the explanation of the model decision.

Sentence Classification Combiner Model (SCC) This second model has a simpler architecture (Figure 7.2), requiring just one transformer model in its pipeline. The input of this transformer is again a sequence of t tokens, i.e., the single sentence s_i. Again, its output is a new representation r_i of that sentence. Such representation is given in input to a *dense* layer to classify the sentiment of the sentence, outputting two probability scores, one for each class. Then, the negative scores are averaged together, and the same for the positive ones, to get a final rating for each class. The prediction of the overall document sentiment will be given by whoever has the greatest final score. Knowing the decision of the model, the sentences are ranked by the inherent probability score. Then, the most relevant ones are extracted to build the summary of the document, serving as an explanation of the model decision.

Parameters In the following, we listed the main features of the two models used in the experiment's session:

- **T1**: For a fair comparison, the first transformer model was the same for both the architectures; we opted to use the pre-trained version of RoBERTa [50];
- **T2**: We used a transformer with two layers, one head per layer; this choice was motivated to facilitate the explainability phase;
- **N**: The maximum number of sentences per document was set to 15; by this way, we ensured that the 75% of the training documents were elaborated in their entirety;
- **t**: The maximum number of tokens per sentence was set to 32, comprehensive of the two special delimiter tokens; by this way, we ensured that the 75% of the training sentences were elaborated without being truncated.

In addition to the two models, we implemented a pre-processing phase consisting of the replacement of the tokens '<\br><\br>' with the newline character, and, obviously, a sentence splitting step. We used the sentence tokenizer provided by NLTK. Furthermore, for documents that do not reach N number of sentences, empty sentences (consisting of just the special tokens) were added up to N. Similar reasoning was applied to sentences that do not reach the t number of tokens: in these cases, the sequences were zero-padded on the right, and an attention mask was applied.

7.3. Experiments

The *SCC* model was trained on the single sentence classification task, with a batch size of 240 sequences. Thus, the transformer together with the *dense* layer has been trained to classify the sentiment of every single sentence. After that, the *average* layer has been added on top of the trained model to perform the document classification (and its explanation) in the test phase. For what concerns the *ExHiT* model, the experiments followed two different fashions, always maintaining a batch size of 8 documents.

7.3.1. Joint Training

In this kind of experiment, the entire model was jointly trained on the document classification task. In this conceptualization, the weights of both the two transformers, the *dense* layer and, eventually, the merging layer (*BiLSTM*) were allowed to be updated during the training phase.

7.3.2. Ablation Study

In this kind of study, a model is modified by adding or removing some features from its architecture. In a first experiment, we added to the model the following features:

- (Sinusoidal) Sentence Positional Embedding (SPE) for T2: following the original works on the transformers architectures, we added the positional embedding to the sentence embedding in input to the second transformer;
- Sentence Masking (SM) for T2: following the principle of the attention mask for the padded token of the sequences in input to the first transformer, we applied an attention mask on the empty sentences added at the bottom of the document.

These two items were added under the hypothesis that encoding the relative position between the sentences and masking the empty sentences may improve the model performance. In a second experiment, we froze the weights from the first transformer also during the training. This experiment was conducted to investigate the following hypothesis: by freezing its weights, no knowledge can be learned by T1 during the training phase; thus, this configuration should force the second transformer to learn the most important features from the document to perform the task it is trained for. Even if it could lead to degradation in the sentiment classification performance, if this assumption was confirmed, it could potentially lead to improving the explanation performances.

The ablation study experiments were conducted using the *ExHiT* model implementing the concatenation merging strategy.

7.4. Results

The proposed models were evaluated for both sentiment analysis and explainability outcomes. In Table 7.1, we reported the sentiment analysis results achieved in terms of accuracy, and precision, recall, and F1-score per class. For the *ExHiT* model, various proposed merging strategies were tested. As the accuracy column highlights, changing the merging strategy does not significantly affects classification performance.

Following the same structure, in Table 7.2 we reported the explainability outcomes in terms of precision averaged over all the documents. The performances are reported for different annotators' agreements, i.e., we built summaries by grouping the sentences for which at least one, two, or three out of the four annotators judged them among the most

Table 7.1.: Sentiment analysis results in terms of accuracy, and precision, recall, and F1-score per class.

Model	Merging Strategy	Accuracy (%)	Precision (%)		Recall (%)		F1 (%)	
			Neg	Pos	Neg	Pos	Neg	Pos
ExHiT	Concatenation	92.59	90.97	**94.34**	**94.56**	90.62	92.73	92.44
	Average	92.35	92.18	92.51	92.54	92.15	92.36	92.33
	Masked Average	92.77	92.07	93.49	93.60	91.94	92.83	92.71
	BiLSTM	92.34	90.97	93.80	94.01	90.67	92.47	93.06
SCC	-	**93.51**	**95.42**	91.75	91.40	**95.62**	**93.37**	**93.65**

important ones. This implies that some annotators' summaries of the j-th document may contain more than three sentences ($N_j > 3$, especially in the first case) or less than three sentences ($N_j < 3$, especially in the latter case). Therefore, we extracted the first N_j sentences in the machine's ranking and evaluated the overlap of these summaries with the annotators' ones. The formula of the introduced explainability metric is reported in Equation (7.2):

$$Precision = \frac{1}{D} \sum_{j}^{D} \frac{TP_j}{N_j} \in [0, 1] \tag{7.2}$$

where D is the number of documents, N_j is the number of sentences in the annotators' summary, and TP_j is the number of well-selected sentences in the system summary of the j-th document. Documents for which N_j was equal to 0 were excluded from the computation. This may happen, in particular, where an agreement of at least three annotators was required. About the *ExHiT* performance, the results of the best layer are reported. In general, the ranking from the first layer slightly outperformed the rankings from the last layer and the rankings obtained by averaging both layers. Furthermore, the empty sentences were removed by the machine rankings.

For what concerns the ablation study on the *ExHiT* model, we combined the configurations described in Section 7.3.2, exploiting only the concatenation merging strategy. We reported the results of these experiments in Table 7.3, both in terms of accuracy and explainability precision. Again, the explainability performances are related with the first layer in general, except when the last layer or an average between both presents (slightly) better summaries. However, in the models implementing the sentence masking for the second transformer we found out a greater degree of agreement between the layers, and it was not rare to see the precision scores from the first layer only slight

Table 7.2.: Explainability performance in terms of precision (averaged over all documents) for different annotators agreements, evaluated on both the annotated documents from training and test sets. For *ExHiT* model the performances from the first layer are reported, except when the rankings from the last layer [1] or from the average of layers [a] have shown better results.

Model	Merging Strategy	Agreement at Least 1 Precision (%)		Agreement at Least 2 Precision (%)		Agreement at Least 3 Precision (%)	
		Test	**Train**	**Test**	**Train**	**Test**	**Train**
ExHiT	Concatenation	53.82%	55.88%[a]	49.15%	45.00%	46.63%	46.45%
	Average	58.04%	57.82%	50.42%	45.92%[1]	45.29%	41.84%
	Masked Average	53.15%[a]	55.79%	45.97%[a]	44.92%	40.66%	39.80%
	BiLSTM	55.51%[a]	55.85%	49.05%[a]	45.24%[a]	43.38%[a]	39.95%
SCC	-	**70.74%**	**65.61%**	**65.22%**	**57.83%**	**55.22%**	**47.52%**

higher than the scores retrieved from the last layer (or the scores obtained by averaging the attention weights from both layers).

Table 7.3.: Ablation study outcomes for the *ExHiT* model, both in terms of accuracy and explainability precision. *SM* stands for *sentence masking*, *SPE* stands for *sentence positional embeddings*. *Frozen T1* indicates that the the weights of T1 were frozen during the training. The model is intended to implement the concatenation merging strategy. As in Table 7.2, results from the last layer or from the average of layers are indicated with the apices [1] and [a], respectively. Otherwise, the reported results are intended to be related with the first layer.

Model	Accuracy (%)	Agreement at Least 1 Precision (%)		Agreement at Least 2 Precision (%)		Agreement at Least 3 Precision (%)	
		Test	**Train**	**Test**	**Train**	**Test**	**Train**
ExHiT	92.59%	53.82%	55.88%[a]	49.15%	45.00%	46.63%	46.45%
+ SM	92.51%	67.24%	68.27%	59.82%	56.17%	54.88%	**57.09%[a]**
+ SPE	92.37%	64.34%	65.35%	58.13%	56.33%	52.19%	56.38%
+ SM + SPE	92.67%	70.27%[a]	**69.11%[1]**	63.65%	**63.50%**	**55.56%**	55.67%[a]
Frozen T1	89.50%	63.43%	68.16%[a]	52.78%[a]	56.00%	44.11%[a]	48.23%[a]
SCC	**93.51%**	**70.74%**	65.61%	**65.22%**	57.83%	55.22%	47.52%

In Tables 7.2 and 7.3, we compared the different models' explainability with what we may call the explainability precision. Such a metric may take place with a priori annotations, i.e., the annotations are made on the original documents. To conduct a more in-depth analysis on the explainability error, we are also proposing a new metric that takes advantage of a posteriori annotations, i.e., the annotations are made on the (N=) 3-sentences summaries retrieved by each model. Here, the annotators were instructed to annotate each sentence as (1), negative (−1) or neutral (0) depending on

the polarity of the document it lets them understand. The final score of the model exploits a sort of *Mean Absolute Error* (*MAE*) for discrete variables. Hence, the total score is computed by following Equation (7.3)

$$1 - \frac{1}{2}MAE = 1 - \frac{1}{2D}\sum_{j}^{D}\frac{1}{N_j}\sum_{i}^{N_j}|c_j - s_{j,i}| \in [0,1] \tag{7.3}$$

where D is the number of documents, N is the number of sentences per summary, c_j is the predicted class of the j-th document, and $s_{j,i}$ is the annotated score of the i-th sentence (of the j-th document), and $\frac{1}{2}$ is a corrective factor to map the range of the *MAE* function (and, therefore, of the score) into an interval of $[0,1]$. In particular, in our case we fixed N_j to be equal to $N = 3$, also excluding the documents with less than 3 sentences from the computation of the total score). In this case, the previous equation may be rewritten in the form of the following equation:

$$1 - \frac{1}{2}MAE = 1 - \frac{1}{2DN}\sum_{j}^{D}\sum_{i}^{N}|c_j - s_{j,i}| \in [0,1] \tag{7.4}$$

The contribution of each sentence $-\frac{1}{2}|c_j - s_{j,i}|$ to the total score is, therefore, equal to 0 if the prediction and the sentence score belong to the same class, -1 if they belong to opposite classes, and $-\frac{1}{2}$ if the sentence is annotated as neutral. Thus, the total score is a real number that lies in the range between 0 (dramatic extreme case, in which all the sentences belongs to the class opposite of the prediction) and 1 (desirable extreme case, in which all sentences belongs to the same class of the prediction). In particular, if the score is equal to $\frac{1}{2}$, it means that all the extracted sentences are evaluated as neutral by the humans (or that the number of well-ranked sentences counterbalances the number of the sentences classified as belonging to the class opposite of the prediction). Thus:

- If $1 - \frac{1}{2}MAE < \frac{1}{2} \Rightarrow$ the model performance is worst than if it chose all neutral sentences;
- If $1 - \frac{1}{2}MAE > \frac{1}{2} \Rightarrow$ the model is going better than if it chose all neutral sentences.

To avoid any bias for the annotators deriving from the prediction of the model or from the other sentences of the document, the predicted class was obscured and the sentences of all the documents were shuffled together. The annotations were performed just for three of the models trained in this work, as reported in Table 7.4. In particular, the results about the *ExHiT*-based systems come from the attention head of the first

layer of the second transformer.

Table 7.4.: Explainability performance in terms of the proposed score, reported in Equation (7.4), and percentage of summary's sentences annotated as neutral. The *ExHiT* models are intended to be implemented with the concatenation merging strategy, and the summaries built by the first layer are analyzed.

Model	$1 - \frac{1}{2}MAE$ (%)		Neutral Rate (%)	
	Test	**Train**	**Test**	**Train**
ExHiT	78.33%	74.67%	26.00%	35.37%
+ SM + SPE	86.50%	82.67%	13.00%	**17.69%**
SCC	**92.67%**	**88.67%**	**11.56%**	19.05%

7.5. Discussion

By analyzing Table 7.1, the *SCC* model seems to achieve a slightly better overall performance. However, it is interesting to notice that *SCC* results are particularly good for the precision for the negative class and the recall for the positive one while achieving the worst performances for their counterpart metrics, for which the best results are obtained by *ExHiT* using the concatenation merging strategy. About Table 7.2, the *ExHiT* explainability results are lower than those achieved by *SCC*, with respect to all the merging strategies. This outcome may be due to an influence of the task on the two models: it may be noticed that the task the second model accomplishes is closer to the one performed by the annotators. It may, therefore, help the *SCC* model in the explainability task. Furthermore, the average merging strategy leads to better performance than the masked one, especially with respect to the test set ($\sim$+5%). This seems to suggest that masking the empty sentences from the average combination after the elaboration of the second transformer does not help the model to better understand the task.

For what concerns the ablation study conducted in this work, the obtained results are very interesting (Table 7.3), in particular, for what regards the explainability performance. In fact, while the sentiment classification did not show significant improvements in terms of accuracy, *ExHiT* models implementing the empty sentence masking have shown significant improvements in terms of explainability precision, achieving results comparable with the *SCC* model. In some cases, it reaches even better outcomes, especially when it is combined with the introduction of the sinusoidal sentence position embeddings, a strategy that has shown explainability improvements even when imple-

mented alone. Furthermore, exploiting a qualitative analysis of the outcomes of the various models, it has been observed how the *ExHiT*-based systems "suffer" from the noisy empty sentences added to achieve the maximum number of the sentence required by the architecture, with the exception of the models implementing the sentence mask. This behavior is highlighted in Table 7.5, reporting the sentences from the same document ranked by two *ExHiT* models, one not implementing the sentence mask and one that does it, respectively. In the example reported here, it is easy to notice the presence of the empty sentences among the first positions of the ranking built by the former, while they are always ignored by the latter model itself and thus put on the bottom part of the constructed ranking.

Table 7.5.: Example of document summary generated by two *ExHiT* models, one implementing sentence masking (**right**) and the other without the sentence mask (**left**). The *index* columns indicate the position of each sentence in the original document.

Index	**ExHiT**	**Index**	**ExHiT + SM**
0	This film was a surprise.	0	This film was a surprise.
6	Jealousy, sexual tension, incest, intrigue, [...]	6	Jealousy, sexual tension, incest, intrigue, [...]
11	Even though her strength and lack of illusion [...]	1	The plot synopsis sounds kinky [...]
13		8	However, I wanted to clarify a point [...]
14		10	The attractive female slave successfully resists [...]
4	The child takes him to the girls [...]	9	I find that there is one.
10	The attractive female slave successfully resists [...]	3	There is that opening scene where [...]
3	There is that opening scene where [...]	11	Even though her strength and lack of illusion [...]
5	He takes advantage of the situation [...]	2	I didn't know what to expect.
7	I've read the other comments here and find little to disagree with.	5	He takes advantage of the situation [...]
2	I didn't know what to expect.	7	I've read the other comments here and find little to disagree with.
12	She, more than any of the other women [...]	4	The child takes him to the girls [...]
9	I find that there is one.	12	She, more than any of the other women [...]
8	However, I wanted to clarify a point [...]	13	
1	The plot synopsis sounds kinky [...]	14	

These considerations seem to suggest that the sentence mask is essential to filter out the noisy empty sentences and, therefore, better understand the task. However, with or without sentence masking, the performances of the classification task of the two kinds of models are not so far from each other. Thus, we hypothesized that, in absence of the sentence masking, the model captures other kinds of information from the text and learns how to use them to perform the sentiment analysis task. Exploiting these features thus leads to good performance in the main task, but being inherently less interpretable for humans it is inevitable for the model to reach worse explainability performance.

The performed ablation study reports interesting results when T1 weights are frozen during the training phase. In this case, the model has shown good improvements in terms of explainability precision. This seems to confirm our hypothesis that freezing the first transformer prevents it from adapting to the task during training and then forces

the second transformer to boost its ability to extract important information from the sentences. In fact, the results in Table 7.3 show that this trick can in part compensate for the absence of sentence masking, at the cost of ∼3 percentage points on the classification accuracy.

For what concerns explainability, in Table 7.4 results in terms of *MAE* are reported for some models. We proposed this kind of metric after visualizing models' outcomes. Figure 7.3 shows an example of document annotation performed by the annotators, in which darker background means higher relevance in the document, computed as the average of agreement among the annotators. Figure 7.4 shows the annotations performed by the *SCC* model on the same example. Again, darker background means higher relevance, in this case in terms of the probability scores.

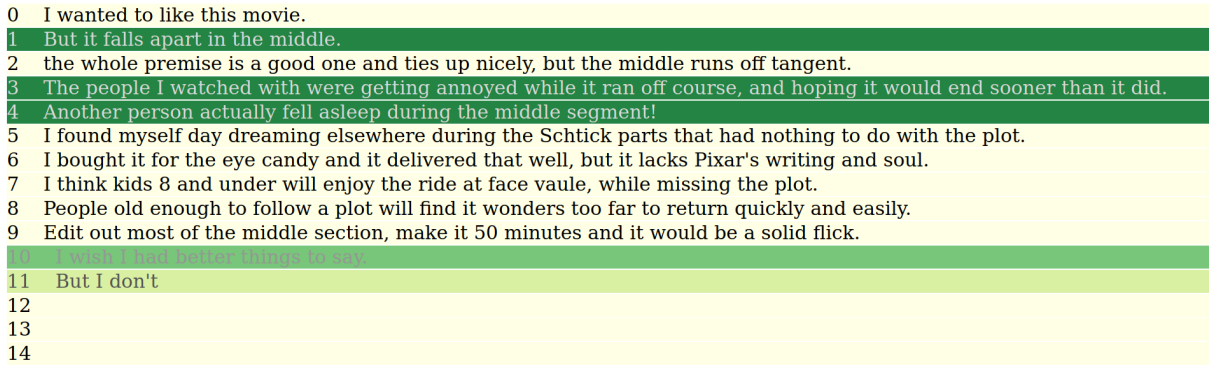

Figure 7.3.: Example of document annotation performed (a priori) by the instructed annotators.

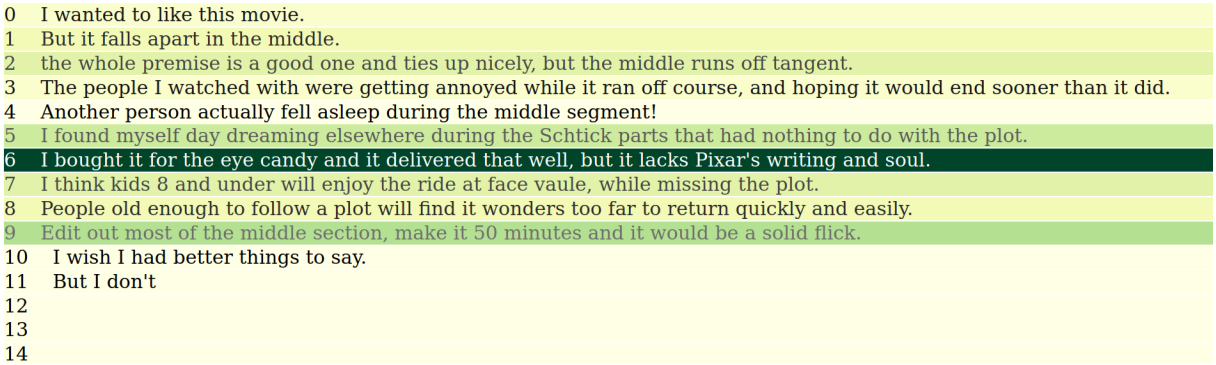

Figure 7.4.: Example of document annotation performed by the *SCC* model during the classification.

As it can be seen, if we consider the case of agreement of at least three out of the four annotators, the explainability precision would be null. In fact, the annotators' summary

would be built by extracting the sentences numbered as {1, 3, and 4} while the model's summary would consist of the sentences numbered as {5, 6, and 9}. However, at a further analysis, it seems clear that the second set of sentences contain a negative sentiment, as well as the first set. Thus, the introduction of such a posteriori metric was necessary to perform a fairer comparison between the models. Unfortunately, being a posteriori metric brings the drawback to perform human annotations on the outcomes of each model we want to compare. For this reason, this analysis was limited only to three of the developed models. The outcomes in terms of this metric (Table 7.4) still show the *SCC* model to outperform the other(s). Very interestingly, for the test dataset, it achieves a very high score, greater than 92%. This seems to suggest that this model not only can achieve a near state-of-the-art accuracy in the sentiment analysis task but also provides a very accurate extractive summary as an explanation of the predicted class. However, to compute this metric the introduction of a new class, the neutral one, was necessary to perform the annotations. The reason is, of course, we cannot exclude a priori that some summary extracted from the template may contain uninformative sentences with respect to the sentiment of the full review. In particular, for both kinds of models, many of the sentences classified by annotators as neutral are excerpts of plot narration. For example, the sentence

> *But success has it's downside, as Macbeth soon finds out, when he has to go to hideous lengths to protect his murderous secret.*

is simply a description of a plot passage in the movie *Macbeth - The tragedy of ambition* and contains no information about the sentiment of the review provided by the user and, therefore, would be impossible for an annotator to classify as black or white. Another kind of sentence extracted by the models is sentences that may gain a sentiment sense only if seen together with the previous or next parts of the document. For example, the sentence

> *You'll be glad you did.*

has no particular sentiment when picked alone and gives no clue about the polarity of the document. In fact, it is licit for an annotator to wonder: *I'll be glad I did what?*. Thus, this sentence alone does not give good hints to guess the sentiment nature of the document. However, if you look at this sentence inside its context

> *Do yourself a favor and avoid this movie at all costs. You'll be glad you did.*

it can be easily noticed that it enforces the negative sense of the previous part of the

discourse. The annotators reported this kind of behavior for the *ExHiT* model, in particular, and this seems to be confirmed by the last two columns of Table 7.4, where a greater percentage of neutral annotations is reported for both training and test set with respect to the *SCC* model. This seems to suggest that the *ExHiT* model it is less suited to perform single sentence-extractive summaries, because of a more contextual understanding of the document classification task. However, the *ExHiT* version implementing both the sentence masking and positional embeddings has been reported to less show this behavior. In fact, its neutral percentage is way lower than the simpler version. In the case of the training set, it is even lower than the *SCC* system. Furthermore, also the proposed score presents a significant improvement. These outcomes suggest, once again, how these components are of great importance to the interpretability of the hierarchical model.

Both the proposed models have achieved good classification results, not so far from the works at the state-of-the-art on the *IMDB* dataset, while also performing an explanation in the form of a summary. The explainability component of the models, achieving good results, as well as the classification one, is a feature that may become essential in several applications. For example, while sentiment analysis may help to mark customer messages and reviews, the explainability part may be helpful to get quick insights about the strengths and the weaknesses of some product or service. Furthermore, both underlying architectures allow their easy adaptation in any document classification task (e.g., topic classification), and may be applied to any language: the only restriction is to use the suitable pre-trained model as *T1*, i.e., a Transformer model that has been pre-trained on the task language (or, at least, in a multilingual fashion). This is an interesting point of view for the next research works to focus on: by using the proposed models it would be possible to achieve the sense of their reasoning and ease the individuation of wrong classified samples. Furthermore, in other scenarios, *ExHiT* may outperform the *SCC* model thanks to its ability to get more insights from the context of the other sentences in the document. Sentiment analysis is, in fact, a task that particularly relies on the lexical meaning of the individual sentences, thus being less influenced by the entire context.

Another attractive idea to follow is to go in-depth with the analyses in the ablation study for the *ExHiT* architecture. We have shown how masking the empty sentences and adding the positional embedding may play an important role for the model, and how freezing the first transformer during training may force the higher part to learn more interpretable features. Future research may extend this kind of study by evaluating

the performance (both in terms of classification and explainability) achieved by deeper models, e.g., adding more layers and self-attention heads to the second transformer. Plus, the explainability at a finer granularity (the tokens level) may be explored by investigating the attention weights of the first transformer for both kinds of architectures, e.g., by highlighting the most important words in each sentence, a strategy already explored in the literature.

8. Breaking Bert: One Sentence makes the Difference

As already stated, since the advent of the Transformers architecture [335], Pre-trained, large neural Language Models (PLMs) such as BERT [23] have become quite ubiquitous in the NLP literature. An advantage of using these models is that their weights can be updated to address multiple downstream tasks with good results [336]. Usually, fine-tuning these pre-trained language models on a downstream task shows improvements in performance with respect to training these models from scratch. A common belief is that pre-training on large corpora allows them to learn linguistic knowledge that is then exploited during fine-tuning. Past works showed how BERT captures syntactic information [61, 62] and other linguistic structures [63, 64], capturing the steps of the traditional NLP pipeline across subsequent layers [65].

Although these results push the popular belief that capturing such information from large upstream corpora underlies the success of these models, such a hypothesis remains unproven. On the contrary, recent studies suggest that such benefits are induced by mechanisms in pre-training not yet fully elucidated. The principal consideration regards the quantity of pre-training data. Recent works showed good performance with models pre-training on small amounts of data before fine-tuning [66, 67]. Other works regard the quality of the pre-training data. Some have shown the benefits in fine-tuning after pre-training with noisy data, such as shuffled texts [68, 69], with data from other domains [67], or even from no-human languages [70], such as amino acid sequences, java scripts, and randomly generated texts. Similar results were obtained by pre-training LSTMs on music [337]; or in different fields such as computer vision [338, 339].

However, the presence in pre-training of texts outside the target test distribution can degrade performance on the downstream task [340]. This limitation can be mitigated when the generic PLMs are adapted to the target test data. An effective and relatively fast way to do this is to *further pre-train* (FPT) with domain- or task-specific data these

broad-coverage models before task-specific fine-tuning [341, 342, 343, 344]. This procedure has led to a family of specialized pre-trained models, showing improvements against their general-domain versions in a range of domains, from medicine (*BioBERT* [57]) and clinics (*ClinicalBERT* [58]), to finance (*FinBERT* [345]) and law (*LEGAL-BERT-FP* [346]), passing by social media (*Rob-RT* [347]) and hate speech (*HateBERT* [348]), among others.

Further pre-training may be seen as an intermediate training phase, between the original pre-training and the downstream fine-tuning, with the usual aim of adapting a general model to a specific domain. While FPT is by now common practice, there does not seem to be a clear picture of which conditions make it fully successful. For example, Mehri et al. [349] and Qiu et al. [350] show that models adapted for dialogue tasks do not consistently outperform their generic counterparts. Gururangan et al. [343] show that using a small amount of data from the task-specific training set to perform FPT results in a more successful fine-tuning step than when using a much larger amount of domain-specific data which is not the actual training set. This raises the question of how to best choose and balance quality and quantity of data for FPT. A similar question is raised by Rietzler et al. [351]: they observe that the size of data used in FPT impacts downstream (sentiment analysis) performance in very different ways according to the domains. For some domains, FPT data only starts to make a difference when extremely large amounts are used, while for other domains smaller amounts are sufficient. In any case, Zhu et al. [352] show that if the size of fine-tuning (FT) data is large enough, the impact of FPT on the final performance on the downstream task is negligible.

With the work presented in this chapter we thus aim at further unpacking (i) the impact of data size and training parameters on FPT, in terms of the measurable difference between a base model and its FPT-ed counterpart; and (ii) the interplay between such differences and the performance of FT-ed models on a downstream task. In other words, we investigate under what conditions FPT yields measurable changes in a large model, and how these changes influence the performance of subsequently fine-tuned models. To this end, we implement a set of experiments where we controlled for the amount of data (number of sentences), the number of epochs, and the learning rate used to FPT a large model, and for the task-specific FT data size. Across all combinations (see Section 8.1 for the full overview of the values and combinations we tested), we observed two sorts of unexpected behaviors: (i) substantially different downstream performances when FT with identical settings the original and the FPT models did not correspond to measurable dif-

ferences in the internal model representations; (ii) FPT with just one single sentence, for one single step and with a low learning rate triggers remarkable performance differences when the models are FT-ed for the downstream task, while the internal representations of the pre-trained models (before and after FPT) are measured as minimally different. This latter observation is rather puzzling, especially considering that the original pre-trained models have been exposed to massive amounts of data; hence the question: *if a minimal, barely detectable update of a large model has consequences on fine-tuning, what can be said about model stability and reported results?* Figure 8.1 graphically represents the experimental settings discussed in this paper. The picture generalises to the whole set of experiments if the size of the data for FPT is changed from "One Sentence" to any other value we used.

The obtained results are counter-intuitive: being already trained on an enormous amount of data, one can imagine one sentence should not make such a difference. Very recently, other works have disclosed numerical instabilities of training neural networks with gradient-based algorithms [353, 354] to which our results may relate. In any case, our work poses new questions on the functioning of BERT-like models and the nature of the FPT paradigm: deepening this research line may help the community's understanding of these technologies. While other works have shown the frailty of fine-tuned BERT-based models against adversarial attacks [355, 356, 357, 358, 359], for the best of our knowledge, this is the first work assessing the impact of one sentence during a training phase on BERT-like models.

8.1. Experiments

For our experiments we use the base BERT model[1] [23]. Although newer models are available, BERT makes a good candidate for this work since it has been substantially used in FPT and fine-tuning procedures, and its manageable size allows us to run all of our configurations with reasonable time and computation. For each experiment, we follow the same two-step approach (Fig. 8.1): first, we further pre-train the original BERT model ($\boldsymbol{B_0}$), using n sentences; in this chapter we describe, in particular, the minimal case where $n = 1$, which already shows surprising behaviors. We perform further pre-training for one step using only the Masked Language Modeling (MLM) training objective, following the same setup as in [23] for the percentage of tokens to

[1] https://huggingface.co/bert-base-cased

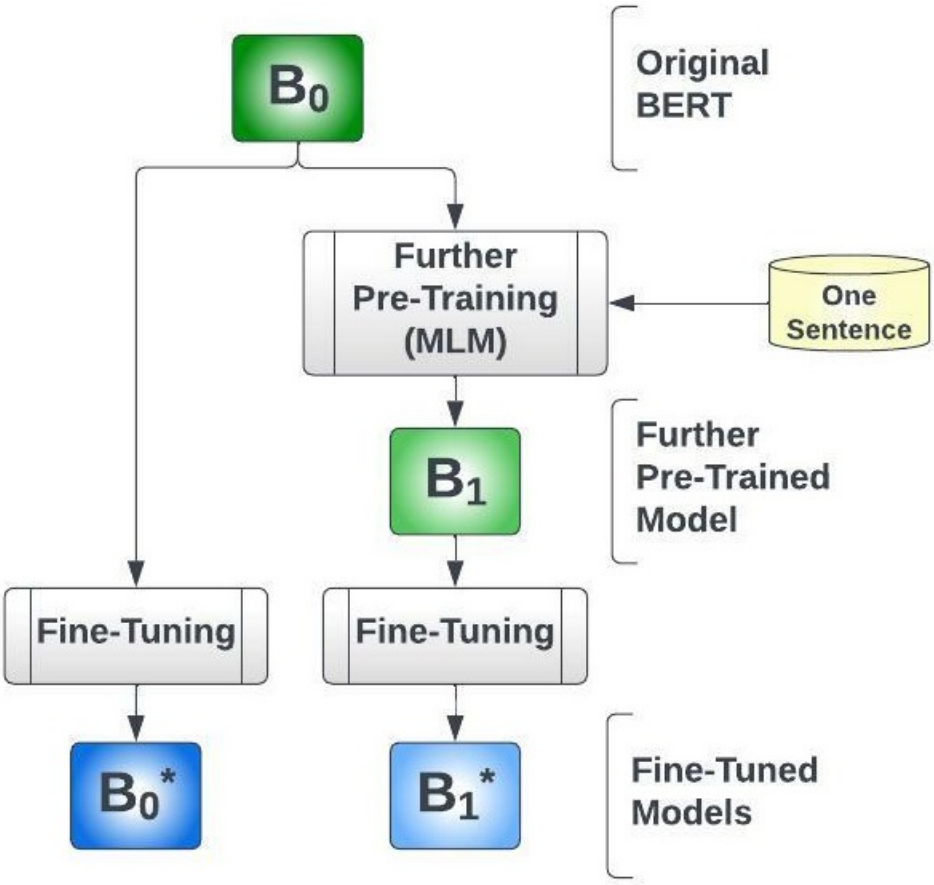

Figure 8.1.: Experiment workflow. We indicate the original BERT as $\boldsymbol{B_0}$ and any further pre-trained model as $\boldsymbol{B_1}$. The '$*$' denotes their fine-tuned versions.

be masked. The outcome of this first step is a new, further pre-trained model ($\boldsymbol{B_1}$). After pre-training, we go through a fine-tuning stage on a specific downstream task using the same amount of data and hyper-parameter settings for the two pre-trained models, $\boldsymbol{B_0}$ and $\boldsymbol{B_1}$, obtaining the two corresponding fine-tuned models $\boldsymbol{B_0^*}$ and $\boldsymbol{B_1^*}$. As downstream task we take binary sentiment classification. As aforementioned, while we performed several experiments with several combinations of parameters, we obtained surprising results with even minimal FPT phase. While we focus on the single-sentence, single-epoch experiments for the rest of the chapter, we report all the parameters and their values taken into account in Table 8.1.

Further pre-training settings For the pre-training phase, we controlled for two variables: data and learning rate. To avoid effects of domain- and task-adaptation, we ensured that the data comes from a different distribution of that used for the fine-

	Further Pre-Training (FPT)	**Fine-Tuning (FT)**
Data size	{1, 10, 100, 1000}	{100, 200, 500}
Training Epochs	{1, 10, 100}	early-stopping
Batch Size	1	8
Learning Rate	$\{1e^{-10}, 5e^{-05}\}$	$5e^{-05}$
Optimizer	Adam	Adam
Scheduler	const	const

Table 8.1.: List of parameters and their values taken into account during the experimental phase.

tuning task (see also Section 6.2). Furthermore, to ensure we could assess if the nature of the pre-training sentence might impact the downstream task, we collected three single sentence datasets (positive, neutral, negative), to which we also added a dataset consisting of a single nonsense sentence. For the learning rate, we employed a value equal to $1e^{-10}$, which is several orders below the values (e.g., $1e^{-04}$ or $1e^{-05}$) usually employed in domain adaptation FPT (e.g., [57, 58, 343]).

Fine-tuning settings In fine-tuning, we control data size and composition, as well as the randomness of initialization and data order. Specifically:

- to avoid hiding the effects of the further pre-training on the downstream task [352], we used a relatively small number of documents; plus, we varied the size to observe its impact, i.e., $n = \{100, 200, 500\}$;
- to avoid the impact of the randomness of initializing the task-specific classifier and of picking the training samples, we initialized five different classifier layers and selected five different training sets[2] (for each sample size n); this led us to fine-tune a total of 25 versions of $\boldsymbol{B_0^*}$ and $\boldsymbol{B_1^*}$ for each fine-tuning sample size.

Finally, to limit the effects of the hyper-parameters during FT, we fixed the optimizer, the learning rate, and the batch size to be constant during each tuning and across all the experiments[3]. We also applied early-stopping (no increase in validation accuracy after 5 epochs).

[2]Since the order of the training samples considerably impacts the tuning process [360], we fixed the order of each fine-tuning set, too.
[3]$opt = Adam$, $lr = 5e^{-05}$ and $bs = 8$

8.2. Data

Regarding the two different training phases, we considered two separated dataset.

Further pre-training data In each experiment, the pre-training dataset consists of a single sentence. To retrieve different sentences, we exploited the *ROCStories* dataset [361]. The choice of such data is two-fold, being a dataset unseen during the original pre-training and belonging to a general domain, we mitigated the effects of already seen samples and those given by an adaptation to the downstream domain. Each document in the dataset consists of 5 sentences, from which we retrieved the longest one to minimize the probabilities of presenting input texts without masked tokens since the masking process as described by [23] is inherently random for each sentence.

To take into account the nature of the sentences with regard to the downstream tasks, we retrieved three sentences with different sentiments, i.e., positive (pos), neutral (neu), and negative (neg). To assess the sentiment of the sentences, we employed the RoBERTa-based model trained on a three-classes sentiment classification task[4] by [347]. Then, we selected the sentences with the highest scores for the given class. After a human assessment of their sentiment, the three datasets we gathered are:

- **pos (+):** *They had a really great day and couldn't wait to go back again!*
- **neu (/):** *He will pitch at home against the Miami Marlins.*
- **neg (-):** *I felt sick as it tasted way too bland and mushy.*

Furthermore, to analyse the contribution of unstructured text, we built a nonsense sentence by replicating the token *'the'*, 32 times.

Fine-tuning data For the downstream task, we employed the *IMDB* dataset [304], which contains positive and negative reviews of movies. To investigate the effects of the downstream training set size, we retrieved datasets in three different sizes, i.e., with 100, 200, or 500 documents. Plus, to investigate the effect of the sample composition, for each FT sample size, we collected five different data combinations, resulting in a total of fifteen downstream tuning sets. For validation, we carved 5,000 documents out of the original training set. We used the entire test set (25,000 samples) for final evaluation.

[4] https://huggingface.co/cardiffnlp/twitter-roberta-base-sentiment

8.3. Evaluation

To assess the impact of our experiments, we first conducted an intrinsic evaluation of the FPT models. For a direct comparison of the original $\boldsymbol{B_0}$ and the corresponding $\boldsymbol{B_1}$ models, we calculated two metrics on the representations of the *[CLS]* tokens of the encoder's last layer, known to be the most subjected to changes after a training phase [362]. The first metric, the *Averaged Infinity-norm* (AIn), quantifies the intrinsic variations in the representations $\boldsymbol{B_0}$ and $\boldsymbol{B_1}$. AIn is a new metric we introduce to capture minimal differences between the representations encoded by two models. Given two models $\boldsymbol{B_i}$ and $\boldsymbol{B_j}$, for each n-th of the $N = 5,000$ samples in the dev, we computed the vector difference $\boldsymbol{\delta^n}$ between the representations $\boldsymbol{x_i^n}$ and $\boldsymbol{x_j^n}$ of dimension d obtained by the two models B_i and B_j

$$\begin{cases} \boldsymbol{\delta^n} = \boldsymbol{x_i^n} - \boldsymbol{x_j^n} \\ \implies \forall k \in \{0...d\}, \delta_k^n = x_{i,k}^n - x_{j,k}^n \end{cases} \tag{8.1}$$

then, we computed the infinity norm (In) of the obtained vector

$$In = ||\boldsymbol{\delta^n}||_\infty = max_{k \in \{0...d\}} |\delta_k^n| \tag{8.2}$$

and finally, we averaged the norms across all the samples

$$AIn = \frac{1}{N} \sum_{n=1}^{N} ||\boldsymbol{\delta^n}||_\infty \tag{8.3}$$

The second metric is the *Representational Similarity Analysis* (RSA), already used in previous work [362, 352]. We calculate RSA on the same $N = 5,000$ *[CLS]* dev samples of the AIn metric. For each model, we computed the $N \times N$ pairwise cosine similarity matrix. Then, we flattened the upper triangular part of the matrices of our further pre-trained models and computed the correlation between the flattened upper triangular part of the BERT-related matrix (the one sharing the same FT setup and data). In line with previous work, we used Pearson correlation.

As extrinsic evaluation we assess and compare the performance of all the fine-tuned models $\boldsymbol{B_0^*}$ and $\boldsymbol{B_1^*}$ on the downstream task. We used *accuracy* since the test set is label-balanced. To assess the diversity of the fine-tuned models, we use a McNemar's test with a significance indicated for $\alpha < 0.05$. By taking into account the positive (1)

and negative (0) predictions of two models $\boldsymbol{B_i^*}$ and $\boldsymbol{B_j^*}$, we first built the contingency matrix as reported in Tab. 8.2.

		$\boldsymbol{B_j^*}$	
		positive (+)	negative (-)
$\boldsymbol{B_i^*}$	positive (+)	a = count(+,+)	b = count(+,-)
	negative (-)	c = count(-,+)	d = count(-,-)

Table 8.2.: Contingency matrix example.

Then, we tested the null hypothesis of marginal homogeneity, i.e., the marginal probabilities of the two models are the same:

$$\begin{cases} \boldsymbol{H_0}: p_b = p_c \\ \boldsymbol{H_1}: p_b \neq p_c \end{cases} \tag{8.4}$$

In cases where the test revealed to be significant ($\alpha < 0.05$), there would be sufficient evidence to reject the null hypothesis $\boldsymbol{H_0}$, i.e., the marginal proportions of the two models under exam are significantly different from each other, which is represented by the alternative hypothesis $\boldsymbol{H_1}$.

8.4. Results

The AIn metric signals differences between any of the FPT-ed models $\boldsymbol{B_1}$ and the original BERT $\boldsymbol{B_0}$, at most at the level of the sixth decimal digit ($\approx (1.87 \pm 0.90) * 10^{-6}$). Even if tiny, this result indicates that some changes took place during the one-sentence-one-step FPT process. On the contrary, these tiny differences are not detected by RSA. In any correlation between $\boldsymbol{B_1}$ and the original BERT $\boldsymbol{B_0}$, the resulting ρ was always 1.0, i.e., suggesting that the models are the same.

Results for the fine-tuned models on the downstream task are in Table 8.3. Since we have 25 different fine-tuned versions for each of the FPT-ed models $\boldsymbol{B_1^*}$ and for $\boldsymbol{B_0^*}$ (see Fine-tuning settings above), we report the maximum, the minimum, and average accuracy scores as observed across all the models. The most surprising result is that in all settings the average performances of the $\boldsymbol{B_1^*}$ models are always significantly better than their original $\boldsymbol{B_0^*}$ counterparts. This is remarkable since the averages are across 25 models, and the difference between the pre-trained models before FT is down to FPT

with a single additional sentence for a single step (with a low learning rate). A less surprising result is that larger fine-tuning training sets yield substantially better performances. This is evident both by the overall average accuracy and the *min* accuracies. This result is in line with the findings in [352], who show that increasing the sizes of the FT training data tends to neutralise the effects of the FPT phase.

FT size	Model	max	min	avg
$n = 100$	$\boldsymbol{B_0^*}$	84.77	69.56	78.18
	$\boldsymbol{B_1^*}$	84.92	66.10	79.76
$n = 200$	$\boldsymbol{B_0^*}$	86.25	78.17	83.76
	$\boldsymbol{B_1^*}$	86.63	77.79	84.55
$n = 500$	$\boldsymbol{B_0^*}$	87.05	83.09	85.82
	$\boldsymbol{B_1^*}$	87.62	85.14	86.48

Table 8.3.: Accuracy (%) with 100/200/500 training samples. Each row shows **max**imum, **min**imum, and average (**avg**) scores across all the 25 variants of the fine-tuned models, either based on the original BERT ($\boldsymbol{B_0^*}$) or on the FPT-ed models ($\boldsymbol{B_1^*}$).

FT size	max	min	avg
$n = 100$	10.26	-9.31	1.57
$n = 200$	3.26	-2.9	0.79
$n = 500$	2.56	-1.52	0.66

Table 8.4.: Absolute difference in percentage points of the fine-tuned models. The differences are computed as $\boldsymbol{B_1^*} - \boldsymbol{B_0^*}$ with the $\boldsymbol{B_1^*}$ and $\boldsymbol{B_0^*}$ models sharing the same FT settings. We report **max**imum, **min**imum, and average (**avg**) distance.

Table 8.4 zooms in on the differences between the FPT-ed and the original BERT models. Specifically, we report the maximum and the minimum difference observed between $\boldsymbol{B_1^*}$ and $\boldsymbol{B_0^*}$ within the very same FT setting. This difference is very large for smaller fine-tuning sizes, while it reduces with the increase of the FT data size. This suggests that FT size has an impact in the stability of the models, as the differences constantly shrink with higher FT sample sizes.

Lastly, under the exact same fine-tuning settings, the differences between the $\boldsymbol{B_1^*}$ models (each FPT-ed with a different single sentence) are minimal when compared to the differences that we observe comparing each $\boldsymbol{B_1^*}$ to their corresponding $\boldsymbol{B_0^*}$ model, suggesting that the impact of the nature of the sentence used for FPT is negligible, even

if such sentence is nonsensical. In other words, fine-tuning a version of BERT which has been further-pretrained with a single nonsensical sentence for a single step yields consistently better results than fine-tuning (with exactly the same settings) original BERT in the context of a sentiment analysis task. Table 8.5 reports the full results for FPT with different sentence types.

$lr_{mlm} = 1e^{-10}$	Accuracy scores (%)											
	$n = 100$				$n = 200$				$n = 500$			
	max	min	$\Delta_{max,min}$	avg	max	min	$\Delta_{max,min}$	avg	max	min	$\Delta_{max,min}$	avg
B_0^*	84.77	69.56	15.21	78.18 ± 3.71	86.25	78.17	8.08	83.76 ± 1.61	87.05	83.09	3.96	85.82 ± 0.85
$B_{1,-}^* - B_0^*$	10.26	-9.31	19.57	1.51 ± 5.27	3.45	-1.91	5.36	1.01 ± 1.45	2.67	-1.62	4.29	0.71 ± 0.95
$B_{1,/}^* - B_0^*$	10.25	-9.31	19.56	1.54 ± 5.35	3.29	-1.91	5.20	0.93 ± 1.31	4.09	-1.69	5.78	0.70 ± 1.24
$B_{1,+}^* - B_0^*$	10.26	-9.32	19.58	1.60 ± 5.27	3.15	-5.87	9.02	0.57 ± 1.88	0.83	-1.05	1.88	0.67 ± 0.61
$B_{1,?}^* - B_0^*$	10.25	-9.31	19.56	1.61 ± 5.25	3.13	-1.91	5.04	0.66 ± 1.32	2.65	-1.73	4.38	0.54 ± 1.02
$B_{1,-}^*$	84.92	66.10	18.82	79.7 ± 3.90	86.58	81.48	5.10	84.77 ± 1.15	87.62	85.4	2.22	86.53 ± 0.63
$B_{1,/}^* - B_{1,-}^*$	5.91	-5.77	11.68	0.03 ± 1.77	1.19	-1.70	2.89	-0.08 ± 0.61	1.78	-1.56	3.34	-0.01 ± 0.80
$B_{1,+}^* - B_{1,-}^*$	3.40	-0.92	4.32	0.09 ± 0.71	1.06	-9.18	10.24	-0.44 ± 1.96	0.83	-1.05	1.88	-0.04 ± 0.61
$B_{1,?}^* - B_{1,-}^*$	3.45	-2.10	5.55	0.10 ± 0.97	0.58	-4.83	5.41	-0.35 ± 1.10	0.70	-1.25	1.95	-0.17 ± 0.56
$B_{1,/}^*$	84.92	66.10	18.82	79.73 ± 3.91	86.66	80.72	5.94	84.69 ± 1.31	87.71	84.96	2.75	86.52 ± 0.7
$B_{1,+}^* - B_{1,/}^*$	4.85	-5.90	10.75	0.06 ± 1.60	0.61	-8.42	9.03	-0.36 ± 1.71	2.10	-1.36	3.46	-0.03 ± 0.72
$B_{1,?}^* - B_{1,/}^*$	4.50	-2.46	6.96	0.07 ± 1.05	0.95	-4.07	5.02	-0.27 ± 0.87	1.88	-1.44	3.32	-0.16 ± 0.67
$B_{1,+}^*$	84.92	66.09	18.83	79.79 ± 3.90	86.58	72.30	14.28	84.32 ± 2.72	87.52	85.13	2.39	86.49 ± 0.61
$B_{1,?}^* - B_{1,+}^*$	3.44	-2.07	5.51	0.01 ± 0.86	4.35	-1.08	5.43	0.09 ± 1.00	0.45	-1.20	1.65	-0.14 ± 0.40
$B_{1,?}^*$	84.92	66.10	18.82	79.80 ± 3.82	87.0	76.65	10.35	84.41 ± 1.92	87.62	85.05	2.57	86.36 ± 0.66

Table 8.5.: Accuracy scores (%) of the fine-tuned models with 100, 200, and 500 samples. Results are reported in terms of the maximum and minimum, the difference between the maximum and minimum $\Delta_{max,min}$, and the average across all the 25 variants of the models. For each block, the first row reports the performance of the single model (B_i^*), either the original Bert or the ones we further pre-trained with a learning rate equal to $1e^{-10}$. Subsequent rows in each block report the differences of performance between models $B_j^* - B_i^*$. The differences are computed between models having the same fine-tuning setup (same classifier initialization, tuning data, and data order).

For each fine-tuning data size, all the $\boldsymbol{B_1^*}$ models always resulted in being significantly different from their $\boldsymbol{B_0^*}$ counterparts (McNemar's test). Interestingly, while most $\boldsymbol{B_1^*}$ models are not significantly different from one another with small training sizes, they instead are with the largest training size. This suggests that, while with larger FT sets all $\boldsymbol{B_1^*}$ performances get closer to the ones obtained when fine-tuning the original BERT, FPT-ed models get more dissimilar from one another. Hence, as discussed above, while our results confirm Zhu et al.'s observation [352] that the larger the FT train set the smaller the measurable impact of FPT, they also show that with a larger FT training set, the $\boldsymbol{B_1^*}$ end up being more different from one another than when using little training data. Such results are confirmed by the training trends of the different models, as reported in Figure 8.2.

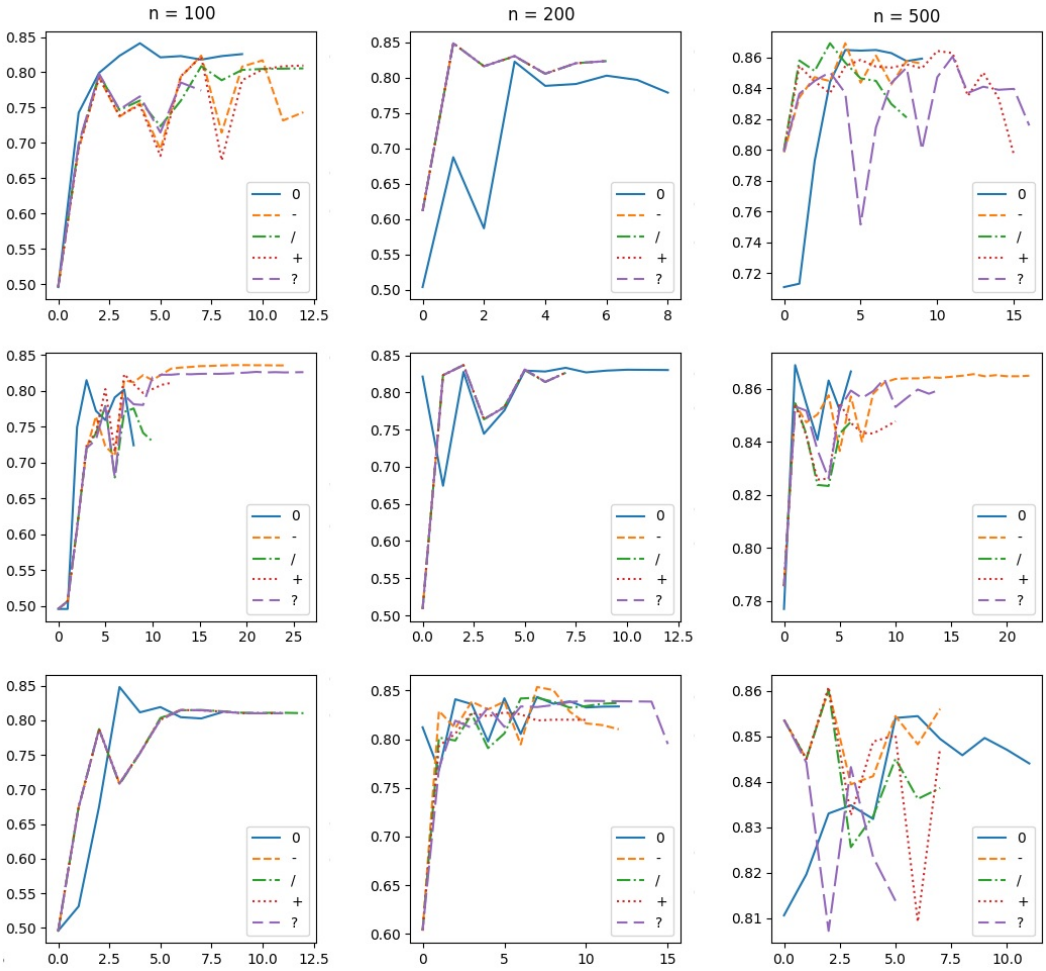

Figure 8.2.: Training trends during fine-tuning of the original BERT and the further pre-trained models to varying the number of samples used in training (along the columns) and the initialization of the classifier (along the rows).

8.5. Discussion

We observed unexpected downstream performances in a BERT model that went through a single-step and single-sentence further pre-training phase, with a low learning rate. Findings that we deem not straightforward to explain are: (i) while the models' internal representations barely differ, fine-tuning the models under exactly the same settings leads to consistently different performances; (ii) on average, all the FPT-ed models outperform their original BERT counterparts when fine-tuned on the downstream task, even when a single nonsensical sentence is used for FPT; (iii) the differences between the $\boldsymbol{B}_1^*$ and $\boldsymbol{B}_0^*$ after FT are always significant; and (iv) although increasing FT data size shrinks the gap in performance between $\boldsymbol{B}_1^*$ and the corresponding $\boldsymbol{B}_0^*$ models, it

also leads to $\boldsymbol{B_1^*}$ models being more dissimilar from one another: with a smaller trainset (100 samples) the resulting $\boldsymbol{B_1^*}$ models are never significantly different, while they are always significantly different with the largest training set (500 samples). These results seem to leave open some questions on model stability and on reliability of claims over the impact of FPT, especially if not many different settings are tested.

We hope these issues will be picked up and further unpacked. Although we analysed such instability over several parameters, a few more parameters may be taken into consideration. For example, the extent of the impacts of using higher learning rates in further pre-training could be addressed. Furthermore, other large(r) language models may turn out to be more robust than (base) BERT: an analysis of these models would improve the relevance of our findings. Furthermore, while we exploited sentiment analysis as the downstream task for providing us with easy control on the characteristics of the FPT inputs, analysing other tasks may strengthen or weaken our findings. In particular, moving from a document-level task to a token-level one (e.g., Named Entity Recognition or Part-of-Speech Tagging) may present entirely different results. This could indicate that PLM have different behaviors when FPT-ed and subsequently FT-ed on different tasks, shedding new lights on how these models work.

Part IV.

Epilogue

9. Conclusions

In this thesis, we explore the role of Natural Language Processing (NLP) in healthcare. In the opening section, we examine the historical developments and recent advancements in the field of NLP, with a particular emphasis on its potential applications in healthcare. To better understand the vast landscape of NLP in healthcare, we narrow our focus to the care of low back pain and related spine diseases, where NLP has shown great promise.

As part of our investigation into *NLP in healthcare*, we identified and undertook two important tasks. To ensure balance, we tackled both NLU and NLG tasks. For the former, in particular, we aimed to developed systems to help healthcare companies identify flaws in their care services. To achieve this, we utilized a common NLU task: sentiment analysis. By gathering data from Internet sources, we obtained the first Italian dataset of healthcare reviews, which has already been used by other researchers. For NLG, we aimed to create a system that simplifies medical texts for patients. Recent studies have highlighted the challenges patients face in comprehending medical texts, which can lead to reduced compliance with therapies and misunderstandings caused by the abundance of uncontrolled information on the web. Our analysis, conducted with both human experts (physicians) and laymen, demonstrated the effectiveness of our system. Our system outperformed previous state-of-the-art systems and achieved results comparable to the gold standard targets.

Furthermore, in the final section of this manuscript, we address one of the major barriers to NLP applications in the sensitive domain of healthcare: explainability. To overcome this challenge, we pursued a two-fold approach. Firstly, we aimed to provide users with explanations of model decisions, thereby increasing their trust in the system and enabling them to identify potential errors. To achieve this, we proposed two models for document classification. Our comparison including human annotators with a well-known benchmark demonstrated that our models performed similarly to previous systems, while extracting high-quality summaries. Secondly, we conducted experiments to analyze the mechanisms underlying the strategies for adapting large language models

to new, specific domains such as healthcare.

Our work in the healthcare field has not only advanced our understanding of NLP's potential applications but has also provided valuable insights into the field itself. For instance,

(i) for ***capturing the patients' perspective***, we discovered that modern Transformer-based models struggle with data imbalance, and thus cannot be assumed to outperform classical NLP methodologies outright.

(ii) In our efforts for ***reducing the expertise gap*** and improve communication between physicians and patients, we implemented and analyzed several Semantic Textual Similarity strategies to address the absence of parallel data, a major issue in sequence-to-sequence tasks. Specifically, we investigated how data of varying quality, measured in terms of similarity between the two texts in each pair, positively impacted the style transfer task. To complement our findings, we conducted an extensive human evaluation phase involving both experts and laypeople, and conducted a qualitative analysis to identify the strengths and limitations of different models and datasets. Additionally, we gained valuable insights regarding the evaluation metrics commonly employed in Text Style Transfer tasks.

(iii) Both our proposed ***hierarchical models***, *ExHiT* and *SCC*, can be applied to various domains and document classification tasks, providing interpretable decisions. We designed these architectures to take advantage of the inherent hierarchy in documents, utilizing a transformer at the intra-sentence token level and one another at the inter-sentences level. To evaluate the built summaries, we proposed two new human-based metrics. This is the first attempt to build a document classification paradigm of models that generate an extractive summary for easy user interpretation. Interestingly, both models have the potential to operate on longer documents, which is a limitation for traditional Transformer architectures. It would be valuable to explore whether they can overcome this limitation by applying them to tasks involving larger documents, such as those containing several hundred tokens or more.

(iv) In the "Breaking Bert" experiments, we conducted minimal-impact further pre-training of BERT by training the model on one sentence for one step with a low learning rate. This approach yielded unexpected outcomes that are not straightforward to explain. Firstly, although the internal representations of the models barely differed, fine-tuning the models under exactly the same settings led to consistently different per-

formances. Secondly, on average, all the FPT-ed models outperformed their original BERT counterparts when fine-tuned on the downstream task, even when a single nonsensical sentence was used for FPT. Thirdly, the differences between the original BERT and the FPT-ed ones after fine-tuning were always significant. Finally, increasing the size of the fine-tuning data set led to the fine-tuned FPT-ed models being more dissimilar from one another, despite shrinking the performance gap between the original BERT and the FPT-ed models. Specifically, when using a smaller training set (100 samples), the resulting FPT-ed models were never significantly different, while they always were with the largest training set (500 samples). These results raise questions about model stability and the reliability of claims regarding the impact of FPT, particularly when only a limited number of settings are tested.

We thus hope, and we do believe, that the materials presented in this manuscript will effectively help the research communities to explore new directions of study, laying new groundwork for the use of technologies such as NLP in domains as complex as healthcare.

Bibliography

[1] Ewoud Pons, Loes MM Braun, MG Myriam Hunink, and Jan A Kors. Natural language processing in radiology: a systematic review. *Radiology*, 279(2):329–343, 2016. 1

[2] Wen-wai Yim, Meliha Yetisgen, William P Harris, and Sharon W Kwan. Natural language processing in oncology: a review. *JAMA oncology*, 2(6):797–804, 2016. 1

[3] Seyedmostafa Sheikhalishahi, Riccardo Miotto, Joel T Dudley, Alberto Lavelli, Fabio Rinaldi, and Venet Osmani. Natural language processing of clinical notes on chronic diseases: systematic review. *JMIR medical informatics*, 7(2):e12239, 2019. 1

[4] Meghan Reading Turchioe, Alexander Volodarskiy, Jyotishman Pathak, Drew N Wright, James Enlou Tcheng, and David Slotwiner. Systematic review of current natural language processing methods and applications in cardiology. *Heart*, 108(12):909–916, 2022. 1

[5] James Mullenbach, Sarah Wiegreffe, Jon Duke, Jimeng Sun, and Jacob Eisenstein. Explainable prediction of medical codes from clinical text. *arXiv preprint arXiv:1802.05695*, 2018. 1

[6] Opim Salim Sitompul, Erna Budhiarti Nababan, Dedy Arisandi, Indra Aulia, and Hengky Wijaya. Template-based natural language generation in interpreting laboratory blood test. *IAENG Int. J. Comput. Sci*, 48(1), 2021. 1

[7] Laith Abualigah, Hamza Essam Alfar, Mohammad Shehab, and Alhareth Mohammed Abu Hussein. Sentiment analysis in healthcare: a brief review. *Recent Advances in NLP: The Case of Arabic Language*, pages 129–141, 2020. 1

[8] Wendy W Chapman, Prakash M Nadkarni, Lynette Hirschman, Leonard W D'Avolio, Guergana K Savova, and Ozlem Uzuner. Overcoming barriers to NLP for clinical text: the role of shared tasks and the need for additional creative solutions. *Journal of the American Medical Informatics Association*, 18(5):540–543, 09 2011. 2, 17

[9] Abeed Sarker, Mohammed Ali Al-Garadi, Yuan-Chi Yang, Jinho Choi, Arshed A Quyyumi, and Greg S Martin. Defining patient-oriented natural language processing: A

new paradigm for research and development to facilitate adoption and use by medical experts. *JMIR Med Inform*, 9(9):e18471, Sep 2021. 2, 17

[10] Andrew D Selbst and Julia Powles. Meaningful information and the right to explanation. *International Data Privacy Law*, 7(4):233–242, 12 2017. 2, 14

[11] Shoujin Wang, Wei Liu, Jia Wu, Longbing Cao, Qinxue Meng, and Paul J Kennedy. Training deep neural networks on imbalanced data sets. In *2016 international joint conference on neural networks (IJCNN)*, pages 4368–4374. IEEE, 2016. 2

[12] Karen Sparck Jones. *Natural Language Processing: A Historical Review*, pages 3–16. Springer Netherlands, Dordrecht, 1994. 7

[13] Joseph Weizenbaum. Eliza - a computer program for the study of natural language communication between man and machine. *Commun. ACM*, 26(1):23–28, jan 1983. 7

[14] Yorick Wilks. The history of natural language processing and machine translation. *Encyclopedia of Language and Linguistics*, page 9, 2005. 7

[15] Jacob Eisenstein. Natural language processing. *Jacob Eisenstein*, 2018. 7, 8

[16] Ronan Collobert, Jason Weston, Léon Bottou, Michael Karlen, Koray Kavukcuoglu, and Pavel Kuksa. Natural language processing (almost) from scratch. *Journal of machine learning research*, 12(ARTICLE):2493–2537, 2011. 8

[17] Hans Peter Luhn. A statistical approach to mechanized encoding and searching of literary information. *IBM Journal of research and development*, 1(4):309–317, 1957. 10

[18] S. E. Robertson and K. Sparck Jones. Relevance weighting of search terms. *Journal of the American Society for Information Science*, 27(3):129–146, 1976. 10

[19] K.Sparck Jones. Experiments in relevance weighting of search terms. *Information Processing & Management*, 15(3):133–144, 1979. 10

[20] Milan Straka and Jana Straková. Tokenizing, POS tagging, lemmatizing and parsing UD 2.0 with UDPipe. In *Proceedings of the CoNLL 2017 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, pages 88–99, Vancouver, Canada, August 2017. Association for Computational Linguistics. 10

[21] Peng Qi, Yuhao Zhang, Yuhui Zhang, Jason Bolton, and Christopher D. Manning. Stanza: A python natural language processing toolkit for many human languages. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 101–108, Online, July 2020. Association for Computational Linguistics. 10

[22] Daniel Gildea. Corpus variation and parser performance. In *Proceedings of the 2001 Conference on Empirical Methods in Natural Language Processing*, 2001. 10

[23] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. 10, 13, 34, 43, 59, 110, 112, 115

[24] Zellig S Harris. Methods in structural linguistics. 1951. 10

[25] Zellig S Harris. Distributional structure. *Word*, 10(2-3):146–162, 1954. 11

[26] Alessandro Lenci. Distributional semantics in linguistic and cognitive research. *Italian journal of linguistics*, 20(1):1–31, 2008. 11

[27] Katrin Erk. Vector space models of word meaning and phrase meaning: A survey. *Language and Linguistics Compass*, 6(10):635–653, 2012. 11

[28] Stephen Clark. Vector space models of lexical meaning. *Handbook of Contemporary Semantics–second edition. Wiley-Blackwell*, 2012. 11

[29] Alessandro Lenci. Distributional models of word meaning. *Annual review of Linguistics*, 4:151–171, 2018. 11

[30] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. Distributed representations of words and phrases and their compositionality. *Advances in neural information processing systems*, 26, 2013. 11

[31] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*, 2013. 11

[32] Tomáš Mikolov, Wen-tau Yih, and Geoffrey Zweig. Linguistic regularities in continuous space word representations. In *Proceedings of the 2013 conference of the north american chapter of the association for computational linguistics: Human language technologies*, pages 746–751, 2013. 11

[33] Marco Baroni, Georgiana Dinu, and Germán Kruszewski. Don't count, predict! a systematic comparison of context-counting vs. context-predicting semantic vectors. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 238–247, 2014. 11

[34] Jeffrey Pennington, Richard Socher, and Christopher D Manning. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543, 2014. 11

[35] Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. Enriching word vectors with subword information, 2016. 11

[36] Julia Hirschberg and Christopher D. Manning. Advances in natural language processing. *Science*, 349(6245):261–266, 2015. 12

[37] Jeffrey L Elman. Finding structure in time. *Cognitive science*, 14, 1990. 12

[38] Mike Schuster and Kuldip K Paliwal. Bidirectional recurrent neural networks. *IEEE transactions on Signal Processing*, 1997. 13

[39] Bryan McCann, James Bradbury, Caiming Xiong, and Richard Socher. Learned in translation: Contextualized word vectors. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. 13

[40] Matthew E Peters et al. Deep contextualized word representations. *arXiv preprint arXiv:1802.05365*, 2018. 13

[41] Jeremy Howard and Sebastian Ruder. Universal language model fine-tuning for text classification. *arXiv preprint arXiv:1801.06146*, 2018. 13

[42] Martin Sundermeyer et al. Lstm neural networks for language modeling. In *Thirteenth annual conference of the international speech communication association*, 2012. 13

[43] Martin Sundermeyer et al. From feedforward to recurrent lstm neural networks for language modeling. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 2015. 13

[44] Yoshua Bengio et al. Learning long-term dependencies with gradient descent is difficult. *IEEE transactions on neural networks*, 1994. 13

[45] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 1997. 13

[46] Kyunghyun Cho et al. Learning phrase representations using rnn encoder-decoder for statistical machine translation. *arXiv preprint arXiv:1406.1078*, 2014. 13

[47] Ashish Vaswani et al. Attention is all you need. In *Advances in neural information processing systems*, 2017. 13, 88

[48] Hugo Larochelle and Geoffrey E Hinton. Learning to combine foveal glimpses with a third-order boltzmann machine. In *Advances in neural information processing systems*, 2010. 13

[49] Dzmitry Bahdanau et al. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*, 2014. 13

[50] Yinhan Liu et al. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*, 2019. 14, 99

[51] Victor Sanh et al. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*, 2019. 14

[52] Alec Radford et al. Improving language understanding by generative pre-training, 2018. 14

[53] Alec Radford et al. Language models are unsupervised multitask learners. 2019. 14

[54] Tom B Brown et al. Language models are few-shot learners. *arXiv preprint arXiv:2005.14165*, 2020. 14

[55] Alex Wang et al. Glue: A multi-task benchmark and analysis platform for natural language understanding. *arXiv preprint arXiv:1804.07461*, 2018. 14

[56] Sinno Jialin Pan and Qiang Yang. A survey on transfer learning. *IEEE Transactions on knowledge and data engineering*, 2009. 14

[57] Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4):1234–1240, 09 2019. 14, 111, 114

[58] Emily Alsentzer, John Murphy, William Boag, Wei-Hung Weng, Di Jindi, Tristan Naumann, and Matthew McDermott. Publicly available clinical BERT embeddings. In *Proceedings of the 2nd Clinical Natural Language Processing Workshop*, pages 72–78, Minneapolis, Minnesota, USA, June 2019. Association for Computational Linguistics. 14, 18, 59, 111, 114

[59] Usman Naseem, Adam G Dunn, Matloob Khushi, and Jinman Kim. Benchmarking for biomedical natural language processing tasks with a domain specific albert. *BMC bioinformatics*, 23(1):1–15, 2022. 14

[60] Yoojoong Kim, Jong-Ho Kim, Jeong Moon Lee, Moon Joung Jang, Yun Jin Yum, Seongtae Kim, Unsub Shin, Young-Min Kim, Hyung Joon Joo, and Sanghoun Song. A pre-

trained bert for korean medical natural language processing. *Scientific Reports*, 12(1):1–10, 2022. 14

[61] Yongjie Lin, Yi Chern Tan, and Robert Frank. Open sesame: Getting inside BERT's linguistic knowledge. In *Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 241–253, Florence, Italy, August 2019. Association for Computational Linguistics. 14, 110

[62] Christopher D. Manning, Kevin Clark, John Hewitt, Urvashi Khandelwal, and Omer Levy. Emergent linguistic structure in artificial neural networks trained by self-supervision. *Proceedings of the National Academy of Sciences*, 117(48):30046–30054, 2020. 14, 110

[63] Alessio Miaschi, Dominique Brunato, Felice Dell'Orletta, and Giulia Venturi. Linguistic profiling of a neural language model. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 745–756, Barcelona, Spain (Online), December 2020. International Committee on Computational Linguistics. 14, 110

[64] Giovanni Puccetti, Alessio Miaschi, and Felice Dell'Orletta. How do BERT embeddings organize linguistic knowledge? In *Proceedings of Deep Learning Inside Out (DeeLIO): The 2nd Workshop on Knowledge Extraction and Integration for Deep Learning Architectures*, pages 48–57, Online, June 2021. Association for Computational Linguistics. 14, 110

[65] Ian Tenney, Dipanjan Das, and Ellie Pavlick. BERT rediscovers the classical NLP pipeline. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4593–4601, Florence, Italy, July 2019. Association for Computational Linguistics. 14, 110

[66] Vincent Micheli, Martin d'Hoffschmidt, and François Fleuret. On the importance of pre-training data volume for compact language models. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7853–7858, Online, November 2020. Association for Computational Linguistics. 14, 110

[67] Kundan Krishna, Saurabh Garg, Jeffrey P Bigham, and Zachary C Lipton. Downstream datasets make surprisingly good pretraining corpora. *arXiv preprint arXiv:2209.14389*, 2022. 14, 110

[68] Koustuv Sinha, Robin Jia, Dieuwke Hupkes, Joelle Pineau, Adina Williams, and Douwe Kiela. Masked language modeling and the distributional hypothesis: Order word matters pre-training for little. In *Proceedings of the 2021 Conference on Empirical Methods in*

Natural Language Processing, pages 2888–2913, Online and Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics. 14, 110

[69] Kundan Krishna, Jeffrey Bigham, and Zachary C. Lipton. Does pretraining for summarization require knowledge transfer? In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 3178–3189, Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics. 14, 110

[70] David Cheng-Han Chiang and Hung yi Lee. Pre-training a language model without human language. *CoRR*, abs/2012.11995, 2020. 14, 110

[71] Yonatan Belinkov and James Glass. Analysis Methods in Neural Language Processing: A Survey. *Transactions of the Association for Computational Linguistics*, 7:49–72, 04 2019. 14

[72] Octavio Loyola-Gonzalez. Black-box vs. white-box: Understanding their advantages and weaknesses from a practical point of view. *IEEE Access*, 2019. 14, 87

[73] David Gunning. Explainable artificial intelligence (xai). *Defense Advanced Research Projects Agency (DARPA), nd Web*, 2017. 14, 87

[74] David Gunning and David Aha. Darpa's explainable artificial intelligence (xai) program. *AI Magazine*, 2019. 14, 87

[75] Alejandro Barredo Arrieta et al. Explainable artificial intelligence (xai): Concepts, taxonomies, opportunities and challenges toward responsible ai. *Information Fusion*. 14, 87

[76] Marina Danilevsky et al. A survey of the state of explainable ai for natural language processing. *arXiv preprint arXiv:2010.00711*, 2020. 14, 88

[77] Marco Tulio Ribeiro et al. " why should i trust you?" explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, 2016. 14, 94

[78] Yujia Zhang et al. " why should you trust my explanation?" understanding uncertainty in lime explanations. *arXiv preprint arXiv:1904.12991*, 2019. 14, 94

[79] Mukund Sundararajan et al. Axiomatic attribution for deep networks. In *International Conference on Machine Learning*, pages 3319–3328, 2017. 14, 94

[80] Sarthak Jain and Byron C. Wallace. Attention is not Explanation. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*,

pages 3543–3556, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. 14

[81] Sarah Wiegreffe and Yuval Pinter. Attention is not not explanation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 11–20, Hong Kong, China, November 2019. Association for Computational Linguistics. 14

[82] Jesse Vig. A multiscale visualization of attention in the transformer model. *arXiv preprint arXiv:1906.05714*, 2019. 15, 93

[83] Yanshan Wang, Liwei Wang, Majid Rastegar-Mojarad, Sungrim Moon, Feichen Shen, Naveed Afzal, Sijia Liu, Yuqun Zeng, Saeed Mehrabi, Sunghwan Sohn, and Hongfang Liu. Clinical information extraction applications: A literature review. *Journal of Biomedical Informatics*, 77:34–49, 2018. 16

[84] Kong Hyoun-Joong. Managing unstructured big data in healthcare system. *hir*, 25(1):1–2, 2019. 16

[85] Cristina Soguero-Ruiz, Kristian Hindberg, Inmaculada Mora-Jiménez, José Luis Rojo-Álvarez, Stein Olav Skrøvseth, Fred Godtliebsen, Kim Mortensen, Arthur Revhaug, Rolv-Ole Lindsetmo, Knut Magne Augestad, and Robert Jenssen. Predicting colorectal surgical complications using heterogeneous clinical data and kernel methods. *Journal of Biomedical Informatics*, 61:87–96, 2016. 16

[86] Kasper Jensen, Cristina Soguero-Ruiz, Karl Oyvind Mikalsen, Rolv-Ole Lindsetmo, Irene Kouskoumvekaki, Mark Girolami, Stein Olav Skrovseth, and Knut Magne Augestad. Analysis of free text in electronic health records for identification of cancer patient trajectories. *Scientific reports*, 7(1):1–12, 2017. 16

[87] Lise Poissant, Jennifer Pereira, Robyn Tamblyn, and Yuko Kawasumi. The Impact of Electronic Health Records on Time Efficiency of Physicians and Nurses: A Systematic Review. *Journal of the American Medical Informatics Association*, 12(5):505–516, 09 2005. 16

[88] Christine Sinsky, Lacey Colligan, Ling Li, Mirela Prgomet, Sam Reynolds, Lindsey Goeders, Johanna Westbrook, Michael Tutty, and George Blike. Allocation of physician time in ambulatory practice: a time and motion study in 4 specialties. *Annals of internal medicine*, 165(11):753–760, 2016. 16

[89] A Névéol, P Zweigenbaum, et al. Clinical natural language processing in 2014: foundational methods supporting efficient healthcare. *Yearbook of Medical Informatics*, 24(01):194–198, 2015. 16

[90] Sumithra Velupillai, Danielle Mowery, Brett R South, Maria Kvist, and Hercules Dalianis. Recent advances in clinical natural language processing in support of semantic analysis. *Yearbook of medical informatics*, 24(01):183–193, 2015. 16

[91] Sumithra Velupillai, Hanna Suominen, Maria Liakata, Angus Roberts, Anoop D. Shah, Katherine Morley, David Osborn, Joseph Hayes, Robert Stewart, Johnny Downs, Wendy Chapman, and Rina Dutta. Using clinical natural language processing for health outcomes research: Overview and actionable suggestions for future advances. *Journal of Biomedical Informatics*, 88:11–19, 2018. 16

[92] Rimma Pivovarov and Noémie Elhadad. Automated methods for the summarization of electronic health records. *Journal of the American Medical Informatics Association*, 22(5):938–947, 04 2015. 16

[93] Yanjun Gao, Dmitriy Dligach, Timothy Miller, Dongfang Xu, Matthew M. M. Churpek, and Majid Afshar. Summarizing patients' problems from hospital progress notes using pre-trained sequence-to-sequence models. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 2979–2991, Gyeongju, Republic of Korea, October 2022. International Committee on Computational Linguistics. 16

[94] Rushan Long, Dan Yang, and Yang Liu. Diseasenet: A novel disease diagnosis deep framework via fusing medical record summarization. *IAENG International Journal of Computer Science*, 49(3), 2022. 16

[95] Robert B Fetter, Youngsoo Shin, Jean L Freeman, Richard F Averill, and John D Thompson. Case mix definition by diagnosis-related groups. *Medical care*, 18(2):i–53, 1980. 16

[96] Rajvir Kaur, Jeewani Anupama Ginige, and Oliver Obst. A systematic literature review of automated icd coding and classification systems using discharge summaries, 2021. 17

[97] Tal Baumel, Jumana Nassour-Kassis, Raphael Cohen, Michael Elhadad, and Noémie Elhadad. Multi-label classification of patient notes a case study on icd code assignment. *arXiv preprint arXiv:1709.09587*, 2017. 17, 18

[98] James Mullenbach, Sarah Wiegreffe, Jon Duke, Jimeng Sun, and Jacob Eisenstein. Explainable prediction of medical codes from clinical text. *arXiv preprint arXiv:1802.05695*, 2018. 17, 18

[99] Damian Pascual, Sandro Luck, and Roger Wattenhofer. Towards BERT-based automatic ICD coding: Limitations and opportunities. In *Proceedings of the 20th Workshop on Biomedical Language Processing*, pages 54–63, Online, June 2021. Association for Computational Linguistics. 17

[100] Ning Zhang and Maciej Jankowski. Hierarchical bert for medical document understanding. *arXiv preprint arXiv:2204.09600*, 2022. 17

[101] Accountability Act. Health insurance portability and accountability act of 1996. *Public law*, 104:191, 1996. 17

[102] Rachel Nosowsky and Thomas J. Giordano. The health insurance portability and accountability act of 1996 (hipaa) privacy rule: Implications for clinical research. *Annual Review of Medicine*, 57(1):575–590, 2006. PMID: 16409167. 17

[103] Özlem Uzuner, Yuan Luo, and Peter Szolovits. Evaluating the State-of-the-Art in Automatic De-identification. *Journal of the American Medical Informatics Association*, 14(5):550–563, 09 2007. 17

[104] Yanjun Gao, Dmitriy Dligach, Leslie Christensen, Samuel Tesch, Ryan Laffin, Dongfang Xu, Timothy Miller, Ozlem Uzuner, Matthew M Churpek, and Majid Afshar. A scoping review of publicly available language tasks in clinical natural language processing. *Journal of the American Medical Informatics Association*, 29(10):1797–1806, 08 2022. 17, 21

[105] Hanna Suominen, Sanna Salanterä, Sumithra Velupillai, Wendy W. Chapman, Guergana Savova, Noemie Elhadad, Sameer Pradhan, Brett R. South, Danielle L. Mowery, Gareth J. F. Jones, Johannes Leveling, Liadh Kelly, Lorraine Goeuriot, David Martinez, and Guido Zuccon. Overview of the share/clef ehealth evaluation lab 2013. In Pamela Forner, Henning Müller, Roberto Paredes, Paolo Rosso, and Benno Stein, editors, *Information Access Evaluation. Multilinguality, Multimodality, and Visualization*, pages 212–231, Berlin, Heidelberg, 2013. Springer Berlin Heidelberg. 17

[106] Joon Lee, Daniel J. Scott, Mauricio Villarroel, Gari D. Clifford, Mohammed Saeed, and Roger G. Mark. Open-access mimic-ii database for intensive care research. In *2011 Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, pages 8315–8318, 2011. 17

[107] Alistair EW Johnson, Tom J Pollard, Lu Shen, Li-wei H Lehman, Mengling Feng, Mohammad Ghassemi, Benjamin Moody, Peter Szolovits, Leo Anthony Celi, and Roger G Mark. Mimic-iii, a freely accessible critical care database. *Scientific data*, 3(1):1–9, 2016. 17

[108] Siddhartha Nuthakki, Sunil Neela, Judy W Gichoya, and Saptarshi Purkayastha. Natural language processing of mimic-iii clinical notes for identifying diagnosis and procedures with neural networks. *arXiv preprint arXiv:1912.12397*, 2019. 18

[109] AK Singh, Mounika Guntu, Ananth Reddy Bhimireddy, Judy W Gichoya, and Saptarshi Purkayastha. Multi-label natural language processing to identify diagnosis and procedure codes from mimic-iii inpatient notes. *arXiv preprint arXiv:2003.07507*, 2020. 18

[110] Andrey Kormilitzin, Nemanja Vaci, Qiang Liu, and Alejo Nevado-Holgado. Med7: A transferable clinical natural language processing model for electronic health records. *Artificial Intelligence in Medicine*, 118:102086, 2021. 18

[111] Zeljko Kraljevic, Thomas Searle, Anthony Shek, Lukasz Roguski, Kawsar Noor, Daniel Bean, Aurelie Mascio, Leilei Zhu, Amos A Folarin, Angus Roberts, et al. Multi-domain clinical natural language processing with medcat: the medical concept annotation toolkit. *Artificial Intelligence in Medicine*, 117:102083, 2021. 18

[112] Carol Friedman and Stephen B Johnson. Natural language and text processing in biomedicine. In *Biomedical Informatics*, pages 312–343. Springer, 2006. 18

[113] Alexa T McCray, Suresh Srinivasan, and Allen C Browne. Lexical methods for managing variation in biomedical terminologies. In *Proceedings of the Annual Symposium on Computer Application in Medical Care*, page 235. American Medical Informatics Association, 1994. 18

[114] Donald AB Lindberg, Betsy L Humphreys, and Alexa T McCray. The unified medical language system. *Yearbook of medical informatics*, 2(01):41–51, 1993. 18

[115] Divya Gopinath, Monica Agrawal, Luke Murray, Steven Horng, David Karger, and David Sontag. Fast, structured clinical documentation via contextual autocomplete. In *Machine Learning for Healthcare Conference*, pages 842–870. PMLR, 2020. 18

[116] Goonmeet Bajaj, Vinh Nguyen, Thilini Wijesiriwardene, Hong Yung Yip, Vishesh Javangula, Amit Sheth, Srinivasan Parthasarathy, and Olivier Bodenreider. Evaluating biomedical word embeddings for vocabulary alignment at scale in the UMLS Metathesaurus using Siamese networks. In *Proceedings of the Third Workshop on Insights from Negative Results in NLP*, pages 82–87, Dublin, Ireland, May 2022. Association for Computational Linguistics. 18

[117] Timothy Baldwin, Paul Cook, Marco Lui, Andrew MacKinlay, and Li Wang. How noisy social media text, how diffrnt social media sources? In *Proceedings of the Sixth International Joint Conference on Natural Language Processing*, pages 356–364, 2013. 18

[118] Alan R Aronson and François-Michel Lang. An overview of MetaMap: historical perspective and recent advances. *Journal of the American Medical Informatics Association*, 17(3):229–236, 05 2010. 18

[119] Guergana K Savova, James J Masanz, Philip V Ogren, Jiaping Zheng, Sunghwan Sohn, Karin C Kipper-Schuler, and Christopher G Chute. Mayo clinical Text Analysis and Knowledge Extraction System (cTAKES): architecture, component evaluation and applications. *Journal of the American Medical Informatics Association*, 17(5):507–513, 09 2010. 18

[120] Kerstin Denecke. Information extraction from medical social media. In *Health Web Science*, pages 61–73. Springer, 2015. 18

[121] Albert Park and Mike Conway. Tracking health related discussions on reddit for public health applications. In *AMIA annual symposium proceedings*, volume 2017, page 1362. American Medical Informatics Association, 2017. 18

[122] Mark Dredze and Michael J Paul. Natural language processing for health and social media. *IEEE Intelligent Systems*, 29(2):64–67, 2014. 18

[123] David Andre Broniatowski, Mark Dredze, Michael J Paul, and Andrea Dugas. Using social media to perform local influenza surveillance in an inner-city hospital: a retrospective observational study. *JMIR public health and surveillance*, 1(1):e4472, 2015. 18

[124] J Danielle Sharpe, Richard S Hopkins, Robert L Cook, and Catherine W Striley. Evaluating google, twitter, and wikipedia as tools for influenza surveillance using bayesian change point analysis: a comparative analysis. *JMIR public health and surveillance*, 2(2):e5901, 2016. 18

[125] Abul Hasan, Mark Levene, David Weston, Renate Fromson, Nicolas Koslover, and Tamara Levene. Monitoring covid-19 on social media: Development of an end-to-end natural language processing pipeline using a novel triage and diagnosis approach. *J Med Internet Res*, 24(2):e30397, Feb 2022. 18

[126] Azadeh Nikfarjam, Abeed Sarker, Karen O'connor, Rachel Ginn, and Graciela Gonzalez. Pharmacovigilance from social media: mining adverse drug reaction mentions using sequence labeling with word embedding cluster features. *Journal of the American Medical Informatics Association*, 22(3):671–681, 2015. 18

[127] John van Stekelenborg, Johan Ellenius, Simon Maskell, Tomas Bergvall, Ola Caster, Nabarun Dasgupta, Juergen Dietrich, Sara Gama, David Lewis, Victoria Newbould, et al. Recommendations for the use of social media in pharmacovigilance: lessons from imi web-radr. *Drug safety*, 42(12):1393–1407, 2019. 18

[128] Lucie M Gattepaille, Sara Hedfors Vidlin, Tomas Bergvall, Carrie E Pierce, and Johan Ellenius. Prospective evaluation of adverse event recognition systems in twitter: Results from the web-radr project. *Drug safety*, 43(8):797–808, 2020. 18

[129] RAFAEL A. CALVO, DAVID N. MILNE, M. SAZZAD HUSSAIN, and HELEN CHRISTENSEN. Natural language processing in mental health applications using non-clinical texts. *Natural Language Engineering*, 23(5):649–685, 2017. 18

[130] Ruba Skaik and Diana Inkpen. Using social media for mental health surveillance: a review. *ACM Computing Surveys (CSUR)*, 53(6):1–31, 2020. 18

[131] Daniel M Low, Laurie Rumker, Tanya Talkar, John Torous, Guillermo Cecchi, and Satrajit S Ghosh. Natural language processing reveals vulnerable mental health support groups and heightened health anxiety on reddit during covid-19: Observational study. *J Med Internet Res*, 22(10):e22635, Oct 2020. 18

[132] Nick Boettcher. Studies of depression and anxiety using reddit as a data source: Scoping review. *JMIR Ment Health*, 8(11):e29487, Nov 2021. 18

[133] Glen Coppersmith, Ryan Leary, Patrick Crutchley, and Alex Fine. Natural language processing of social media as screening for suicide risk. *Biomedical Informatics Insights*, 10:1178222618792860, 2018. PMID: 30158822. 18

[134] Eldar Yeskuatov, Sook-Ling Chua, and Lee Kien Foo. Leveraging reddit for suicidal ideation detection: A review of machine learning and natural language processing techniques. *International Journal of Environmental Research and Public Health*, 19(16), 2022. 18

[135] Graciela Gonzalez-Hernandez, Abeed Sarker, Karen O'Connor, and Guergana Savova. Capturing the patient's perspective: a review of advances in natural language processing of health-related text. *Yearbook of medical informatics*, 26(01):214–227, 2017. 19, 21

[136] Susannah Fox et al. *The social life of health information, 2011.* California Healthcare Foundation, 2011. 19

[137] Felix Greaves, Christopher Millett, et al. Consistently increasing numbers of online ratings of healthcare in england. *J Med Internet Res*, 14(3):e94, 2012. 19, 39

[138] Guodong Gordon Gao, Jeffrey S McCullough, Ritu Agarwal, and Ashish K Jha. A changing landscape of physician quality reporting: analysis of patients' online ratings of their physicians over a 5-year period. *Journal of medical Internet research*, 14(1):e2003, 2012. 19, 39

[139] Kerstin Denecke and Yihan Deng. Sentiment analysis in medical settings: New opportunities and challenges. *Artificial Intelligence in Medicine*, 64(1):17–27, 2015. 19

[140] Colin Camerer, George Loewenstein, and Martin Weber. The curse of knowledge in economic settings: An experimental analysis. *Journal of political Economy*, 97(5):1232–1254, 1989. 19, 50

[141] Rita D. Zielstorff. Controlled vocabularies for consumer health. *Journal of Biomedical Informatics*, 36(4):326–333, 2003. Building Nursing Knowledge through Informatics: From Concept Representation to Data Mining. 19, 50, 53

[142] Monique A Sager, Aditya M Kashyap, Mila Tamminga, Sadhana Ravoori, Christopher Callison-Burch, and Jules B Lipoff. Identifying and responding to health misinformation on reddit dermatology forums with artificially intelligent bots using natural language processing: Design and evaluation study. *JMIR Dermatol*, 4(2):e20975, Sep 2021. 19

[143] Parth Patwa, Shivam Sharma, Srinivas Pykl, Vineeth Guptha, Gitanjali Kumari, Md Shad Akhtar, Asif Ekbal, Amitava Das, and Tanmoy Chakraborty. Fighting an infodemic: Covid-19 fake news dataset. In *International Workshop on Combating On line Ho st ile Posts in Regional Languages dur ing Emerge ncy Si tuation*, pages 21–29. Springer, 2021. 19

[144] Raymond Francis Sarmiento and Franck Dernoncourt. *Improving Patient Cohort Identification Using Natural Language Processing*, pages 405–417. Springer International Publishing, Cham, 2016. 20

[145] Victor M Castro, Dmitriy Dligach, Sean Finan, Sheng Yu, Anil Can, Muhammad Abd-El-Barr, Vivian Gainer, Nancy A Shadick, Shawn Murphy, Tianxi Cai, et al. Large-scale identification of patients with cerebral aneurysms using natural language processing. *Neurology*, 88(2):164–168, 2017. 20

[146] Sunghwan Sohn, Yanshan Wang, Chung-Il Wi, Elizabeth A Krusemark, Euijung Ryu, Mir H Ali, Young J Juhn, and Hongfang Liu. Clinical documentation variations and nlp system portability: a case study in asthma birth cohorts across institutions. *Journal of the American Medical Informatics Association*, 25(3):353–359, 2018. 20

[147] Xiaosong Wang, Yifan Peng, Le Lu, Zhiyong Lu, Mohammadhadi Bagheri, and Ronald M. Summers. Chestx-ray8: Hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017. 20

[148] Michael L Mauldin. Chatterbots, tinymuds, and the turing test: Entering the loebner prize competition. In *AAAI*, volume 94, pages 16–21, 1994. 20

[149] Tricia S. Tang, Martha M. Funnell, Morton B. Brown, and Jacob E. Kurlander. Self-management support in "real-world" settings: An empowerment-based intervention. *Patient Education and Counseling*, 79(2):178–184, 2010. 20

[150] Pritika Parmar, Jina Ryu, Shivani Pandya, João Sedoc, and Smisha Agarwal. Health-focused conversational agents in person-centered care: a review of apps. *npj Digital Medicine*, 5(1), 2022. Cited by: 2; All Open Access, Gold Open Access, Green Open Access. 20

[151] Essam H Houssein, Rehab E Mohamed, and Abdelmgeid A Ali. Machine learning techniques for biomedical natural language processing: a comprehensive review. *IEEE Access*, 2021. 21

[152] Tianlin Zhang, Annika M. Schoene, Shaoxiong Ji, and Sophia Ananiadou. Natural language processing applied to mental illness detection: a narrative review. *npj Digital Medicine*, 5(1), 2022. Cited by: 5; All Open Access, Gold Open Access, Green Open Access. 21

[153] Jimmy S. Chen and Sally L. Baxter. Applications of natural language processing in ophthalmology: present and future. *Frontiers in Medicine*, 9, 2022. Cited by: 0; All Open Access, Gold Open Access, Green Open Access. 21

[154] Yanjun Gao, Dmitriy Dligach, Timothy Miller, John Caskey, Brihat Sharma, Matthew M Churpek, and Majid Afshar. Dr.bench: Diagnostic reasoning benchmark for clinical natural language processing, 2022. 21

[155] Luca Bacco, Fabrizio Russo, Luca Ambrosio, Federico D'Antoni, Luca Vollero, Gianluca Vadalà, Felice Dell'Orletta, Mario Merone, Rocco Papalia, and Vincenzo Denaro. Natural language processing in low back pain and spine diseases: A systematic review. *Frontiers in Surgery*, 9, 2022. 22

[156] Aimin Wu, Lyn March, Xuanqi Zheng, Jinfeng Huang, Xiangyang Wang, Jie Zhao, Fiona M Blyth, Emma Smith, Rachelle Buchbinder, and Damian Hoy. Global low back pain prevalence and years lived with disability from 1990 to 2017: estimates from the global burden of disease study 2017. *Annals of translational medicine*, 8(6), 2020. 22

[157] Leah J Jeffries, Steve F Milanese, and Karen A Grimmer-Somers. Epidemiology of adolescent spinal pain: a systematic overview of the research literature. *Spine*, 32(23):2630–2637, 2007. 22

[158] Federico Balagué, Anne F Mannion, Ferran Pellisé, and Christine Cedraschi. Non-specific low back pain. *The lancet*, 379(9814):482–491, 2012. 22

[159] P Croft, AS Rigby, R Boswell, J Schollum, and A Silman. The prevalence of chronic widespread pain in the general population. *The Journal of rheumatology*, 20(4):710–713, 1993. 22

[160] How-Ran Guo, Shiro Tanaka, William E Halperin, and Lorraine L Cameron. Back pain prevalence in us industry and estimates of lost workdays. *American journal of public health*, 89(7):1029–1035, 1999. 22

[161] Jeffrey N Katz. Lumbar disc disorders and low-back pain: socioeconomic factors and consequences. *JBJS*, 88(suppl_2):21–24, 2006. 22

[162] Federico D'Antoni, Fabrizio Russo, Luca Ambrosio, Luca Vollero, Gianluca Vadalà, Mario Merone, Rocco Papalia, and Vincenzo Denaro. Artificial intelligence and computer vision in low back pain: A systematic review. *International journal of environmental research and public health*, 18(20):10909, 2021. 23

[163] Federico D'Antoni, Fabrizio Russo, Luca Ambrosio, Luca Bacco, Luca Vollero, Gianluca Vadalà, Mario Merone, Rocco Papalia, and Vincenzo Denaro. Artificial intelligence and computer aided diagnosis in chronic low back pain: A systematic review. *International Journal of Environmental Research and Public Health*, 19(10), 2022. 23

[164] Penny F Whiting, Anne WS Rutjes, Marie E Westwood, Susan Mallett, Jonathan J Deeks, Johannes B Reitsma, Mariska MG Leeflang, Jonathan AC Sterne, Patrick MM Bossuyt, and QUADAS-2 Group*. Quadas-2: a revised tool for the quality assessment of diagnostic accuracy studies. *Annals of internal medicine*, 155(8):529–536, 2011. 26

[165] Olivier Q Groot, Paul T Ogink, Jacobien H Oosterhoff, and Andrew L Beam. Natural language processing and its role in spine surgery: A narrative review of potentials and challenges. In *Seminars in Spine Surgery*, page 100877. Elsevier, 2021. 27

[166] Waleed Brinjikji, Patrick H Luetmer, Bryan Comstock, Brian W Bresnahan, LE Chen, RA Deyo, Safwan Halabi, JA Turner, AL Avins, K James, et al. Systematic literature review of imaging features of spinal degeneration in asymptomatic populations. *American Journal of Neuroradiology*, 36(4):811–816, 2015. 29

[167] W Katherine Tan, Saeed Hassanpour, Patrick J Heagerty, Sean D Rundell, Pradeep Suri, Hannu T Huhdanpaa, Kathryn James, David S Carrell, Curtis P Langlotz, Nancy L Organ, et al. Comparison of natural language processing rules-based and machine-learning systems to identify lumbar spine imaging findings related to low back pain. *Academic radiology*, 25(11):1422–1432, 2018. 29, 32, 33, 34

[168] Jinglan Mu, Andrea D Furlan, Wai Yee Lam, Marcos Y Hsu, Zhipeng Ning, and Lixing Lao. Acupuncture for chronic nonspecific low back pain. *Cochrane Database of Systematic Reviews*, (12), 2020. 29

[169] Riccardo Miotto, Bethany L Percha, Benjamin S Glicksberg, Hao-Chih Lee, Lisanne Cruz, Joel T Dudley, and Ismail Nabeel. Identifying acute low back pain episodes in primary care practice from clinical notes: Observational study. *JMIR medical informatics*, 8(2):e16878, 2020. 29, 33, 34

[170] Philip C Robinson, Sjef van der Linden, Muhammad A Khan, and William J Taylor. Axial spondyloarthritis: concept, construct, classification and implications for therapy. *Nature Reviews Rheumatology*, 17(2):109–118, 2021. 30

[171] Sizheng Steven Zhao, Chuan Hong, Tianrun Cai, Chang Xu, Jie Huang, Joerg Ermann, Nicola J Goodson, Daniel H Solomon, Tianxi Cai, and Katherine P Liao. Incorporating natural language processing to improve classification of axial spondyloarthritis using electronic health records. *Rheumatology*, 59(5):1059–1065, 2020. 30, 32, 33, 34

[172] Jessica A Walsh, Yijun Shao, Jianwei Leng, Tao He, Chia-Chen Teng, Doug Redd, Qing Treitler Zeng, Zachary Burningham, Daniel O Clegg, and Brian C Sauer. Identifying axial spondyloarthritis in electronic medical records of us veterans. *Arthritis care & research*, 69(9):1414–1420, 2017. 30, 32, 33

[173] Jessica A Walsh, Shaobo Pei, Gopi Penmetsa, Jared Lareno Hansen, Grant W Cannon, Daniel O Clegg, and Brian C Sauer. Identification of axial spondyloarthritis patients in a large dataset: the development and validation of novel methods. *The Journal of rheumatology*, 47(1):42–49, 2020. 30, 33

[174] Michael Travis Caton, Walter F Wiggins, Stuart R Pomerantz, and Katherine P Andriole. Effects of age and sex on the distribution and symmetry of lumbar spinal and neural foraminal stenosis: a natural language processing analysis of 43,255 lumbar mri reports. *Neuroradiology*, 63(6):959–966, 2021. 30, 32, 33

[175] Michael Travis Caton, Walter F Wiggins, Stuart R Pomerantz, and Katherine P Andriole. The composite severity score for lumbar spine mri: a metric of cumulative degenerative disease predicts time spent on interpretation and reporting. *Journal of digital imaging*, pages 1–9, 2021. 30, 32, 33

[176] Tue Secher Jensen, Jaro Karppinen, Joan S Sorensen, Jaakko Niinimäki, and Charlotte Leboeuf-Yde. Vertebral endplate signal changes (modic change): a systematic literature review of prevalence and association with non-specific low back pain. *European Spine Journal*, 17(11):1407–1422, 2008. 30

[177] Hannu T Huhdanpaa, W Katherine Tan, Sean D Rundell, Pradeep Suri, Falgun H Chokshi, Bryan A Comstock, Patrick J Heagerty, Kathryn T James, Andrew L Avins, Srdjan S Nedeljkovic, et al. Using natural language processing of free-text radiology reports to identify type 1 modic endplate changes. *Journal of digital imaging*, 31(1):84–90, 2018. 30, 32, 33

[178] Hamid Hassanzadeh, Joshua Bell, Manminder Bhatia, and Varun Puvanesarajah. Incidental durotomy in lumbar spine surgery; risk factors, complications, and perioperative management. *JAAOS-Journal of the American Academy of Orthopaedic Surgeons*, pages 10–5435, 2021. 30

[179] Hisatoshi Ishikura, Satoshi Ogihara, Hiroyuki Oka, Toru Maruyama, Hirohiko Inanami, Kota Miyoshi, Ko Matsudaira, Hirotaka Chikuda, Seiichi Azuma, Naohiro Kawamura, et al. Risk factors for incidental durotomy during posterior open spine surgery for degenerative diseases in adults: a multicenter observational study. *PLoS One*, 12(11):e0188038, 2017. 31

[180] Aditya V Karhade, Michiel ER Bongers, Olivier Q Groot, Erick R Kazarian, Thomas D Cha, Harold A Fogel, Stuart H Hershman, Daniel G Tobert, Andrew J Schoenfeld, Christopher M Bono, et al. Natural language processing for automated detection of incidental durotomy. *The Spine Journal*, 20(5):695–700, 2020. 31, 32, 33, 35

[181] Jeff Ehresman, Zach Pennington, Aditya V Karhade, Sakibul Huq, Ravi Medikonda, Andrew Schilling, James Feghali, Andrew Hersh, A Karim Ahmed, Ethan Cottrill, et al. Incidental durotomy: predictive risk model and external validation of natural language process identification algorithm. *Journal of Neurosurgery: Spine*, 33(3):342–348, 2020. 31, 32, 33

[182] Aditya V Karhade, Michiel ER Bongers, Olivier Q Groot, Thomas D Cha, Terence P Doorly, Harold A Fogel, Stuart H Hershman, Daniel G Tobert, Sunita D Srivastava, Christopher M Bono, et al. Development of machine learning and natural language processing algorithms for preoperative prediction and automated identification of intraoperative vascular injury in anterior lumbar spine surgery. *The Spine Journal*, 21(10):1635–1642, 2021. 31, 32, 33

[183] Raymund B Dantes, Shuai Zheng, James J Lu, Michele G Beckman, Asha Krishnaswamy, Lisa C Richardson, Sheri Chernetsky-Tejedor, and Fusheng Wang. Improved identification of venous thromboembolism from electronic medical records using a novel information extraction software platform. *Medical care*, 56(9):e54, 2018. 31, 32, 33

[184] Kai-Uwe LewandrowskI, Narendran Muraleedharan, Steven Allen Eddy, Vikram Sobti, Brian D Reece, Jorge Felipe Ramírez León, and Sandeep Shah. Feasibility of deep learning algorithms for reporting in routine spine magnetic resonance imaging. *International Journal of Spine Surgery*, 14(s3):S86–S97, 2020. 31, 32

[185] Fabio Galbusera, Andrea Cina, Tito Bassani, Matteo Panico, and Luca Maria Sconfienza. Automatic diagnosis of spinal disorders on radiographic images: Leveraging existing unstructured datasets with natural language processing. *Global Spine Journal*, page 21925682211026910, 2021. 31, 32, 34

[186] Aditya V Karhade, Michiel ER Bongers, Olivier Q Groot, Thomas D Cha, Terence P Doorly, Harold A Fogel, Stuart H Hershman, Daniel G Tobert, Andrew J Schoenfeld, James D Kang, et al. Can natural language processing provide accurate, automated reporting of wound infection requiring reoperation after lumbar discectomy? *The Spine Journal*, 20(10):1602–1609, 2020. 31, 32, 33, 35

[187] Aditya V Karhade, Ophelie Lavoie-Gagne, Nicole Agaronnik, Hamid Ghaednia, Austin K Collins, David Shin, and Joseph H Schwab. Natural language processing for prediction of readmission in posterior lumbar fusion patients: which free-text notes have the most utility? *The Spine Journal*, 2021. 31, 32, 33

[188] Wendy W Chapman, Will Bridewell, Paul Hanbury, Gregory F Cooper, and Bruce G Buchanan. A simple algorithm for identifying negated findings and diseases in discharge summaries. *Journal of biomedical informatics*, 34(5):301–310, 2001. 33

[189] Shuai Zheng, James J Lu, Nima Ghasemzadeh, Salim S Hayek, Arshed A Quyyumi, and Fusheng Wang. Effective information extraction framework for heterogeneous clinical reports using online machine learning and controlled vocabularies. *JMIR medical informatics*, 5(2):e12, 2017. 33

[190] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008, 2017. 34

[191] David M Blei, Andrew Y Ng, and Michael I Jordan. Latent dirichlet allocation. *the Journal of machine Learning research*, 3:993–1022, 2003. 34

[192] Katherine P Liao, Jiehuan Sun, Tianrun A Cai, Nicholas Link, Chuan Hong, Jie Huang, Jennifer E Huffman, Jessica Gronsbell, Yichi Zhang, Yuk-Lam Ho, et al. High-throughput multimodal automated phenotyping (map) with application to phewas. *Journal of the American Medical Informatics Association*, 26(11):1255–1262, 2019. 34

[193] Felice Dell'Orletta, Giulia Venturi, Andrea Cimino, and Simonetta Montemagni. T2kˆ 2: a system for automatically extracting and organizing knowledge from texts. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 2062–2070, 2014. 35

[194] Katerina T Frantzi and Sophia Ananiadou. The c-value/nc-value domain-independent method for multi-word term extraction. *Journal of Natural Language Processing*, 6(3):145–179, 1999. 35

[195] Luca Bacco, Andrea Cimino, Luca Paulon, Mario Merone, and Felice Dell'Orletta. A machine learning approach for sentiment analysis for italian reviews in healthcare. *Computational Linguistics CLiC-it 2020*, 630(699):16, 2020. 39

[196] Pierpaolo Basile, Danilo Croce, Valerio Basile, and Marco Polignano. Overview of the evalita 2018 aspect-based sentiment analysis task (absita). In *EVALITA Evaluation of NLP and Speech Tools for Italian*, pages 1–10. CEUR, 2018. 39

[197] Alessandra Teresa Cignarella, Simona Frenda, Valerio Basile, Cristina Bosco, Viviana Patti, Paolo Rosso, et al. Overview of the evalita 2018 task on irony detection in italian tweets (ironita). In *Sixth Evaluation Campaign of Natural Language Processing and Speech Tools for Italian (EVALITA 2018)*, volume 2263, pages 1–6. CEUR-WS, 2018. 39

[198] Francesco Barbieri, Valerio Basile, Danilo Croce, Malvina Nissim, Nicole Novielli, and Viviana Patti. Overview of the evalita 2016 sentiment polarity classification task. In *Proceedings of third Italian conference on computational linguistics (CLiC-it 2016) & fifth evaluation campaign of natural language processing and speech tools for Italian. Final Workshop (EVALITA 2016)*, 2016. 39

[199] Judit Bar-Ilan. Data collection methods on the web for infometric purposes - a review and analysis. *Scientometrics*, 50(1):7–32, 2001. 40

[200] Bo Zhao. Web scraping. *Encyclopedia of big data*, pages 1–3, 2017. 40

[201] Stephen J Mooney, Daniel J Westreich, and Abdulrahman M El-Sayed. Epidemiology in the era of big data. *Epidemiology (Cambridge, Mass.)*, 26(3):390, 2015. 40

[202] Richard Baron Penman, Timothy Baldwin, and David Martinez. Web scraping made simple with sitescraper. *Penman Web Scraping*, pages 1–10, 2009. 40

[203] De S Sirisuriya et al. A comparative study on web scraping. 2015. 40

[204] Leonard Richardson. Beautiful soup documentation. *Dosegljivo: https://www. crummy. com/software/BeautifulSoup/bs4/doc/.[Dostopano: 7. 7. 2018]*, 2007. 40

[205] Gábor László Hajba. Using beautiful soup. In *Website Scraping with Python*, pages 41–96. Springer, 2018. 40

[206] Saif M Mohammad, Svetlana Kiritchenko, and Xiaodan Zhu. Nrc-canada: Building the state-of-the-art in sentiment analysis of tweets. *arXiv preprint arXiv:1308.6242*, 2013. 42

[207] Rong-En Fan, Kai-Wei Chang, Cho-Jui Hsieh, Xiang-Rui Wang, and Chih-Jen Lin. Liblinear: A library for large linear classification. *the Journal of machine Learning research*, 9:1871–1874, 2008. 42

[208] Andrea Cimino and Felice Dell'Orletta. Tandem lstm-svm approach for sentiment analysis. In *of the Final Workshop 7 December 2016, Naples*, page 172, 2016. 42

[209] Andrea Cimino, Lorenzo De Mattei, and Felice Dell'Orletta. Multi-task learning in deep neural networks at evalita 2018. *Proceedings of the 6th evaluation campaign of Natural Language Processing and Speech tools for Italian (EVALITA'18)*, pages 86–95, 2018. 43

[210] John Platt et al. Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. *Advances in large margin classifiers*, 10(3):61–74, 1999. 46

[211] Debora Nozza, Federico Bianchi, and Dirk Hovy. What the [mask]? making sense of language-specific bert models. *arXiv preprint arXiv:2003.02912*, 2020. 48

[212] Marco Polignano, Pierpaolo Basile, Marco De Gemmis, Giovanni Semeraro, and Valerio Basile. Alberto: Italian bert language understanding model for nlp challenging tasks based on tweets. In *6th Italian Conference on Computational Linguistics, CLiC-it 2019*, volume 2481, pages 1–6. CEUR, 2019. 48, 49

[213] Leonardo Ranaldi, Michele Mastromattei, Dario Onorati, Elena Sofia Ruzzetti, Francesca Fallucchi, and Fabio Massimo Zanzotto. Kermit for sentiment analysis in italian healthcare reviews. In *CLiC-it*, 2021. 49

[214] Fabio Massimo Zanzotto, Andrea Santilli, Leonardo Ranaldi, Dario Onorati, Pierfrancesco Tommasino, and Francesca Fallucchi. Kermit: complementing transformer architectures with encoders of explicit syntactic interpretations. In *Proceedings of the 2020 conference on empirical methods in natural language processing (EMNLP)*, pages 256–267, 2020. 49

[215] Leonardo Ranaldi, Francesca Fallucchi, Andrea Santilli, and Fabio Massimo Zanzotto. Kermit: Visualizing neural network activations on syntactic trees. In *Research Conference on Metadata and Semantics Research*, pages 139–147. Springer, 2022. 49

[216] Maria Chiara Martinis, Chiara Zucco, and Mario Cannataro. An italian lexicon-based sentiment analysis approach for medical applications. BCB '22, New York, NY, USA, 2022. Association for Computing Machinery. 49

[217] C. Hutto and Eric Gilbert. Vader: A parsimonious rule-based model for sentiment analysis of social media text. *Proceedings of the International AAAI Conference on Web and Social Media*, 8(1):216–225, May 2014. 49

[218] Sharon Swee-Lin Tan and Nadee Goonawardene. Internet health information seeking and the patient-physician relationship: a systematic review. *Journal of medical Internet research*, 19(1):e5729, 2017. 50

[219] Allison Tong, Andrew S. Levey, Kai-Uwe Eckardt, Samaya Anumudu, Cristina M. Arce, Amanda Baumgart, Louese Dunn, Talia Gutman, Tess Harris, Liz Lightstone, Nicole Scholes-Robertson, Jenny I. Shen, David C. Wheeler, David M. White, Martin Wilkie, Jonathan C. Craig, Michel Jadoul, and Wolfgang C. Winkelmayer. Patient and caregiver perspectives on terms used to describe kidney health. *Clinical Journal of the American Society of Nephrology*, 15(7):937–948, 2020. 50

[220] Alexandra King. Poor health literacy: a'hidden'risk factor. *Nature Reviews Cardiology*, 7(9):473–474, 2010. 50

[221] Yixin Cao, Ruihao Shui, Liangming Pan, Min-Yen Kan, Zhiyuan Liu, and Tat-Seng Chua. Expertise style transfer: A new task towards better communication between experts and laymen. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1061–1071, Online, July 2020. Association for Computational Linguistics. 51, 53, 54, 56, 63, 64, 66, 67

[222] Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880, Online, July 2020. Association for Computational Linguistics. 51, 57

[223] Huiyuan Lai, Antonio Toral, and Malvina Nissim. Thank you BART! rewarding pre-trained models improves formality style transfer. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 484–494, Online, August 2021. Association for Computational Linguistics. 51, 52, 64

[224] Huiyuan Lai, Antonio Toral, and Malvina Nissim. Generic resources are what you need: Style transfer tasks without task-specific parallel training data. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 4241–4254, Online and Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics. 51, 52, 53, 64

[225] Nils Reimers and Iryna Gurevych. Sentence-bert: Sentence embeddings using siamese bert-networks. *arXiv preprint arXiv:1908.10084*, 2019. 51, 59

[226] Leon A Gatys, Alexander S Ecker, and Matthias Bethge. A neural algorithm of artistic style. *arXiv preprint arXiv:1508.06576*, 2015. 52

[227] Haozhi Huang, Hao Wang, Wenhan Luo, Lin Ma, Wenhao Jiang, Xiaolong Zhu, Zhifeng Li, and Wei Liu. Real-time neural style transfer for videos. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 783–791, 2017. 52

[228] Yongcheng Jing, Yezhou Yang, Zunlei Feng, Jingwen Ye, Yizhou Yu, and Mingli Song. Neural style transfer: A review. *IEEE Transactions on Visualization and Computer Graphics*, 26:3365–3385, 11 2020. 52

[229] Yu Hwan Kim, Se Hyun Nam, Seung Baek Hong, and Kang Ryoung Park. Gra-gan: Generative adversarial network for image style transfer of gender, race, and age. *Expert Systems with Applications*, 198:116792, 2022. 52

[230] Ondřej Cífka, Umut Şimşekli, and Gaël Richard. Groove2groove: One-shot music style transfer with supervision from synthetic data. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 28:2638–2650, 2020. 52

[231] Sreetama Mukherjee and Manjunath Mulimani. Composeinstyle: Music composition with and without style transfer. *Expert Systems with Applications*, 191:116195, 2022. 52

[232] Sudha Rao and Joel Tetreault. Dear sir or madam, may i introduce the gyafc dataset: Corpus, benchmarks and metrics for formality style transfer. *arXiv preprint arXiv:1803.06535*, 2018. 52

[233] Tong Niu and Mohit Bansal. Polite dialogue generation without parallel data. *Transactions of the Association for Computational Linguistics*, 6:373–389, 2018. 52

[234] Aman Madaan, Amrith Setlur, Tanmay Parekh, Barnabas Poczos, Graham Neubig, Yiming Yang, Ruslan Salakhutdinov, Alan W Black, and Shrimai Prabhumoye. Politeness transfer: A tag and generate approach. *arXiv preprint arXiv:2004.14257*, 2020. 52

[235] Tianxiao Shen, Tao Lei, Regina Barzilay, and Tommi Jaakkola. Style transfer from non-parallel text by cross-alignment. *Advances in neural information processing systems*, 30, 2017. 52

[236] Di Jin, Zhijing Jin, Zhiting Hu, Olga Vechtomova, and Rada Mihalcea. Deep learning for text style transfer: A survey. *Computational Linguistics*, pages 1–51, 2021. 52, 57

[237] Martina Toshevska and Sonja Gievska. A review of text style transfer using deep learning. *IEEE Transactions on Artificial Intelligence*, 3(5):669–684, 2022. 52

[238] Zhiting Hu, Zichao Yang, Xiaodan Liang, Ruslan Salakhutdinov, and Eric P. Xing. Toward controlled generation of text. In Doina Precup and Yee Whye Teh, editors, *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 1587–1596. PMLR, 06–11 Aug 2017. 52, 63

[239] Zhenxin Fu, Xiaoye Tan, Nanyun Peng, Dongyan Zhao, and Rui Yan. Style transfer in text: Exploration and evaluation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32, 2018. 52

[240] Yanai Elazar and Yoav Goldberg. Adversarial removal of demographic attributes from text data. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 11–21, Brussels, Belgium, October-November 2018. Association for Computational Linguistics. 52

[241] Guillaume Lample, Sandeep Subramanian, Eric Smith, Ludovic Denoyer, Marc'Aurelio Ranzato, and Y-Lan Boureau. Multiple-attribute text rewriting. In *International Conference on Learning Representations*, 2019. 52, 53

[242] Juncen Li, Robin Jia, He He, and Percy Liang. Delete, retrieve, generate: a simple approach to sentiment and style transfer. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1865–1874, New Orleans, Louisiana, June 2018. Association for Computational Linguistics. 53, 64

[243] Matthew Shardlow and Raheel Nawaz. Neural text simplification of clinical letters with a domain specific phrase table. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 380–389, Florence, Italy, July 2019. Association for Computational Linguistics. 53, 63

[244] Qing Zeng-Treitler, Sergey Goryachev, Hyeoneui Kim, Alla Keselman, and Douglas Rosendale. Making texts in electronic health records comprehensible to consumers: a prototype translator. In *AMIA Annual Symposium Proceedings*, volume 2007, page 846. American Medical Informatics Association, 2007. 53

[245] Wei-Hung Weng, Yu-An Chung, and Peter Szolovits. Unsupervised clinical language translation. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 3121–3131, 2019. 53

[246] VG Vinod Vydiswaran, Qiaozhu Mei, David A Hanauer, and Kai Zheng. Mining consumer health vocabulary from community-generated text. In *AMIA Annual Symposium Proceedings*, volume 2014, page 1150. American Medical Informatics Association, 2014. 53

[247] Enrico Manzini, Jon Garrido-Aguirre, Jordi Fonollosa, and Alexandre Perera-Lluna. Mapping layperson medical terminology into the human phenotype ontology using neural machine translation models. *Expert Systems with Applications*, 204:117446, 2022. 53

[248] Rico Sennrich, Barry Haddow, and Alexandra Birch. Improving neural machine translation models with monolingual data. *arXiv preprint arXiv:1511.06709*, 2015. 53

[249] Shrimai Prabhumoye, Yulia Tsvetkov, Ruslan Salakhutdinov, and Alan W. Black. Style transfer through back-translation. *CoRR*, abs/1804.09000, 2018. 53

[250] Ella Rabinovich, Shachar Mirkin, Raj Nath Patel, Lucia Specia, and Shuly Wintner. Personalized machine translation: Preserving original author traits. *arXiv preprint arXiv:1610.05461*, 2016. 53

[251] Ning Dai, Jianze Liang, Xipeng Qiu, and Xuanjing Huang. Style transformer: Unpaired text style transfer without disentangled latent representation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5997–6007, Florence, Italy, July 2019. Association for Computational Linguistics. 53, 64

[252] Chulun Zhou, Liangyu Chen, Jiachen Liu, Xinyan Xiao, Jinsong Su, Sheng Guo, and Hua Wu. Exploring contextual word-level style relevance for unsupervised style transfer. *arXiv preprint arXiv:2005.02049*, 2020. 53

[253] Zhijing Jin, Di Jin, Jonas Mueller, Nicholas Matthews, and Enrico Santus. IMaT: Unsupervised text attribute transfer via iterative matching and translation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3097–3109, Hong Kong, China, November 2019. Association for Computational Linguistics. 53

[254] Junyu Luo, Zifei Zheng, Hanzhong Ye, Muchao Ye, Yaqing Wang, Quanzeng You, Cao Xiao, and Fenglong Ma. A benchmark dataset for understandable medical language translation. *arXiv preprint arXiv:2012.02420*, 2020. 53

[255] Alistair EW Johnson, Tom J Pollard, Lu Shen, Li-wei H Lehman, Mengling Feng, Mohammad Ghassemi, Benjamin Moody, Peter Szolovits, Leo Anthony Celi, and Roger G Mark. Mimic-iii, a freely accessible critical care database. *Scientific data*, 3(1):1–9, 2016. 53

[256] Wenda Xu, Michael Saxon, Misha Sra, and William Yang Wang. Self-supervised knowledge assimilation for expert-layman text style transfer. *arXiv preprint arXiv:2110.02950*, 2021. 53, 54, 61

[257] Mikel Artetxe and Holger Schwenk. Massively Multilingual Sentence Embeddings for Zero-Shot Cross-Lingual Transfer and Beyond. *Transactions of the Association for Computational Linguistics*, 7:597–610, 09 2019. 53

[258] Aizhan Imankulova, Takayuki Sato, and Mamoru Komachi. Improving low-resource neural machine translation with filtered pseudo-parallel corpus. In *Proceedings of the 4th Workshop on Asian Translation (WAT2017)*, pages 70–78, Taipei, Taiwan, November 2017. Asian Federation of Natural Language Processing. 54

[259] Aizhan Imankulova, Takayuki Sato, and Mamoru Komachi. Filtered pseudo-parallel corpus improves low-resource neural machine translation. *ACM Trans. Asian Low-Resour. Lang. Inf. Process.*, 19(2), oct 2019. 54

[260] Huiyuan Lai, Antonio Toral, and Malvina Nissim. Multilingual pre-training with language and task adaptation for multilingual text style transfer. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 262–271, Dublin, Ireland, May 2022. Association for Computational Linguistics. 54

[261] Laurens van den Bercken, Robert-Jan Sips, and Christoph Lofi. Evaluating neural text simplification in the medical domain. In *The World Wide Web Conference*, WWW '19, page 3286–3292, New York, NY, USA, 2019. Association for Computing Machinery. 54

[262] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA, July 2002. Association for Computational Linguistics. 54, 64

[263] Benjamin Marie and Atsushi Fujita. Efficient extraction of pseudo-parallel sentences from raw monolingual data using word embeddings. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 392–398, 2017. 54

[264] Shaolin Zhu, Yong Yang, and Chun Xu. Extracting parallel sentences from nonparallel corpora using parallel hierarchical attention network. *Computational Intelligence and Neuroscience*, 2020, 2020. 54

[265] Eleftheria Briakou, Sweta Agrawal, Joel Tetreault, and Marine Carpuat. Evaluating the evaluation metrics for style transfer: A case study in multilingual formality transfer. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 1321–1336, Online and Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics. 54

[266] Huiyuan Lai, Jiali Mao, Antonio Toral, and Malvina Nissim. Human judgement as a compass to navigate automatic metrics for formality transfer. In *Proceedings of the 2nd Workshop on Human Evaluation of NLP Systems (HumEval)*, pages 102–115, Dublin, Ireland, May 2022. Association for Computational Linguistics. 54, 64, 82

[267] Laura Vásquez-Rodríguez, Matthew Shardlow, Piotr Przybyła, and Sophia Ananiadou. Investigating text simplification evaluation. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 876–882, 2021. 54

[268] Chandrayee Basu, Rosni Vasu, Michihiro Yasunaga, Sohyeong Kim, and Qian Yang. Automatic medical text simplification: Challenges of data quality and curation. 2021. 54

[269] Ashwin Devaraj, Iain Marshall, Byron Wallace, and Junyi Jessy Li. Paragraph-level simplification of medical texts. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4972–4984, Online, June 2021. Association for Computational Linguistics. 54

[270] Yue Guo, Wei Qiu, Yizhong Wang, and Trevor Cohen. Automated lay language summarization of biomedical scientific reviews. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 160–168, 2021. 54

[271] Natalia Grabar and Rémi Cardon. Clear–simple corpus for medical french. In *Proceedings of the 1st Workshop on Automatic Text Adaptation (ATA)*, pages 3–9, 2018. 54

[272] Yanshan Wang, Sunyang Fu, Feichen Shen, Sam Henry, Ozlem Uzuner, Hongfang Liu, et al. The 2019 n2c2/ohnlp track on clinical semantic textual similarity: overview. *JMIR Medical Informatics*, 8(11):e23375, 2020. 55

[273] Yanshan Wang, Naveed Afzal, Sijia Liu, Majid Rastegar-Mojarad, Liwei Wang, Feichen Shen, Sunyang Fu, and Hongfang Liu. Overview of the biocreative/ohnlp challenge

2018 task 2: clinical semantic textual similarity. *Proceedings of the BioCreative/OHNLP Challenge*, 2018, 2018. 55

[274] Yanshan Wang, Naveed Afzal, Sunyang Fu, Liwei Wang, Feichen Shen, Majid Rastegar-Mojarad, and Hongfang Liu. Medsts: a resource for clinical semantic textual similarity. *Language Resources and Evaluation*, 54(1):57–72, 2020. 55, 67

[275] Clara H McCreery, Namit Katariya, Anitha Kannan, Manish Chablani, and Xavier Amatriain. Effective transfer learning for identifying similar questions: matching user questions to covid-19 faqs. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 3458–3465, 2020. 55

[276] Luca Soldaini and Nazli Goharian. Quickumls: a fast, unsupervised approach for medical concept extraction. In *MedIR workshop, sigir*, pages 1–4, 2016. 56

[277] Olivier Bodenreider. The unified medical language system (umls): integrating biomedical terminology. *Nucleic acids research*, 32(suppl_1):D267–D270, 2004. 56

[278] Jeff Johnson, Matthijs Douze, and Hervé Jégou. Billion-scale similarity search with GPUs. *IEEE Transactions on Big Data*, 7(3):535–547, 2019. 59

[279] Matthew Henderson, Rami Al-Rfou, Brian Strope, Yun-Hsuan Sung, László Lukács, Ruiqi Guo, Sanjiv Kumar, Balint Miklos, and Ray Kurzweil. Efficient natural language response suggestion for smart reply. *ArXiv*, abs/1705.00652, 2017. 60

[280] R. Hadsell, S. Chopra, and Y. LeCun. Dimensionality reduction by learning an invariant mapping. In *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06)*, volume 2, pages 1735–1742, 2006. 60

[281] Tianyu Gao, Xingcheng Yao, and Danqi Chen. SimCSE: Simple contrastive learning of sentence embeddings. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6894–6910, Online and Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics. 60

[282] Geoffrey E Hinton, Nitish Srivastava, Alex Krizhevsky, Ilya Sutskever, and Ruslan R Salakhutdinov. Improving neural networks by preventing co-adaptation of feature detectors. *arXiv preprint arXiv:1207.0580*, 2012. 60

[283] Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. Dropout: A simple way to prevent neural networks from overfitting. *Journal of Machine Learning Research*, 15(56):1929–1958, 2014. 60

[284] Vu Cong Duy Hoang, Philipp Koehn, Gholamreza Haffari, and Trevor Cohn. Iterative back-translation for neural machine translation. In *Proceedings of the 2nd Workshop on*

Neural Machine Translation and Generation, pages 18–24, Melbourne, Australia, July 2018. Association for Computational Linguistics. 63

[285] Guillaume Klein, Yoon Kim, Yuntian Deng, Jean Senellart, and Alexander Rush. OpenNMT: Open-source toolkit for neural machine translation. In *Proceedings of ACL 2017, System Demonstrations*, pages 67–72, Vancouver, Canada, July 2017. Association for Computational Linguistics. 63

[286] Sai Surya, Abhijit Mishra, Anirban Laha, Parag Jain, and Karthik Sankaranarayanan. Unsupervised neural text simplification. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2058–2068, Florence, Italy, July 2019. Association for Computational Linguistics. 63

[287] Fuli Luo, Peng Li, Jie Zhou, Pengcheng Yang, Baobao Chang, Zhifang Sui, and Xu Sun. A dual reinforcement learning framework for unsupervised text style transfer. In *Proceedings of the 28th International Joint Conference on Artificial Intelligence*, pages 5116–5122, 2019. 64

[288] Abhilasha Sancheti, Kundan Krishna, Balaji Vasan Srinivasan, and Anandhavelu Natarajan. Reinforced rewards framework for text style transfer. In *Advances in Information Retrieval*, pages 545–560, 2020. 64

[289] Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. Bertscore: Evaluating text generation with bert. In *International Conference on Learning Representations*, 2020. 64

[290] Thibault Sellam, Dipanjan Das, and Ankur Parikh. BLEURT: Learning robust metrics for text generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7881–7892, Online, July 2020. Association for Computational Linguistics. 64

[291] Ricardo Rei, Craig Stewart, Ana C Farinha, and Alon Lavie. COMET: A neural framework for MT evaluation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2685–2702, Online, November 2020. Association for Computational Linguistics. 64

[292] Yoon Kim. Convolutional neural networks for sentence classification. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1746–1751, Doha, Qatar, October 2014. Association for Computational Linguistics. 64

[293] Julian Salazar, Davis Liang, Toan Q. Nguyen, and Katrin Kirchhoff. Masked language model scoring. In *Proceedings of the 58th Annual Meeting of the Association for Compu-*

tational Linguistics, pages 2699–2712, Online, July 2020. Association for Computational Linguistics. 65

[294] Jacob Cohen. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20(1):37–46, 1960. 67

[295] Eleftheria Briakou, Sweta Agrawal, Ke Zhang, Joel Tetreault, and Marine Carpuat. A review of human evaluation for style transfer. In *Proceedings of the 1st Workshop on Natural Language Generation, Evaluation, and Metrics (GEM 2021)*, pages 58–67, Online, August 2021. Association for Computational Linguistics. 68

[296] Jacob Cohen. Weighted kappa: nominal scale agreement provision for scaled disagreement or partial credit. *Psychological bulletin*, 70(4):213, 1968. 68

[297] Sophie Vanbelle. A new interpretation of the weighted kappa coefficients. *Psychometrika*, 81(2):399–410, 2016. 68, 79

[298] J Richard Landis and Gary G Koch. The measurement of observer agreement for categorical data. *biometrics*, pages 159–174, 1977. 78

[299] Patrick E Shrout. Measurement reliability and agreement in psychiatry. *Statistical methods in medical research*, 7(3):301–317, 1998. 79

[300] Lorenzo De Mattei, Michele Cafagna, Felice Dell'Orletta, and Malvina Nissim. Invisible to people but not to machines: Evaluation of style-aware HeadlineGeneration in absence of reliable human judgment. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 6709–6717, Marseille, France, May 2020. European Language Resources Association. 82

[301] Kia Dashtipour et al. A novel context-aware multimodal framework for persian sentiment analysis. *arXiv preprint arXiv:2103.02636*, 2021. 87

[302] Luca Bacco, Andrea Cimino, Felice Dell'Orletta, and Mario Merone. Explainable sentiment analysis: A hierarchical transformer-based extractive summarization approach. *Electronics*, 10(18), 2021. 88

[303] Luca Bacco, Andrea Cimino, Felice Dell'Orletta, and Mario Merone. Extractive summarization for explainable sentiment analysis using transformers. In *Sixth International Workshop on eXplainable SENTIment Mining and EmotioN deTection*, 2021. 88

[304] Andrew Maas et al. Learning word vectors for sentiment analysis. In *Proceedings of the 49th annual meeting of the association for computational linguistics: Human language technologies*, pages 142–150, 2011. 88, 95, 115

[305] Bing Liu. Sentiment analysis and opinion mining. *Synthesis lectures on human language technologies*, 5(1):1–167, 2012. 89

[306] Mary McGlohon et al. Star quality: Aggregating reviews to rank products and merchants. In *Fourth international AAAI conference on weblogs and social media*, 2010. 89

[307] Hafed Zarzour, Mahmoud Al-Ayyoub, Yaser Jararweh, et al. Sentiment analysis based on deep learning methods for explainable recommendations with reviews. In *2021 12th International Conference on Information and Communication Systems (ICICS)*, pages 452–456. IEEE, 2021. 89

[308] Shilpa Gite, Hrituja Khatavkar, Ketan Kotecha, Shilpi Srivastava, Priyam Maheshwari, and Neerav Pandey. Explainable stock prices prediction from financial news articles using sentiment analysis. *PeerJ Computer Science*, 7:e340, 2021. 89

[309] Mahesh Joshi et al. Movie reviews and revenues: An experiment in text regression. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 293–296, 2010. 89

[310] Minqing Hu and Bing Liu. Mining opinion features in customer reviews. In *AAAI*, volume 4, pages 755–760, 2004. 90

[311] Thiago De Sousa Silveira et al. Using aspect-based analysis for explainable sentiment predictions. In *CCF International Conference on Natural Language Processing and Chinese Computing*, 2019. 90

[312] Stefano Baccianella et al. Sentiwordnet 3.0: an enhanced lexical resource for sentiment analysis and opinion mining. In *Lrec*, 2010. 90

[313] Erik Cambria et al. Senticnet: A publicly available semantic resource for opinion mining. In *AAAI fall symposium: commonsense knowledge*, 2010. 90

[314] Yongfeng Zhang et al. Explicit factor models for explainable recommendation based on phrase-level sentiment analysis. In *Proceedings of the 37th international ACM SIGIR conference on Research & development in information retrieval*, 2014. 90

[315] Anping Zhao and Yu Yu. Knowledge-enabled bert for aspect-based sentiment analysis. *Knowledge-Based Systems*, page 107220, 2021. 91

[316] Wafaa S El-Kassas et al. Automatic text summarization: A comprehensive survey. *Expert Systems with Applications*, page 113679, 2020. 91, 92

[317] Raghavendra Pappagari et al. Hierarchical transformers for long document classification. In *2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, 2019. 92

[318] Andraž Pelicon et al. Zero-shot learning for cross-lingual news sentiment classification. *Applied Sciences*, 2020. 93

[319] Xingxing Zhang et al. Hibert: Document level pre-training of hierarchical bidirectional transformers for document summarization. *arXiv preprint arXiv:1905.06566*, 2019. 93

[320] Shusheng Xu et al. Unsupervised extractive summarization by pre-training hierarchical transformers. *arXiv preprint arXiv:2010.08242*, 2020. 93

[321] Olga Kovaleva et al. Revealing the dark secrets of bert. *arXiv preprint arXiv:1908.08593*, 2019. 94

[322] Elena Voita et al. Context-aware neural machine translation learns anaphora resolution. *arXiv preprint arXiv:1805.10163*, 2018. 94

[323] Yoav Goldberg. Assessing bert's syntactic abilities. *arXiv preprint arXiv:1901.05287*, 2019. 94

[324] Thomas Wolf. Some additional experiments extending the tech report "assessing berts syntactic abilities" by yoav goldberg. Technical report, 2019. 94

[325] John Hewitt and Christopher D Manning. A structural probe for finding syntax in word representations. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4129–4138, 2019. 94

[326] Alessandro Raganato et al. An analysis of encoder representations in transformer-based machine translation. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, 2018. 94

[327] Jesse Vig and Yonatan Belinkov. Analyzing the structure of attention in a transformer language model. *arXiv preprint arXiv:1906.04284*, 2019. 94

[328] Elena Voita et al. Analyzing multi-head self-attention: Specialized heads do the heavy lifting, the rest can be pruned. *arXiv preprint arXiv:1905.09418*, 2019. 94

[329] Leopold Franz et al. A deep learning pipeline for patient diagnosis prediction using electronic health records. *arXiv preprint arXiv:2006.16926*, 2020. 94

[330] Gaël Letarte et al. Importance of self-attention for sentiment analysis. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, 2018. 94

[331] Francesco Bodria et al. Explainability methods for natural language processing: Applications to sentiment analysis (discussion paper). 2020. 94

[332] Klaus Krippendorff. Estimating the reliability, systematic error and random error of interval data. *Educational and Psychological Measurement*, 1970. 95

[333] Klaus Krippendorff. Content analysis: An introduction to its methodology (2 nd thousand oaks, 2004. 96

[334] Klaus Krippendorff. Reliability in content analysis: Some common misconceptions and recommendations. *Human communication research*, 2004. 96

[335] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. Attention is all you need. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. 110

[336] Yaru Hao, Li Dong, Furu Wei, and Ke Xu. Visualizing and understanding the effectiveness of BERT. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 4143–4152, Hong Kong, China, November 2019. Association for Computational Linguistics. 110

[337] Isabel Papadimitriou and Dan Jurafsky. Learning Music Helps You Read: Using transfer to study linguistic structure in language models. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6829–6839, Online, November 2020. Association for Computational Linguistics. 110

[338] Hirokatsu Kataoka, Kazushige Okayasu, Asato Matsumoto, Eisuke Yamagata, Ryosuke Yamada, Nakamasa Inoue, Akio Nakamura, and Yutaka Satoh. Pre-training without natural images. In *Proceedings of the Asian Conference on Computer Vision (ACCV)*, November 2020. 110

[339] Alaaeldin El-Nouby, Gautier Izacard, Hugo Touvron, Ivan Laptev, Hervé Jegou, and Edouard Grave. Are large-scale datasets necessary for self-supervised pre-training? *arXiv preprint arXiv:2112.10740*, 2021. 110

[340] Yonatan Oren, Shiori Sagawa, Tatsunori B. Hashimoto, and Percy Liang. Distributionally robust language modeling. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 4227–4237, Hong Kong, China, November 2019. Association for Computational Linguistics. 110

[341] Taesun Whang, Dongyub Lee, Chanhee Lee, Kisu Yang, Dongsuk Oh, and HeuiSeok Lim. An effective domain adaptive post-training method for bert in response selection, 2019. 111

[342] Chi Sun, Xipeng Qiu, Yige Xu, and Xuanjing Huang. How to fine-tune bert for text classification? In Maosong Sun, Xuanjing Huang, Heng Ji, Zhiyuan Liu, and Yang Liu, editors, *Chinese Computational Linguistics*, pages 194–206, Cham, 2019. Springer International Publishing. 111

[343] Suchin Gururangan, Ana Marasović, Swabha Swayamdipta, Kyle Lo, Iz Beltagy, Doug Downey, and Noah A. Smith. Don't stop pretraining: Adapt language models to domains and tasks. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8342–8360, Online, July 2020. Association for Computational Linguistics. 111, 114

[344] Chien-Sheng Wu, Steven C.H. Hoi, Richard Socher, and Caiming Xiong. TOD-BERT: Pre-trained natural language understanding for task-oriented dialogue. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 917–929, Online, November 2020. Association for Computational Linguistics. 111

[345] Yi Yang, Mark Christopher Siy UY, and Allen Huang. FinBERT: A Pretrained Language Model for Financial Communications. *arXiv e-prints*, page arXiv:2006.08097, June 2020. 111

[346] Ilias Chalkidis, Manos Fergadiotis, Prodromos Malakasiotis, Nikolaos Aletras, and Ion Androutsopoulos. LEGAL-BERT: The muppets straight out of law school. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 2898–2904, Online, November 2020. Association for Computational Linguistics. 111

[347] Francesco Barbieri, Jose Camacho-Collados, Luis Espinosa Anke, and Leonardo Neves. TweetEval: Unified benchmark and comparative evaluation for tweet classification. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1644–1650, Online, November 2020. Association for Computational Linguistics. 111, 115

[348] Tommaso Caselli, Valerio Basile, Jelena Mitrović, and Michael Granitzer. HateBERT: Retraining BERT for abusive language detection in English. In *Proceedings of the 5th*

Workshop on Online Abuse and Harms (WOAH 2021), pages 17–25, Online, August 2021. Association for Computational Linguistics. 111

[349] Shikib Mehri, Mihail Eric, and Dilek Hakkani-Tür. Dialoglue: A natural language understanding benchmark for task-oriented dialogue. *CoRR*, abs/2009.13570, 2020. 111

[350] Yao Qiu, Jinchao Zhang, and Jie Zhou. Different strokes for different folks: Investigating appropriate further pre-training approaches for diverse dialogue tasks. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 2318–2327, Online and Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics. 111

[351] Alexander Rietzler, Sebastian Stabinger, Paul Opitz, and Stefan Engl. Adapt or get left behind: Domain adaptation through BERT language model finetuning for aspect-target sentiment classification. *CoRR*, abs/1908.11860, 2019. 111

[352] Qi Zhu, Yuxian Gu, Lingxiao Luo, Bing Li, Cheng Li, Wei Peng, Minlie Huang, and Xiaoyan Zhu. When does further pre-training MLM help? an empirical study on task-oriented dialog pre-training. In *Proceedings of the Second Workshop on Insights from Negative Results in NLP*, pages 54–61, Online and Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics. 111, 114, 116, 118, 119

[353] Yuxin Sun, Dong Lao, Ganesh Sundaramoorthi, and Anthony Yezzi. Surprising instabilities in training deep networks and a theoretical analysis. *arXiv preprint arXiv:2206.02001*, 2022. 112

[354] Clemens Karner, Vladimir Kazeev, and Philipp Christian Petersen. Limitations of gradient descent due to numerical instability of backpropagation. *arXiv preprint arXiv:2210.00805*, 2022. 112

[355] Di Jin, Zhijing Jin, Joey Tianyi Zhou, and Peter Szolovits. Is bert really robust? a strong baseline for natural language attack on text classification and entailment. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(05):8018–8025, Apr. 2020. 112

[356] Yuan Zang, Bairu Hou, Fanchao Qi, Zhiyuan Liu, Xiaojun Meng, and Maosong Sun. Learning to attack: Towards textual adversarial attacking in real-world situations. *CoRR*, abs/2009.09192, 2020. 112

[357] Lichao Sun, Kazuma Hashimoto, Wenpeng Yin, Akari Asai, Jia Li, Philip S. Yu, and Caiming Xiong. Adv-bert: Bert is not robust on misspellings! generating nature adversarial samples on bert. *CoRR*, abs/2003.04985, 2020. 112

[358] Linyang Li, Ruotian Ma, Qipeng Guo, Xiangyang Xue, and Xipeng Qiu. BERT-ATTACK: Adversarial attack against BERT using BERT. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6193–6202, Online, November 2020. Association for Computational Linguistics. 112

[359] Anne Dirkson, Suzan Verberne, and Wessel Kraaij. Breaking bert: Understanding its vulnerabilities for biomedical named entity recognition through adversarial attack. *arXiv preprint arXiv:2109.11308*, 2021. 112

[360] Jesse Dodge, Gabriel Ilharco, Roy Schwartz, Ali Farhadi, Hannaneh Hajishirzi, and Noah Smith. Fine-Tuning Pretrained Language Models: Weight Initializations, Data Orders, and Early Stopping. *arXiv e-prints*, page arXiv:2002.06305, February 2020. 114

[361] Nasrin Mostafazadeh, Nathanael Chambers, Xiaodong He, Devi Parikh, Dhruv Batra, Lucy Vanderwende, Pushmeet Kohli, and James Allen. A corpus and cloze evaluation for deeper understanding of commonsense stories. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 839–849, San Diego, California, June 2016. Association for Computational Linguistics. 115

[362] Amil Merchant, Elahe Rahimtoroghi, Ellie Pavlick, and Ian Tenney. What happens to BERT embeddings during fine-tuning? In *Proceedings of the Third BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP*, pages 33–44, Online, November 2020. Association for Computational Linguistics. 116

Contributions in Computer Science and Bioengineering

Artificial Intelligence in Low Back Pain

Reference *Artificial Intelligence and Computer Aided Diagnosis in Chronic Low Back Pain: A Systematic Review* D'Antoni, F., Russo, F., Ambrosio, L., Bacco, L., Vollero, L., Vadalá, G., Dell'Orletta, F., Merone, M., Papalia, R., & Denaro, V. (2022), In *International Journal of Environmental Research and Public Health* https://doi.org/10.3390/ijerph19105971

Abstract Low Back Pain (LBP) is currently the first cause of disability in the world, with a significant socioeconomic burden. Diagnosis and treatment of LBP often involve a multidisciplinary, individualized approach consisting of several outcome measures and imaging data along with emerging technologies. The increased amount of data generated in this process has led to the development of methods related to artificial intelligence (AI), and to computer-aided diagnosis (CAD) in particular, which aim to assist and improve the diagnosis and treatment of LBP. In this manuscript, we have systematically reviewed the available literature on the use of CAD in the diagnosis and treatment of chronic LBP. A systematic research of PubMed, Scopus, and Web of Science electronic databases was performed. The search strategy was set as the combinations of the following keywords: "Artificial Intelligence", "Machine Learning", "Deep Learning", "Neural Network", "Computer Aided Diagnosis", "Low Back Pain", "Lumbar", "Intervertebral Disc Degeneration", "Spine Surgery", etc. The search returned a total of 1536 articles. After duplication removal and evaluation of the abstracts, 1386 were excluded, whereas 93 papers were excluded after full-text examination, taking the number of eligible articles to 57. The main applications of CAD in LBP included classification and regression. Classification is used to identify or categorize a disease, whereas regression is used to

produce a numerical output as a quantitative evaluation of some measure. The best performing systems were developed to diagnose degenerative changes of the spine from imaging data, with average accuracy rates >80%. However, notable outcomes were also reported for CAD tools executing different tasks including analysis of clinical, biomechanical, electrophysiological, and functional imaging data. Further studies are needed to better define the role of CAD in LBP care.

Decision Support Systems

Reference *Layered Meta-Learning Algorithm for Predicting Adverse Events in Type 1 Diabetes* In *IEEE Access* https://doi.org/10.1109/ACCESS.2023.3237992

Abstract Type 1 diabetes mellitus (T1D) is a chronic disease that, if not treated properly, can lead to serious complications. We propose a layered meta-learning approach based on multi-expert systems to predict adverse events in T1D. The base learner is composed of three deep neural networks and exploits only continuous glucose monitoring data as an input feature. Each network specializes in predicting whether the patient is about to experience hypoglycemia, hyperglycemia, or euglycemia. The output of the experts is passed to a meta-learner to provide the final model classification. In addition, we formally introduce a novel parameter, α, to evaluate the advance by which a prediction is performed. We evaluate the proposed approach on both a public and a private dataset and implement it on an edge device to test its feasibility in real life. On average, on the Ohio T1DM dataset, our system was able to predict hypoglycemia events with a time gain of 22.8 minutes, hyperglycemia ones with an advance of 24.0 minutes. Our model not only outperforms presented models in the literature in terms of events predicted with sufficient advance, but also with regard to the number of false positives, achieving on average 0.45 and 0.46 hypo- and hyperglycemic false alarms per day, respectively. Furthermore, the meta-learning approach effectively improves performance in a new cohort of patients by training only the meta-learner with a limited amount of data. We believe our approach would be an essential ally for the patients to control the glycemic fluctuations and adjust their insulin therapy and dietary intakes, enabling them to speed up decision-making and improve personal self-management, resulting in a reduced risk of acute and chronic complications. As our last contribution, we assessed the validity of the approach by exploiting only blood glucose variations as well as in

combination with the information of the insulin boluses, the skin temperature, and the galvanic skin response. In general, we have observed that providing other information but CGM leads to slightly lower performances with respect to considering CGM alone.

Reference *Machine Learning analysis of High-Grade Serous Ovarian Cancer proteomic dataset reveals novel candidate biomarkers*, Farinella, F., Merone, M., Bacco, L., Capirchio, A., Ciccozzi, M., & Caligiore, D. (2022), In *Scientific Reports* https://doi.org/10.1038/s41598-022-06788-2

Abstract Ovarian cancer is one of the most common gynecological malignancies, ranking third after cervical and uterine cancer. High-Grade Serous Ovarian Cancer (HGSOC) is one of the most aggressive subtype, and the late onset of its symptoms leads in most cases to an unfavourable prognosis. Current predictive algorithms used to estimate the risk of having Ovarian Cancer fail to provide sufficient sensitivity and specificity to be used widely in clinical practice. The use of additional biomarkers or parameters such as age or menopausal status to overcome these issues showed only weak improvements. It is necessary to identify novel molecular signatures and the development of new predictive algorithms able to support the diagnosis of HGSOC, and at the same time, deepen the understanding of this elusive disease, with the final goal of improving patient survival.
Here, we apply a Machine Learning-based pipeline to an open-source HGSOC Proteomic dataset to develop a decision support system (DSS) that displayed high discerning ability on a dataset of HGSOC biopsies. The proposed DSS consists of a double-step feature selection and a decision tree, with the resulting output consisting of a combination of three highly discriminating proteins: TOP1, PDIA4, and OGN, that could be of interest for further clinical and experimental validation. Furthermore, we took advantage of the ranked list of proteins generated during the feature selection steps to perform a pathway analysis to provide a snapshot of the main deregulated pathways of HGSOC.
The datasets used for this study are available in the Clinical Proteomic Tumor Analysis Consortium (CPTAC) data portal (https://cptac-data-portal.georgetown.edu/).

Pharmaceutical Sciences

Reference *The impact of the intestinal microbiota and the mucosal permeability on three different antibiotic drugs*, Palombo G., Merone M., Altomare A., Gori M., Terradura C., Bacco L., Del Chierico F., Putignani L., Cicala M., Guarino M., & Piemonte, V. (2021), https://doi.org/10.1016/j.ejps.2021.105869

Abstract **Background** The totality of bacteria, protozoa, viruses and fungi that lives in the human body is called microbiota. Human microbiota specifically colonizes the skin, the respiratory and urinary tract, the urogenital tract and the gastrointestinal system. This study focuses on the intestinal microbiota to explore the drug-microbiota relationship and, therefore, how the drug bioavailability changes in relation to the microbiota biodiversity to identify more personalized therapies, with the minimum risk of side effects.
Methods To achieve this goal, we developed a new mathematical model with two compartments, the intestine and the blood, which takes into account the colonic mucosal permeability variation - measured by Ussing chamber system on human colonic mucosal biopsies - and the fecal microbiota composition, determined through microbiota 16S rRNA sequencing analysis. Both of the clinical parameters were evaluated in a group of Irritable Bowel Syndrome patients compared to a group of healthy controls.
Key Results The results show that plasma drug concentration increases as bacterial concentration decreases, while it decreases as intestinal length decreases too.
Conclusions The study provides interesting data since in literature there are not yet mathematical models with these features, in which the importance of intestinal microbiota, the "*forgotten organ*", is considered both for the subject health state and in the nutrients and drugs metabolism.

Biometrics

Reference *Single beat ECG-based Identification System: development and robustness test in different working conditions*, Sorvillo, R., Bacco, L., Merone, M., Zompanti, A., Santonico, M., Pennazza, G., & Iannello, G. (2021), In *Proceedings of IEEE Metrology for Industry 4.0 and IoT (MetroInd4.0&IoT)* https://doi.org/10.1109/MetroInd4.0IoT51437.2021.9488474

Abstract One-lead electrocardiogram (ECG) tracings have already shown to be a good candidate as a feature for a biometric identification system. Also, the reduced computational burden and the fact that it can ensure that the subject is alive put the ECG ahead of currently used biometric features. Most of the literature provides studies exploiting acquisitions made with clinical instrumentation, preceded by invasive preparation of the subject, in a structured environment and with the subjects at rest. These conditions are not very feasible for an application in a real-world context. Therefore, we are proposing a system that is performant with acquisitions collected with (non-invasive) non-clinical instrumentation and in an unstructured environment, and that is robust to variations of the psycho-physical state of the subjects (i.e. at rest or under mental or physical stress). To do so, we developed an acquisition protocol that we followed to collect a new dataset to evaluate our method. The proposed system achieved up to the 97% of single segments (beats) classification accuracy when the test segments come from the same kind of acquisition procedure of the training beats. The same result was obtained by training and testing by combining the three trials. An 88% and 68% of accuracy were achieved by testing the system under mental and physical stress conditions, respectively, while trained at the rest state. Our findings suggest that the proposed method may be put at the base of a future application in a real-world context.